UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF FLORIDA

FLORIDA STATE CONFERENCE OF THE NATIONAL ASSOCIATION FOR THE ADVANCEMENT OF COLORED PEOPLE (NAACP), as an organization and representative of its members; *et al.*,

Plaintiffs,

Vs.

KURT S. BROWNING, in his official capacity as Secretary of State for the State of Florida,

Defendant.

Civil No. 4:07cv402 SPM/WCS

**DECLARATION OF
ANDREW BORTHWICK
IN SUPPORT OF
PLAINTIFFS' MOTION FOR A
PRELIMINARY INJUNCTION**

Pursuant to 28 U.S.C. § 1746, I, Andrew Borthwick, hereby declare as follows:

1. I am Principal Scientist at Spock Networks, Inc. ("Spock"), a company specializing in the indexing and classifying of public information about people from diverse sources. Our work depends on identifying, selecting, and matching discrete information about individuals from publicly available databases and websites. My current office is located at 1450 Veterans Boulevard, Redwood City, California 94063.

2. I am also a member of the Board of Directors, and former President and Chief Executive Officer, of ChoiceMaker Technologies, Inc. ("ChoiceMaker"), a company that I co-founded in 1998. ChoiceMaker is a data quality company specializing in the design and development of record-matching software. I submit this declaration in support of Plaintiffs' Motion for a Preliminary Injunction.

Background

3. I earned a B.A. from Oberlin College in 1988, graduating Phi Beta Kappa. I earned an M.S. and Ph.D. in Computer Science from New York University. My Ph.D. was awarded in 1999. My doctoral dissertation discussed a maximum entropy approach to named entity recognition, which in broad terms involves a learning technology that builds a model of the human decision-making process for identifying and categorizing proper names, in order to find names in context in newspaper text. For example, my dissertation discussed an approach useful for distinguishing articles about Calvin Klein, the individual, from articles about Calvin Klein, the company. I co-founded ChoiceMaker to apply the technology discussed in my dissertation to the record-matching field.

4. My academic expertise is in the fields of record-matching, machine learning, and computational linguistics. Of greatest relevance here is my background in the first, "record matching," which is a common term of art in the field of information science (also referred to as "informatics"). Record-matching refers to the process of identifying entries in a database (known as "records") that pertain to the same entity or person represented in other records, either in the same database or in another database. Often, these records are matched by comparing particular categories of data (known as "fields") within each record, such as a person's first name, last name, and date of birth. The science and challenge of record-matching involves identifying matching entries in the face of errors in one or both sets of data, or of inconsistencies between the data. I was

awarded U.S. Patent No. 6,523,019 in 2003 for a machine learning approach to record-matching, and U.S. Patent No. 7,152,060 in 2006 for a method of decreasing the processing time required for accurate record-matching through more efficient searching. I also am the co-inventor of one other process concerning record-matching with patent applications pending before the U.S. Patent and Trademark Office.

5. In 1998, I founded ChoiceMaker, and I was joined as co-founder by my business partner, Arthur Goldberg. ChoiceMaker is a data quality company specializing in record-matching software. Among other functions, ChoiceMaker software allows entities to identify corresponding records within and between databases. From 2000 to 2005, ChoiceMaker won a prestigious series of Small Business Innovation Research grants from the National Science Foundation for its research into a maximum entropy approach to approximate record matching.

6. In 2005, ChoiceMaker was awarded a contract with the U.S. Centers for Disease Control ("CDC") to use our record-matching system for the National Electronic Disease Surveillance System. This CDC surveillance project tracks the emergence and incidence of diseases in all 50 states in order to facilitate reporting to the CDC and to identify outbreaks of infectious diseases. The ChoiceMaker system was also purchased by nine states to track the academic records of every student in K through 12 public education, to support implementation of the No Child Left Behind Act. ChoiceMaker has also won multiple contracts with the New York City Department of Health to track immunizations of children, lead tests, and the incidence of communicable

diseases – which requires matching records from laboratories, hospitals, and clinics, in the face of numerous errors. ChoiceMaker was also used for the World Trade Center Health Registry to compile a register of everyone near the World Trade Center on 9/11 in order to track long-term health issues. ChoiceMaker was also awarded a contract with the New South Wales, Australia, Department of Health to develop a system for epidemiological research.[1]

7. In 2007, I became Principal Scientist at Spock Networks, Inc., which was founded in or around early 2006. Spock Networks is the premier online leader in personal search technologies, helping users find and discover people. Its success depends on its ability to accurately gather information on individuals, and sophisticated record linkage techniques are essential to this goal. That is, the principal Spock Networks product rapidly, reliably, and accurately finds public information that "matches" a specific individual while excluding information from individuals who appear very similar but are not in fact the same person.

8. I have published on, among other things, techniques involved in record-matching, including articles for peer-reviewed conferences in 1999 and 2004. A complete list of my publications is included in my *curriculum vitae*, a copy of which is attached as Exhibit A. I was invited to speak on record-matching at the First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, in conjunction with the

[1] ChoiceMaker has never been, and is not now, affiliated with ChoicePoint, Inc., Database Technologies (DBT Online), or any of their successors in interest.

Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining's Ninth International Conference on Knowledge Discovery and Data Mining, in Washington, D.C., on July 17, 2003. In addition, I have given presentations at the Massachusetts Institute of Technology's annual International Conference on Information Quality, the annual Information Quality Conference, and the annual National Immunization Conference, among others.

9. I regularly review publications in the record-matching field, in order to ensure that I am well versed in the current state of the art. Representative publications of this sort include *Institute of Electrical and Electronics Engineers ("IEEE") Transactions on Pattern Analysis and Machine Intelligence* and *IEEE Transactions on Knowledge and Data Engineering*. I am a member of the IEEE, the world's leading professional association for the advancement of technology, and the Association for Computing Machinery, an international scientific and educational organization dedicated to advancing the arts, sciences, and applications of information technology. I have submitted an expert declaration before a court in only one other matter, a case in Washington State that also pertained to the fallibility of record-matching processes in the voter registration context. My billing rate for this matter is $150 per hour.

Summary of Conclusions

10. For this case, I was asked to examine whether record-matching protocols used in Florida's voter registration program will result in a significant number of errors. In particular, I was asked to explain whether and why, in my professional opinion, errors endemic to information gathering, entry and maintenance, along with immaterial differences and inconsistencies across different databases, will result in the failure to match information from two different sources pertaining to the same individual. In the technical language of my field, the question I explored was whether an exact comparison of the characters in multiple fields of two different records, each filled out (or "populated") at different times by manual data entry from different handwritten forms or from information given orally to a data entry clerk, would likely result in substantial numbers of "false negatives" – that is, registration records for which the name and identifying information do not "match" the name and information in another database when, in fact, both records reflect the same person. For example, a "false negative" will result when a woman registers to vote with her married name, her maiden name is listed on her Social Security record, and the two records fail to "match."

11. To prepare to give my opinions in this case, I read the applicable federal and State laws and regulations, including relevant portions of the Help America Vote Act (and in particular 42 U.S.C. § 15483) and Fla. Stat. § 97.053(6), my own academic and professional work, and relevant publications in the field. Those publications include papers in the *Statistical Research Report* series of the U.S. Bureau of

the Census, articles published in peer reviewed journals such as *Computers and Biomedical Research* (now known as the *Journal of Biomedical Informatics*) and the *Journal of the American Medical Informatics Association* (AMIA), and proceedings of the AMIA's annual symposium. I also reviewed documents provided to the Plaintiffs, including letters from the Florida Secretary of State to the Advancement Project and others, describing the registration process in Florida.

12. To learn more specifically about Florida's record-matching process, I also reviewed documents including the *"Social Security Verification" System Specification* distributed by the American Association of Motor Vehicle Administrators in August 2004. In particular, I reviewed the Help America Vote Verification ("HAVV") process (also referred to as a "transaction") and the matching protocol described on pages 26 and 27 of that document; a copy of the relevant pages is attached as Exhibit B. I also reviewed the Guide to the Florida Voter Registration System ("FVRS") dated September 7, 2005, and the registration protocol described on pages 41-46 of that document; a copy of the relevant pages is attached as Exhibit C. I also reviewed the public presentations of Peter Monaghan, the Social Security Administration's Senior Advisor to the Office of Programs, dated February 6, 2006, and February 12, 2007, relating to audits that the Social Security Administration has performed relating to its matching of voter registration information. Copies are attached as Exhibits D and E. I reviewed similar reports of audits of the voter registration information matching process by election officials in New York City. A copy is attached as Exhibit F.

13. Based on my academic and professional experience with record matching, and additional research and analysis I performed, including my review of relevant materials from Washington State, and based on my understanding of Florida's matching protocols, I conclude that the voter registration matching processes used for Florida voters will result in a significant number of "false negatives": records that pertain to the same individual but which are unable to be matched. If Florida effectively conditions registration on a successful "match," a significant number of valid voter registration applications will be rejected as a result. These errors are likely to occur even if the applicants do not make any mistakes or provide any incorrect information on their registration forms. It also is my opinion, based on studies of similar protocols, that the rate of such "false negative" errors will be substantial, and could reach as high as 30% overall. Indeed, the Social Security Administration has reported that its attempts to match voter registration records to its own records have thus far failed 46.2% of the time.

<u>The Process of Record Matching in Florida</u>

14. I understand that in 2006, the Secretary of State of Florida, in conjunction with other State and federal agencies, began to attempt to match certain information provided on new voter registration forms with information stored on other databases. Based on documents that I have reviewed pertaining to Florida's voter registration process, I outline in broad terms below my understanding of the process of record matching as it is performed in Florida.

15. *First,* I understand that citizens fill out a voter registration form by hand with their identifying information. That information includes: name, date of birth, and either a driver's license number (or non-driver's ID card number) or the last four digits of their Social Security number (if the applicant has such a number). A true and correct copy of Florida's registration form available online from the website of Florida's Secretary of State is attached hereto as Exhibit G. Registrants may also supply their identifying information orally to a data entry clerk. The completed forms are then submitted to State or county officials.

16. *Second,* I understand that data entry operators working for the State or county will input the data contained on the voter registration forms into one or more databases that serve as temporary, electronic storage for such new registration records.

17. *Third,* I understand that each new electronic registration record will be submitted to State officials, who will cause certain pieces of information in the registration record to be compared automatically either to the Social Security Administration database or to the State Department of Highway Safety and Motor Vehicles ("HSMV") database. This is done in an attempt to "match" the information contained in the voter registration record to the information contained in the database. For matching with the Social Security Administration database, my understanding is that Florida is using a protocol in which each character of the first name, each character of the last name, each character of the year of birth, each character of the month of birth, and

each character of the last four digits of the applicant's Social Security number, as entered in the voter registration record, must match exactly with the corresponding character or the corresponding field of the Social Security Administration database. For matching with the HSMV database, my understanding is that Florida is using a protocol in which at least the driver's license or non-driver's identification card number and the first four letters of the first name and last name must match exactly with the characters of the corresponding field in the HSMV database.

18. *Fourth,* I understand that if the State finds a "match," the person who filled out the form, if otherwise eligible, will be registered to vote. If the State does not find a "match," the person who filled out the application will not at that point be registered to vote, although I understand that in certain circumstances, there may be some further review, and that elections officials will attempt to correspond with such applicants to try to resolve the problem.

19. Based on this information, it is my opinion that record matching as I understand it will be conducted in Florida will likely result in high numbers of false negatives. That is, I am confident that attempts to match information in new voter registration records to information in other State and federal databases will fail for reasons unrelated to the accuracy of information provided by the applicants or the eligibility of applicants to vote.

Common Errors Related to Record Matching

20. There are several reasons why large databases are prone to errors that make the process of record matching imperfect. The point is a rather basic one, but it has profound consequences when attempting to match individual records in one large database with records in another database: typos, misplaced information, incorrectly transcribed data, and immaterial spelling and punctuation differences in either one or both of the databases may result in two records for the same person not "matching."

21. **Data Submission.** Errors within individual records of large databases may be caused by mistakes in data submission. I am not referring to false information, but mistakes in the form in which the data is submitted. Such mistakes can include minor errors made by individuals filling out forms, such as writing information in one place when the information should be written in another. These immaterial mistakes may appear in the registration record or in the government database being matched – or in both. For example, a person may write her day of birth in a space reserved for the month of birth. If this were to occur in Florida, that person's birth date as entered from her registration form will not match the birth date as recorded on the database with which the State will compare her identifying information.

22. **Data Entry.** Errors within individual records of large databases may also be caused by mistakes in the process of entering the data in the computer. Such errors may occur when an operator strikes an incorrect key, incorrectly hears information

given orally, or incorrectly reads information from a form. For example, a data entry operator may type an "a" when an "o" is written, or type a "d" when a "c" and an "l" are written together. Common data entry errors also include:

- omitting characters (*e.g.*, "JOHN" becomes "JON");
- adding characters (*e.g.*, "OWEN" becomes "OWENS");
- transposing characters (*e.g.*, "SIERRA" becomes "SEIRRA,");
- substituting characters (*e.g.*, "THOMAS" becomes "TH0MAS"); or
- any combination of the above.

23. Other data entry errors may occur when an operator enters information in the wrong field (*e.g.*, inverts day and month in fields provided for the date of birth). Operators also separate compound last names into the "middle name" and "last name" fields or, conversely, combine a middle and last name into a single last name (*e.g.*, "GABRIEL" "GARCÍA" "MÁRQUEZ" becomes "GABRIEL" "GARCÍA MÁRQUEZ"). Such errors include:

- omitting fields (*e.g.*, "MARIE MAUDE" becomes "MARIE");
- adding fields (*e.g.*, "JAMES THOMAS" becomes "JAMES J THOMAS" or "MR JAMES THOMAS" or "CAPT JAMES THOMAS");
- transposing fields (*e.g.*, "JAMES THOMAS" becomes "THOMAS JAMES", or "LU BAO" becomes "BAO LU");
- substituting fields (*e.g.*, "JIMMY THOMAS" becomes "JAMES THOMAS");
- improperly separating fields (*e.g.*, "JEAN-CLAUDE" becomes "JEAN" "CLAUDE");

- improperly combining fields (*e.g.*, "DEBBIE" "WASSERMAN" "SCHULZ" becomes "DEBBIE" "WASSERMAN-SCHULZ"); or
- any combination of the above.

24. I am a member of the Association for Computing Machinery, and I have reviewed relevant portions of the ACM's February 2006 study *Statewide Databases of Registered Voters*, a copy of which is attached as Exhibit H. This study recognizes both that "[m]ost errors in individual database records occur during data entry," and that "[w]hile quality control systems and appropriate supervision of data entry may reduce data entry errors, some errors will inevitably occur. . . . Changes that are primarily entered in other state databases – such as changes in marital status and court approved name changes – also compound the challenge to accuracy." Exh. H at 21.

25. In my extensive experience working with databases containing similar kinds of personal information, the errors described in paragraphs 22 and 23 can be quite common. One reliable study found that the names of 23-37% of the patients in several medical databases were misspelled in at least one database record; a copy of this study is attached as Exhibit I. Another study reported that approximately 26% of records in a Florida social service database included city names with apparent misspellings, including more than 40 different spellings of "Fort Lauderdale"; the study's relevant pages are attached as Exhibit J.

26. If any one or more of the errors described in paragraphs 22 and 23 were to occur in Florida in the registration record itself and/or in the database with which

the record will be matched, the name as entered from the individual's registration record will not exactly match the name as recorded in the database with which the State will compare the individual's registration information.

27. Data entry operators commonly commit errors when they input names, but they also commit many of the same types of errors when they input numbers. Such errors are specifically acknowledged to occur with respect to Social Security Numbers. The leading expert on record matching for the U.S. Bureau of the Census estimates that in one large California employment database, given these types of errors "[o]ver a period of twenty years, the records [associated] with each individual can expect to contain *at least two errors* where the [Social Security Number] has been mis-keyed or transcribed improperly" (emphasis added). A copy of this publication is attached as Exhibit K.

28. **Data Maintenance, Storage, Transfer, and Transformation.** Once a record is created for an individual applicant, the State must maintain, store, transfer and, often, transform the data contained in that record. Federal and State officials must perform similar tasks with respect to data contained in the Social Security Administration and HSMV databases. These processes are also prone to error, for example, when computer viruses cause file corruption; when the data input locally, in Florida's 67 county election management systems, is transferred to the State; and when database fields are added, modified or deleted and, accordingly, data is split, changed, or

consolidated. In my experience, such transfers can lead to unintended changes in the underlying data.

29. The ACM's study, *Statewide Databases of Registered Voters*, also recognizes that glitches can create problems in large databases. As the study states:

> Databases also can be inaccurate or unreliable because of computer viruses, programming errors, and system failures. For example, in 2003 the Maryland Motor Vehicle Administration (MVA) offices were attacked by a computer worm. The worm shut down the MVA's computers and telecommunication systems, cutting them off from all forms of remote communication and disrupting operations in all 23 MVA offices located throughout the state. A second event occurred on January 20, 2004, when the MVA could not process work on the mainframe computer for about an hour after opening. The problem was characterized as a computer glitch.

Exh. H at 24.

30. There is no single standard industry algorithm or process for maintaining, storing, or transforming data; different entities use different processes for these purposes. As noted on page 14 of the *"Social Security Verification" System Specification*, for example, there will be "many different types of computers on the [AAMVA] network, each possibly having a different data-encoding scheme." Different entities using different conventions, or transferring data using different encoding systems may, because of incompatibilities, cause modifications in the data they maintain that will lead to unmatched information.

31. If any of these modifications were to occur in Florida, affecting the registration record itself and/or the database the record will be matched with, the information as entered from the individual's registration record will not exactly match the information as recorded in the database with which the State will compare her identifying information.

32. **System Errors**. Online computer systems intermittently experience system errors or other "down time." The Social Security Administration is not immune to these errors; the *"Social Security Verification" System Specification* describes "program problems, network interface errors, database errors, program aborts, [and] the more common system error[,] when the SSA file is off-line." Exh. B at 17.

33. If such system errors occur when a Florida registration record is submitted, at least during the error period, the information on that record will not be able to be matched with information in the offline database.

34. **Natural Data Inconsistency**. In addition to the errors described above, the process of matching information in different records itself produces false negatives because of superficial discrepancies between those records that do not reflect inaccurate information. For example, names are not truly standardized, nor are they fixed. People adopt nicknames, use shortened names, pick up or drop middle names, take their spouse's names, and/or change the spelling of their transliterated names – and they do so even in formal government documents. In addition, different applicants or different data entry operators (and even the same people on different occasions) may transliterate

non-English characters in different ways. Thus, two records for the same person may show different names, like a maiden name or married name. Similarly, data entry operators often use default assumptions to fill in missing information (*e.g.*, choosing the first of the month when no day of the month is given).

35. Common examples of natural data inconsistencies that may cause false negatives include:

- nicknames (*e.g.*, "ELIZABETH" versus "LIZ");
- maiden names (*e.g.*, "REBECCA JONES" versus "REBECCA SMITH");
- husband's names (*e.g.*, "MRS. JOHN SMITH" versus "MRS. REBECCA SMITH");
- punctuation (*e.g.*, "O'BRIEN" versus "O BRIEN" or "OBRIEN")
- compound last names (*e.g.*, "GABRIEL" "GARCÍA MÁRQUEZ" versus "GABRIEL" "GARCÍA" "MÁRQUEZ");
- first or middle initials (*e.g.*, "F. SCOTT FITZGERALD" versus "FRANCIS S. FITZGERALD");
- name change due to religious conversion (*e.g.*, "MUHAMMAD ALI" versus "CASSIUS CLAY"); or
- any combination of the above.

36. Similar data inconsistencies arise when confronting names common within certain ethnic communities:

- immigrants adopting "Americanized" names, for all purposes or just some purposes (*e.g.*, "GRACE KIM" versus "HYUN KIM");
- name change due to different status in the community (*e.g.*, in Burmese, "MAUNG TIN" (for younger men) versus "U TIN" (for married men));

- mistaking a title for a first name (*e.g.*, "MAUNG TIN" versus "TIN");
- transliterated names or diacriticals (*e.g.*, "MUHAMMAD" with "MOHAMMED," or "SCHRÖDER" with "SCHRODER" or "SCHROEDER";
- alternative spellings (*e.g.*, "DE LA CRUZ" with "DELACRUZ"); or
- any combination of the above.

37. In my extensive experience working with databases containing similar kinds of personal information, the discrepancies described in paragraphs 35 and 36 can be quite common. If any of the natural discrepancies described in paragraphs 35 and 36 were to occur in Florida, creating immaterial differences between the registration record itself and the database the record will be matched with, the information as entered from the individual's registration record will not exactly match the information as recorded on the database with which the State will compare her identifying information.

38. My wife's name provides an example of how such trivial differences can cause record-matching problems. My wife usually represents herself as "Sarah C. Borthwick," and signs personal checks that way. But she is registered to vote in New York as "Sarah E. Caguiat Borthwick." Her New York driver's license shows her name as "Caguiat-Borthwick, S". And she appears in Social Security Administration records as "Sarah E. Caguiat." If she attempted to register to vote in Florida as "Sarah Borthwick," the information in her application would likely not match information in either the driver's license or Social Security databases.

Errors Common in Particular Communities

39. Certain errors contributing to difficulties in record-matching are more prevalent among particular racial and ethnic communities. For example, in Hispanic or Latino communities, it is common to use either maternal or paternal last names, or both. These names are often supplied inconsistently by the individual or entered inconsistently by the data entry operator such that the "middle name" and "last name" fields in the resulting record are inverted, separated, or combined. For example, "José Luis Rodriguez Zapatero" might have "Zapatero," "Rodriguez," or "Rodriguez Zapatero" entered as his last name.

40. In African-American communities, names derived through modification of more traditional spellings are more common than in other racial or ethnic communities. These names are more likely to be misspelled in data entry. For example, one study reports that "Jazmine," "Jasmin," and "Jazmin" are all girls' names much more common among African-Americans. A copy of this study is attached as Exhibit L. These names may all be misspelled as "Jasmine" in data entry, thus creating errors when an exact character-by-character match protocol is applied. Moreover, names that are unique to a particular individual are also more common in African-American communities. The same study cited above, for example, found that African Americans in California are six times more likely to have a unique name than are Caucasians. These names may be unfamiliar to data entry personnel (of any race or ethnicity), and are more likely to be misspelled in a database.

41. Transposition of the "first" name and "last" name is more common with regard to individuals of Chinese descent, many of whom present their family name first and their given name second, contrary to the usual American practice. A data entry operator might not know which name in "Lu Bao" is the first name and which is the second, and enter it based on any variety of conventions, such as assuming that the first name listed is the given name. If Mr. Lu's name is transposed in one record, that name will not match exactly to the other record. In my experience, individuals of Chinese descent also frequently adopt names considered to be common "Western names," but use these "Western names" inconsistently in official records. The following example illustrates both phenomena: a Chinese woman named "Wang Fei" might inconsistently put her first name before her last name (*i.e.*, "Fei Wang"); use a Western form of her first name (*e.g.*, "Faye Wang") or a Western name not derived from her first name (*e.g.*, "Grace Wang"); and/or use a Western form for both her first and last names (*e.g.*, "Faye Wong").

42. In communities that do not use the Roman alphabet in their primary language, such as East Asian, Middle Eastern, Hellenic, and Slavic communities – or communities using diacritical marks not found in English, such as the umlaut or tilde – inconsistent transliterations are common. Arabic names like "Mohammed," for example, are transcribed differently depending on the country of origin. Three variants include "Muhammad," "Mohamed," and "Mahomet."

43. The transposition of the date and month of birth is more commonly found with regard to recent immigrants, who may be accustomed to presenting dates in a day-month-year convention, which is commonly used in Europe, Africa, the Middle East, and Asia. Thus, someone whose date of birth is May 6, 1980 might input her name as "6/5/1980," and since that is a valid date under the American month-day-year convention, her record will reflect that her birth date is June 5, 1980.

44. Mismatched surnames due to a maiden name or married name are more common, of course, with regard to women. Thus, to use my wife as an example again, whether she registers to vote using a compound last name without a hyphen, a compound last name with a hyphen, or my last name, the information entered from her registration record will not exactly match the information in the Social Security Administration's database, where she appears under her maiden name.

The Impact of Errors and Non-Standardized Data on Record Matching

45. Attempts to match records using exact, character-by-character matching – referred to in the industry as "deterministic" matching – are highly sensitive to all of the errors and discrepancies described above. The failure to match information because of such errors and discrepancies would result in false negatives – *i.e.*, the failure to match database entries that in fact belong to the same individual.

46. Data from several reliable studies show that, in similar circumstances, false negative rates generated by deterministic matching protocols can

reasonably be expected in the range of 20-30%. For example, in one reliable study, the U.S. Bureau of the Census suggested that using a deterministic match on census data would have resulted in a false negative rate of about 25%; a copy of this study is attached as Exhibit M. Another reliable health care study found a false negative rate of about 22% using a deterministic protocol. Exh. I at 503-04. And yet another reliable study found that a deterministic protocol missed 17-30% of records belonging to the same individuals; a copy is attached as Exhibit N.

47. Deterministic matching protocols have shown similar failure rates in practice. For example, through mid-June of 2006, I understand that Washington State compared information on new voter registration forms to information maintained in motor vehicle and Social Security records through a deterministic matching protocol. Through this process, no match was found for 16% of new forms statewide, and 30% of the records submitted in the state's most populous county (King County) were unable to be matched. These "no match" rates are consistent with the false negative rates in the studies above. That is, it would be consistent with these accounts to find that 30% of the forms submitted in King County failed to match information in the motor vehicles or Social Security databases, but actually represented individuals accounted for in the motor vehicles or Social Security databases.

48. As noted above, I have reviewed the *Social Security Verification System Specification* prepared by the American Association of Motor Vehicle Administrators in August 2004, and, in particular, the HAVV transaction described on

pages 26 and 27 of that document. Exh. B at 26-27. As described in that document, the HAVV transaction uses a deterministic match protocol in which a system will attempt to match the last name, first name, month of birth, year of birth, and last four digits of the Social Security number of a target record to the same elements of records in the Social Security Administration database. A successful match will be reported only when each character of each such field in the target record matches precisely each character of each corresponding field in the Social Security Administration database. Pursuant to the same document, I understand that an unsuccessful match will be coded as a "system error," "invalid input data," or "no match found"; no more specific information will be returned to the state indicating why a match could not be found, more precisely locating the source of the error.

49. The HAVV protocol is not designed to account for, and will not readily account for, the errors described above; the protocol for matching with the HSMV database is similarly susceptible to the same errors described above. Moreover, particularly but not exclusively in the HAVV protocol, the requirement that *multiple* fields exactly match compounds the error rate expected for an exact match on any individual field.

50. Page 10 of the above-noted 2006 presentation of Mr. Monaghan, the Social Security Administration's Director of Information Exchange, states that no match was found in 28.5% of 143,000 queries submitted in the period before his presentation. Exh. D at 10. Page 9 of Mr. Monaghan's 2007 presentation reports that of

the 2.6 million queries submitted by 2007, no match was found in 46.2% of the queries. Exh. E at 9.

51. Assuming that the Social Security Administration used the deterministic HAVV transaction to seek matches for voter registration records, the reported 28.5% "no match" rate described in paragraph 49 is consistent with the rate of false negatives found in other published accounts of deterministic matching. The 46.2% "no match" rate reported in 2007 is greater than the false negative rate reported in many other accounts, but given the acknowledged errors in the Social Security database, the multiple points at which error may be introduced in the process of entering voter registration data, and the HAVV protocol's use of deterministic matches on multiple fields in combination, it is not unreasonable to believe that the 46.2% "no match" rate in fact represents false negatives. That is, it would be consistent with these accounts to find that 46.2% of the queries submitted to the Social Security Administration failed to match information in the Social Security database, but actually represent individuals accounted for in the Social Security database.

52. Moreover, in my opinion, some voters will probably not be provided an effective opportunity to resolve a "false negative." For example, although I understand that Florida election officials may attempt to correspond with unmatched registrants, data entry errors impacting name and address will probably prevent some correspondence from reaching its intended target.

53. I have reviewed the May 15, 2003 appraisal of Virchow Krause & Company, a prominent Midwest accounting and consulting firm retained by the Wisconsin State Elections Board to evaluate project proposals for Wisconsin's statewide voter registration database. A copy of the relevant portion of this appraisal is attached as Exhibit O. I agree with the appraisal's conclusions regarding the difficulty and likely effect of matching in this context:

> Name matching and validation issues are very complex (e.g., matching Margie L. Smith with Margaret Smith), and are made even more complex when aliases and name changes are considered. . . . Even a 1% error rate on an interface validating names, driver license numbers, etc. could generate tens of thousands of bad matches in an error log, well beyond any ability for the [state, county, or local] users to manually verify the errors. . . . [¶] All vendors suggested that incomplete or unmatched records be ignored, because the time to resolve, cost to resolve, and potential for error and disenfranchisement was too high.

Exh. O at 20.

54. In sum, the matching systems that I understand Florida is using, described in paragraphs 17 and 48, are prone to many errors – especially false negatives. In my opinion, such systems would generate failed matches for individuals who are, in fact, legitimately represented in the target database. When comparing two data sources of significant size – as Florida is doing here – records representing the same individual would fail to match even if the Secretary of State used protocols representing the best available technology. If matching is effectively a prerequisite to registration, the use of any match process will result in eligible voters being denied the right to vote.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct, and that this Declaration was executed on September 15, 2007 in Palo Alto, California.

ANDREW BORTHWICK

CERTIFICATE OF SERVICE

Undersigned counsel herby certifies that a copy of the foregoing *Declaration* was served via HAND DELIVERY this 17th of September , 2007 upon the following:

Kurt Browning, Defendant
Secretary of State
Florida Department of State
R.A. Gray Building
500 South Bronough Street
Tallahassee, FL 32399-0250

GREENBERG TRAURIG, P.A.

GLENN T. BURHANS, JR.
FLA. BAR NO. 605867
101 EAST COLLEGE AVENUE
TALLAHASSEE, FLORIDA 32301
TEL. (850) 222-6891
FAX (850) 681-0207

ANDREW BORTHWICK, Ph.D.
1453 Kings Lane
Palo Alto, CA 94303

Summary: Computer Science Ph.D. with deep experience in natural language processing, machine learning, and approximate record matching. Business skills include 8 years' experience founding and growing a technology startup.

EXPERIENCE

SPOCK NETWORKS, Redwood City, CA — May, 2007 – Present

Principal Scientist

- Play key scientific role in rapidly-growing early-stage people search engine:
  - Research and develop processes to extract biographical information about people from diverse public websites for the purpose of generating profiles for the Spock website
  - Research and develop process to link profiles of people so that the Spock website will have only one profile for each real-world individual

CHOICEMAKER TECHNOLOGIES, INC., New York, NY

CTO — August 2006 – May, 2007

- Conceived, coordinated, and implemented a wide range of R&D projects:
  - Built probabilistic models using maximum entropy machine learning technology
  - Researched new approaches to approximate string matching
  - Researched and coded new algorithms for enhancing the functionality and robustness of ChoiceMaker's record matching processes

CEO — July 1998-August 2006

Business Accomplishments

- Founded the company to commercialize a machine learning approach to the problem of approximate record matching based on my Ph.D. research. For instance, our software can determine that records for "Andrew Borthwick" and "Andy Barthwick" represent the same person.
- Coordinated the successful deployment of systems with the Centers for Disease Control, New York City Department of Health, NY, IA, KS, MO, NE, NM, PA, SC, and WY State Education Departments, Regulatory DataCorp, Phoenix-ESI, and other clients.
- Hands-on work in marketing, including helping to develop marketing collaterals and the creation of design and content for the website.
- Actively involved in sales, including technical sales support and contract negotiation.
- Managed the firm's finances. Carefully monitored cash flow to enable firm to grow on minimal equity investments. Secured equity, debt, and government grant financing for the firm.
- Led all HR activities, including hiring, evaluation, and dismissal.

Technical Accomplishments

- Wrote papers and gave numerous presentations (including one invited talk) on record matching to help promote the firm.
- Principal investigator of three National Science Foundation Small Business Innovation Research grants. Grants provided $1M for research into machine learning approaches to approximate record matching, high speed record matching, and the design of the ClueMaker programming language for record matching.
- Awarded US Patent 6,523,019 for machine learning approach to record matching. Also awarded U.K. patent for same work.
- Principal designer of high speed algorithm for real time "blocking", the first stage of matching in which a database is searched for possible candidate matches. Real-time algorithm received U.S. Patent #7,152,060. Algorithm takes a completely different approach to this problem from the main line of published research.
- Managed the ChoiceMaker product. Conceived and implemented a strategy which led to the construction of the "ChoiceMaker Developer" IDE for the creation and testing of record matching models and the "ChoiceMaker Server" system which our clients deploy in production.
- Personally coded ChoiceMaker version 1.0 in C++ using a flexible object oriented framework to describe the features, histories, and futures making up the system. Made heavy use of the STL. Used Perl for scripting the core modules. Deployed the system to the New York City Department of Health for use in production.
- Built a maximum entropy "estimator" in C++ for computing the weight to be used for features in a maximum entropy model. Estimator was used in ChoiceMaker 1.0, CM 2.0, and was licensed to two Japanese research institutions.

EXHIBIT
A

MORGAN STANLEY, New York, NY 1993-2002

Systems Consultant

- Working only one day per week, was critical designer and maintainer of the Information Services Allocation Model (ISAM). ISAM is a highly sophisticated system which solves matrix algebra equations describing the circular movement of money within IT in order to equitably allocate over $1 billion in annual IT costs to the rest of the firm.
- Designed most of the major and minor upgrades for ISAM, which were implemented by a team of three full-time programmers. The system grew greatly in functionality and importance over nine years.
- Duties included making presentations to explain the functionality of the existing system, clarifying user requirements, designing enhancements, answering user questions, fixing bugs, and Y2K.
- One of a small number of consultants put on a *must-retain* list during a switchover of consulting agencies.

IBM WATSON LABORATORY, Yorktown Heights, NY Summer 1997

Summer Intern

- Researched maximum entropy language modeling for a voice-operated air travel reservation system.

MORGAN STANLEY, New York, NY 1988-1993

Programmer/Team Leader/Business Analyst

- Designed and managed the project which built ISAM.
- Supervised a team of four while working closely with the users in IT Finance to take the project from a few pages of notes, diagrams, and equations to a finished product.
- Personally coded the mathematical heart of the system in APL.

EDUCATION

Courant Institute of Mathematical Sciences, New York University, New York, NY September 1999

- M.S. and Ph.D., Computer Science
- Thesis title: "A Maximum Entropy Approach to Named Entity Recognition"
- Specialized in Machine Learning and Natural Language Processing
- Advisor: Prof. Ralph Grishman
- 3.74 GPA
- Invented and constructed a system to detect proper names ("named entities") in newspaper text. Built the first system to combine the output of multiple hand-coded information extraction systems within a maximum entropy framework. System placed fourth out of twelve in a DOD evaluation after only four person-months of effort. Rapidly ported the system to Japanese and performed well in a Japanese named entity evaluation, where it was the only system written by a non-speaker of Japanese.

Oberlin College, Oberlin, OH May 1988

- Bachelor of Arts, History
- 3.64 GPA
- Phi Beta Kappa
- Comfort Starr Prize for Excellence in History

PUBLICATIONS: Data quality and record matching

Patents

- Andrew Borthwick, Martin Buechi, and Arthur Goldberg. *Automated Database Blocking and Record Matching.* U.S. Patent #7,152,060. Filed April 11, 2003. Awarded December 19, 2006.
- Andrew Borthwick. *A Probabilistic Record Linkage Model Derived from Training Data.* U.S. Patent #6,523,019. Filed Oct. 28, 1999. Awarded February 18, 2003. Also awarded U.K. Patent #2,371,901.
- Co-inventor of one other pending patent.

Papers

- Andrew Borthwick. *The ChoiceMaker 2 Record Matching System.* ChoiceMaker Technologies white paper. November, 2004.
- Vikki Papadouka, Paul Schaeffer, Amy Metroka, Andrew Borthwick, Parisa Taranifar, Jessica Leighton, Angel Aponte, Ruron Liao, Alexandra Ternier, Stephen Friedman, and Noam Arzt. *Integrating the New York City Immunization Registry and the Childhood Blood Lead Registry.* Peer reviewed paper. Journal of Public Health Management and Practice. November, 2004.

- Andrew Borthwick and Maggie Soffer. *Business Requirements of a Record Matching System.* Peer reviewed paper. Massachusetts Institute of Technology's Ninth International Conference on Information Quality (MIT ICIQ), Cambridge, MA. September 7, 2004.
- Martin Buechi, Andrew Borthwick, Adam Winkel, and Arthur Goldberg. *ClueMaker: A Language for Approximate Record Matching.* Peer reviewed paper. Massachusetts Institute of Technology's Eighth International Conference on Information Quality (MIT ICIQ), Cambridge, MA. August 27, 2003.
- Andrew Borthwick, Martin Buechi, and Arthur Goldberg. *Key Concepts in the ChoiceMaker 2 Record Matching System.* Peer reviewed paper. First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, in conjunction with the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC. July 17, 2003.

Miscellaneous

- Andrew Borthwick. Expert witness testimony on record matching issues. *Washington Association of Churches, et. al. v. Reed.* May 24, 2006. Prepared in association with the Brennan Center for Social Justice and pro bono attorneys from Paul Weis. Resulted in an injunction brought against Washington State to block inaccurate record matching from being used on voter registration rolls. See www.choicemaker.com/content/news/press/20061009_testimony.php3 for details and www.brennancenter.org/dynamic/subpages/download_file_36559.pdf for my testimony.
- Andrew Borthwick. *When Accuracy Counts.* Podcast interview with Claudia Imhoff of the Data Warehouse Institute. May, 2006.
- Andrew Borthwick. *The Design and Testing of a Record Matching System.* Slides and abstract. 17th Information Quality Conference. Houston, Texas. September 21, 2005.
- Andrew Borthwick. *Record Linkage Industrial Trends.* Invited talk. First workshop on Data Cleaning, Record Linkage, and Object Consolidation in conjunction with the ACM SIG KDD's Ninth International Conference on Knowledge Discovery and Data mining. Washington, D.C., August 27, 2003.
- Andrew Borthwick and Deborah Walker. *Applications of Record Matching Techniques for a Lead-Immunization Registry Integration Project.* "35th National Immunization Conference", Slides and abstract, Atlanta, Georgia, May 2001.

PUBLICATIONS: Computational Linguistics

Links to all of the below can be found at scholar.google.com

- Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition.* Ph.D. thesis, New York University, New York, New York, September 1999.
- Andrew Borthwick. *A Japanese Named Entity Recognizer Constructed by a Non-Speaker of Japanese.* "Proceedings of the IREX Workshop", Tokyo, Japan, August 1999.
- Andrew Bothwick, John Sterling, Eugene Agichtein, and Ralph Grishman. *Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition.* "Proceedings of the Sixth Workshop on Very Large Corpora", August 1998.
- Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. *NYU: Description of the MENE Named Entity System as used in MUC-7.* "Proceedings of the Seventh Message Understanding Conference (MUC-7)", Fairfax, Virginia, April 1998.

TECHNICAL EXPERTISE

Operating Systems

- Windows, Unix/Linux, MVS (IBM Mainframe)

Technologies

- Data quality, record matching (a.k.a. "entity resolution", "deduplication", "record linkage"), high speed processing of large databases, computational linguistics, algorithms, software architecture, compilers

Programming Languages

- Java, Python, ClueMaker® (Co-Inventor of proprietary language), C++ (including Standard Template Library), Perl, C, Adabas/Natural, APL

Software Packages

- ChoiceMaker 2, Eclipse, MS Project, Visio, QuickBooks, MS Office

Foreign Languages

- Reading knowledge of French, some German

AAMVA

American Association of
Motor Vehicle Administrators

Social Security Verification (SSV)

System Specification

Release 2.0.0

August 2004

4301 Wilson Boulevard, Suite 400 • Arlington, VA 22203 • (703) 522-1300 • Fax (703) 522-1553

EXHIBIT

B

Table Of Contents

1 INTRODUCTION

1.1 Document Objective

The purpose of this document is to describe the data flows and transactions in the Social Security Verification application. The document is written for the Social Security Administration (SSA) and Jurisdictions who must develop Social Security On-line Verification (SSOLV) and Help America Vote Act Verification (HAVV) systems. The document contains the information necessary for a software application development team to:

- Write an implementation plan;
- Determine requirements for their application, based on nationwide requirements;
- Construct a framework for the design of their system implementation.

Because the requirements are written at a high level, a detailed implementation specification should be produced by the developers to describe how the system will be implemented in their environment.

The original release of this document focused on SSOLV implementation at SSA. This release also contains information for users who make SSOLV inquiries to SSA and a full description of the HAVV transaction.

1.2 Getting Help (1-888-AAMVA-80)

Questions regarding this document or the application itself should be directed to:

Operations Department:

Hours:	8:00 a.m.- 6:00 p.m. Eastern Time
Telephone:	1-888-AAMVA-80
Fax:	(703) 522-1553
Address:	AAMVA, Inc. 4301 Wilson Boulevard, Suite 400 Arlington, Virginia 22203
Website:	www.aamva.org
e-mail:	opsdept@aamva.org

2 APPLICATION DESCRIPTION

The Social Security Verification application (SSV) consists of two transactions: the Social Security On-line Verification (SSOLV) and the Help America Vote Verification (HAVV). Each transaction contains two messages: an inquiry message sent to the Social Security Administration (SSA) and a response message returned by the SSA

2.1 Social Security On-line Verification (SSOLV)

Driver licenses and identification cards issued by US Motor Vehicle Agencies (MVA's) have become the U.S. standard for identification. In order to curb the fraudulent issuance of driver license and identification cards, the MVA's carefully review documentation that is presented to them to verify the identity of the individual. The Social Security Card is one form of identification that is reviewed. In most jurisdictions, the Social Security Number (SSN) is also used as the standard to uniquely identify individuals on the licensing records.

To minimize the fraudulent issuance of the driver license or identification card, the MVA's need a way to verify the information contained on the card is valid. The SSN verification needs to be performed on-line while the applicant is still at the MVA counter but prior to the issuance of the driver license or identification card.

The Social Security On-line Verification (SSOLV) transaction has been developed to allow authorized MVA's to have on-line access to SSA for SSN verification. Using this transaction, a MVA electronically sends SSA a person's name, date of birth (DOB) and SSN. SSA then compares this data to what is on its Master File. SSA will then respond back to the inquiring MVA, indicating how much of the MVA-submitted data matched against the SSA file.

2.2 Help America Vote Verification (HAVV)

Section 303 of Public Law 107-252 (Help America Vote Act of 2002) requires States and localities to develop centralized, computerized voter databases and to verify voter registration information. Individuals registering to vote must provide their driver's license number to the State election agency. If the registrant has no driver's license, they must supply the last four digits of their SSN. The statute requires that the chief State election official and the officials responsible for the State motor vehicle authorities to enter into agreements to match voter registration information with MVA information. The statute further requires the MVA officials and the Commissioner of Social Security to reach agreements for the purpose of verifying name, date of birth, the last four digits of the SSN, and any information recorded in SSA's records about the death of an individual.

The Help America Vote Verification (HAVV) transaction allows a MVA to submit an inquiry to SSA. The SSA verifies the information and responds back to the MVA with the results.

The Primary Address is the 2-character code for jurisdictions and AAMVA processing sites (normally the postal abbreviation for jurisdictions), and the Interface Code is used to distinguish between multiple systems at a single site. For example, in some states the driver licensing and vehicle registration systems are operated on different physical machines. The Interface Codes would be different for each.

AAMVA manages the overall AMIE UserID system, and is therefore responsible for assigning all values as necessary for the GAP Code.

The User Extension field of the Primary User description can be used at the discretion of the users, within the normal parameters for AMIE Messages (See the section on General Rules for AMIE Message Composition). This field is frequently used to identify a particular workstation that originated the message and therefore should receive the response. Other uses are possible depending on the needs of the users. Usage of this field should be limited to the Transaction Originator because it is the pass-through field.

4.4.3 General Rules for AMIE Message Composition

Data in an AMIE message may consist of any printable character. This means that non-printable bytes are not allowed in any AMIE message. This limitation has been imposed due to the architecture of the AAMVAnet network, which consists of many different types of computers on the network, each possibly having a different data-encoding scheme.

For example, the AT&T NETWORK SERVICES and its mainframes store character data in EBCDIC, while Unisys, Bull, and most other computer types store character data in ASCII. Translation between these code sets is performed as part of the network transmission to or from an ASCII based machine. The translation occurs by replacing a bit pattern from one code set with the corresponding bit pattern from the other code set. As the translation is performed to each byte of data traveling on the data path without regard to the content of the data, non-printable data would be corrupted when the bit patterns were replaced as if the byte contained character data.

Translation adulteration aside, each different machine type stores computational numeric data in a format native to the processor. Assuming numeric data could move between AAMVA nodes without adulteration, the data would probably be unusable by the destination node unless the origination and destination nodes happen to be compatible machine types.

For example, floating point decimal data on an AT&T NETWORK SERVICES mainframe is stored in a specific pattern of bits within two, four, or eight bytes, depending on the resolution required. Elements such as the exponent and mantissa are assigned to certain bits and are represented in defined ways. The same number on a VAX machine is stored with a different bit pattern, different exponent bases, and different byte order. Moving a floating-point decimal data item from an AT&T NETWORK SERVICES platform to a VAX would not yield usable data on the VAX. The reverse is also true.

Eventually, exceptions to this rule may be required to allow movement of complex data in an efficient manner, possibly using encoding and compression schemes. At that time specific exceptions will be defined and will be documented to an extent that potentially affected users will be aware of their limitations. However, the general rule will still apply to all other messages that may be sent between nodes running on different computer types.

To ensure only printable bytes exist in a message, you must initialize all unused areas of each block with spaces. This ensures that un-addressable areas, such as the reserved bytes at the end of most, contain valid AMIE data. The unused fields should also be initialized to spaces regardless of the data type of the field. For example, a date field is normally numeric, yet if the field is not a valid part of the message being built, the field should contain spaces rather than zeroes. Do not initialize AMIE blocks or fields to LOW-VALUES or HIGH-VALUES, as these are binary zeroes or ones, respectively, and do not represent printable data.

All application data elements must contain printable characters that can be used in both ASCII and common versions of EBCDIC. The printable characters are:

```
space
a to z
A to Z
0 to 9
! " # $ % & ' ( ) * + , - . / : ; < = > ? @
```

Other characters are not printable in ASCII and US-EBCDIC, so should be excluded. The user will need to determine if the non-printable characters will be omitted or if they will substitute an other character. The recommendation for the Spanish 'Ñ' and 'ñ', is to convert the character to 'N' and 'n' before sending the data.

4.4.4 Application Text Blocks

For this system, the text block pool of an AMIE message contains the following block types:

- Message Exchange Control block (02/2). One Message Exchange Control (MEC) block will be present on each message. See the Message Exchange Control Block section for details.
- Business Application blocks (09/1, 10/1).
- Return-as-received blocks (98/3). Zero to five return-as-received blocks may be used, and they are used by the transaction originator.
- Error blocks (99/1). Zero to five error blocks are used, depending on the number of errors detected. See the Error Handling Section for details.

Because the blocks are sent in the Type/Sub-type number order, the text blocks will be sent in the order shown above.

Most blocks are used once within a message. However, instances exist where an AMIE text block is used multiple times within a message. These multiple repetitions exist when:

- A field is too long to fit in a single 61-byte block. A 108-byte address is transmitted in two AMIE text blocks. The first 61 bytes are sent in the first block and the final 47 bytes are sent in the second block.
- The application data is needed multiple times, where a single occurrence of the data will fit onto one block. The number of blocks will correspond to the number of occurrences of the data. The data is needed multiple time times; however, the total length of the data to be repeated exceeds one block. In these situations, the number of AMIE text blocks used is the product of the number of blocks used to hold a single occurrence, times the number of occurrences.

To be unique the Text Block Key will use an incremented line number to distinguish between the multiple occurrences of block types and maintain the sort sequence.

4.4.5 Message Format of Fields

All dates sent in the application specific blocks of the messages are passed as eight character fields in 'ccyymmdd' form, (e.g., '19951231'). All numbers sent in a message are passed in an unpacked form with leading zeros (e.g., a field with 6 integer digits with a value of '1,234', is transmitted as '001234', in an alpha numeric field).

For elements that require specific values (such as codes), the fields transmitted must contain the standard values, as defined in the data dictionary.

4.5 Error Handling Specifications

The error handling procedure describes a convention by which every message error will be processed, both by the entity that detected the error and the entity that originated the message. The errors can be categorized as follows:

- network errors;
- system errors, such as program aborts, files off line, or similar conditions;
- processing errors which are caused by faulty application data in the message

When an error is detected, the message that encountered or contained the error is returned to the sender. There are several flags and fields in the message structure that can convey information regarding errors or unusual circumstances. Depending on the severity of the problem, different combinations of the error flags/fields are used. Information can be found in the following areas:

GNCBER - NCB ERROR CODE
Set to 'Y' (yes)

GNETST – NETWORK STATUS
Set to a value other than zero, that describes the error.

GAPPST - APPLICATION STATUS

Set to a value other than space or zero, that describes the error.

GERUEC – UNI ERROR CODE or
GERCDO – ERROR CODE

Set to a value other than space, that describes an error.

GERMSO - ERROR MESSAGE DESCRIPTION

A 54 character text field containing the description of the error.

4.5.1 Network Errors

Network errors occur when the origination or destination entity drops from the network or the network itself encounters a failure. There are established availability requirements that minimize occurrences of this nature, but occasionally a failure occurs.

When the originating entity is not connected or the network is completely down, the error is normally detectable and the message can be set-aside for later transmission. The Unified Network Interface (UNI) provides this service.

If the destination node is down, the network (NCS) will return the message to the originator with an indication of the error (NCB error code = 'U' for Undeliverable) and the message can be set aside for later transmission. If the destination application is down, UNI can detect the error, notify the originator, and set aside the message for later transmission.

4.5.2 System Errors

In this application system errors may be reported in one of two ways:

- Generic system errors
- SSA file off-line

A generic system error is an error with the system itself, such as program problems, network interface errors, database errors, program aborts, etc. To the extent possible, message recipients should try to detect these conditions and return the original message with the appropriate indicators to inform the originator of the problem (NCB error code = 'Y', processing status = '01', error block attached indicating the error and application status set to appropriate code, if applicable).

The more common system error occurs when the SSA file is off-line. In this instance, the SSA application will return the SSA Verification Response (HS) message with the SSA Verification Response Code set to '9'. Other system errors detected within the SSA application will also be reported with the Response Code set to '9' on the Verification Response message.

4.5.3 Processing Errors

The SSA will not edit data received in the incoming Verification Request (SS) nor will it return corrected information. Therefore, the only error a SOI should encounter would be that of a network or system error.

4.6 Application Layer Network Interface Software

The Application Layer Network Interface Software (ALNIS) is generically defined as a software application residing on the host computer. The main function is the translation between the AMIE message structure and a data element and the message structure used by the application. The application data structure is provided in COBOL and C formats. It also provides a variety of other application interface support features. The interface between the application and the ANLIS is usually platform dependent. An example of ALNIS software is AAMVA's Unified Network Interface (UNI) software package.

4.6.1 AAMVA's Unified Network Interface (UNI)

Unified Network Interface (UNI) provides critical services for jurisdictions' applications. The UNI was developed by AAMVA for its customers running applications requiring data transfer in the AAMVAnet Message Interchange Envelope (AMIE) electronic data interchange (EDI) format. Although using AAMVA's network interface tool is not a requirement, most users will choose to implement the system using the Unified Network Interface (UNI). UNI has several valuable functions available to assist users (such as message control, routing validation, logging, audit trails, and message grouping). A jurisdiction's network interface team needs to understand UNI's functions to avoid duplicating those functions within the application.

The purpose of this section is to supplement the UNI documentation by calling attention to several UNI features that have been found particularly useful. Although they are documented in the UNI Application Developer's Reference, we have included a brief synopsis here along with suggested settings, where applicable.

4.6.2 Message Retry

AAMVA recommends that users configure the parameter list of all on-line update messages to attempt up to three retries in the event the messages are undeliverable. When set, UNI retry is performed automatically. Users should keep in mind that automatic retry may not be appropriate for messages where the state prefers to control retries either manually or programmatically through the application (as may be the case with inquiry messages).

The PARM-CNT-RETRY-MAX field in the UNI parameter list controls the maximum number of times that UNI will attempt to send an outbound message to its destination. This is a 1-digit numeric field, so valid values range from '0' to '9'.

If the number of retries is set to '0' and the outbound message is returned as undeliverable, UNI will not retry the message. If the number of retries is set to a non-zero value, UNI will hold the message in its undeliverable message file until such time as UNI determines that the destination's node or application is again available. UNI actively checks the status of retry destinations and does not attempt a retry until a positive status is attained. UNI checks the status of all other nodes on the network by issuing IN messages at regular intervals and interrogating the RN responses. The default interval is 20 minutes, but this is configurable. UNI will attempt to resend until it has exhausted the maximum number of retries designated.

4.6.3 Hard Manual Down

A hard manual down causes UNI to treat a destination node as though it were down even when it is not. This can be used, for example, when a state must store on-line transactions while its load file is being processed. Issuing a hard manual down on the destination node causes on-line transactions to that node to go to the message pending process given message retry is configured. Transactions will continue to queue up in message pending until the hard manual down is manually removed. As stated earlier, it is very important to pace messages being released from message pending.

Hard manual downs are issued from the UTT200 Network/Application Status screen by adding the site ID of the destination to be downed to the application status list. First, enter an action code of 'A', the network ID of the destination, and an application code of '11'. The down reason will be set to 'soft manual' by the system. To change the down reason to 'hard manual', enter an action code of 'M'. The 'M' action code toggles between a soft and a hard manual down. To delete a hard manual down, enter an action code of 'D'. Message pending will initiate release of messages at the next IN/RN interval.

Before issuing a hard manual down, states should estimate the amount of space needed to store the message pending file. Steps should be taken to ensure that enough space will be available to hold the estimated number of pending messages.

4.6.4 Message Locator

When a transaction is initiated, UNI generates a unique identifier for the message called a message locator. UNI uses the message locator to match messages with their responses. When contacting the AAMVA Operations help desk for support, it is important that you provide the message locator. The message locator provides a means for the AAMVA Operations help desk to find the specific message or messages causing the problem.

The message locator is found in the first 26 bytes of the MEC block. It is comprised of a date/time/sequence number along with the message type.

A sample message locator and its components are shown below:

```
000502132312001    1UNISS
```

where:

'000502' is the date
'132312' is the time
'0001' is the sequence number
' ' is a constant
'1' is the occurrence of the destination in the PARM-DESC-TABLE-DEST of the parameter list
'UNI' is a constant
'SS' is the message type

4.6.5 Call List

UNI provides a parameter list and call list to interface between the jurisdiction's application and the network. The call list data is converted to the AMIE structure before it is sent to network and vice-versa. The parameter list provides a means for matching response messages to inquiry messages, routing messages and store and forward features. The parameter and call lists use a flat file format which make it easy for developers to address the elements.

4.6.6 Driver Call List Layout

In the Driver Call List, there is a record type indicator (CLMF-DESC-RECORD-TYPE) that is populated by UNI when a message is received. This indicator is used to identify how much of the variable length Call List is being used. In this application UNI sets the indicator to "R", "L" or "S". When the indicator contains a:
"L" the type of record is a long record. In this situation the address is included.
"S" the type of record is a short record. In this situation no address is included.
"R" the type of record is a return as received.
So before addressing elements residing in the extended part of the call list, check the record type indicator to ensure a long call list has been delivered.

4.6.7 UNI Platforms Supported

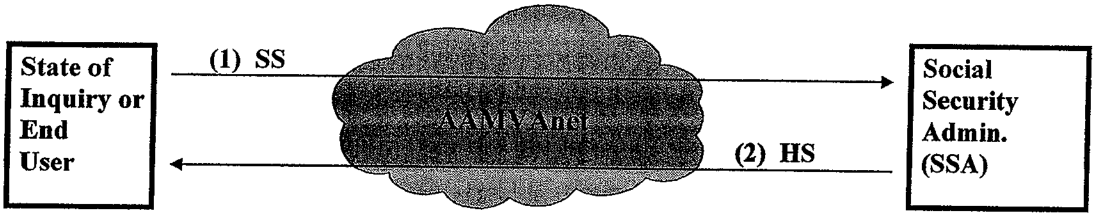
AAMVA's web site (www.aamva.org) has a complete up-to-date listing of supported platforms.

5 SSV TRANSACTIONS

5.1 Social Security On-line Verification (SSOLV) Tansaction

Purpose: The SSOLV Transaction is used by an authorized MVA (End User) to request the verification of an SSN provided by an applicant or that is found on the MVA's database to aid in the prevention of fraudulent identification issuance.

Transaction Message Flow Diagram



1. The MVA (End User) formats the request into the AMIE format and forwards the it to the SSA through the AAMVAnet network.

2. SSA receives the request and responds to the State of Inquiry (SOI) with the verification data in the AMIE format.

Note. For detailed information on the message formats, the AMIE blocks and the data elements, refer to Appendixes A, B, C and D.

5.1.1 'SS' - SSA Verification Request Message

5.1.1.1 State of Inquiry (SOI) Processing Requirements:

The SOI must provide the following data elements to successfully process the SSA Verification Request (SS):

- Social Security Number (DDVSSN) Required
- Driver Name (DDVNM4) Required (See the SSA Name Formatting Rules in the Appendix)
- Driver Date of Birth (DDVDOB) Required

In addition the SOI may include the following elements:

- Jurisdiction (DDLJU1) Optional

- Driver License Number (DDLNUM) Optional
- Return as Received (GRREC2) Optional

NOTE: Do not attempt to verify SSNs allocated by user applications (e.g. the CDLIS substitute and pseudo-SSN), because the SSA will always respond that such SSNs are invalid.

5.1.1.2 Social Security Administration (SSA) Processing Requirements:

Upon receiving the SSA Verification Request (SS), the Social Security Administration (SSA) will search for the requested record in its database.

NOTE: The SSA will not edit or check for errors in the SS message, it only verifies the data present.

5.1.1.2.1 SSOLV Name Match Criteria

A name (see the SSA Name Formatting Rules in the Appendix) provided by the MVA will be accepted as verified, if a match is made using any the following criteria:

1. If the first seven positions of the surname (e.g.: last name) and the first and middle name initials match exactly.

2. If only one initial is provided, the first seven positions of the surname must match and the initial provided will match the first initial of either the first or middle name.

3. The first four positions of the input first name and the first four positions of the file first name match.

4. If no first name is provided, the first four positions of the surname and the first and middle name initials must match.

5. A one letter difference or transposition of two adjacent letters in the first seven positions of the surname and the first and middle name initials match exactly (AB=AB) or are transposed (AB=BA).

6. A one letter difference or transposition of two adjacent letters in the first seven positions of the surname and

a) the first or middle initial of the MVA name match that of the first name initial of the SSA name when only one initial is present on SSA files (AB=A or BA=A); or
b) the first initial of the MVA name matches the first or middle initial of the SSA name when only one initial is present on the MVA record (A=AB; A=BA; B=BA; B=AB); or
c) the MVA first name initial matches the SSA first name initial and the MVA middle name initial disagrees with the SSA middle name initial, but matches the first initial of another

surname for a female (AB SM@TH = AG SMITH X REF - Brown, sex = female, i.e. a maiden name check).

7. An extraneous or missing letter is present in the first seven positions of the MVA surname and the MVA first name initial matches the SSA first or middle name initial.

Extraneous Letter	Missing Letter
A JJOHNSO = A JOHNSON	AR OHNSTON = A JOHNSTON
A J@OHNSO = A JOHNSON	A JHNSTON = A JOHNSTON
A JOSHNSO = A JOHNSON	A JONSTON = A JOHNSTON
B JOHHNSO = AB JOHNSON	A JOHSTON = A JOHNSTON
B JOHNOSO = AB JOHNSON	B JOHNTON = AB JOHNSTON
B JOHNSTO = AB JOHNSON	B JOSHSTN = AB JOHNSTON

8. A compound surname may only be verified using one surname. If the single MVA surname contains more than three letters SSA will compare it to up to 13 positions of the SSA name. SSA will compare positions 1-7, then 2-8, then 3-9, then 4-10, then 5-11, then 6-12, and finally 7-13. If a match occurs on any one of these comparisons, the compound surname will be verified.

5.1.1.2.2 SSOLV Date of Birth (DOB) Match Criteria

A DOB will be verified if it matches the SSA DOB using the following criteria:

1. The year of birth on the MVA record matches the year of birth on the SSA record exactly. The day and month are ignored.

2. The year of birth on the MVA record differs from the SSA DOB +/- one year and the month on the MVA record matches the SSA month.

5.1.1.2.3 SSOLV SSN Match Criteria

The SSN sent on the verification request will only be reported as verified if it matches the SSN found on the SSA record exactly.

5.1.2 'HS' - SSA Verification Response Message

5.1.2.1 Social Security Administration (SSA) Processing Requirements:

After checking for a record in its database, the SSA will send the SSA Verification Response (HS) message to the SOI with the SSA Verification Response Code (GMSVRC) in the MEC block.

The following is a list of SSA Verification Response Codes returned and a description of their meaning:

Code	Description
1	SSN, Name and DOB verified
2	Invalid SSN
3	Name did not verify, DOB is valid
4	DOB did not verify, Name is valid
5	Name and DOB did not verify
6	Unable to process request - go to the local Social Security office for more information
9	System Error. Unable to process at this time

5.1.2.2 State of Inquiry (SOI) Processing Requirements:

The SOI should examine the SSA Verification Response Code (GMSVRC) on the HS message.

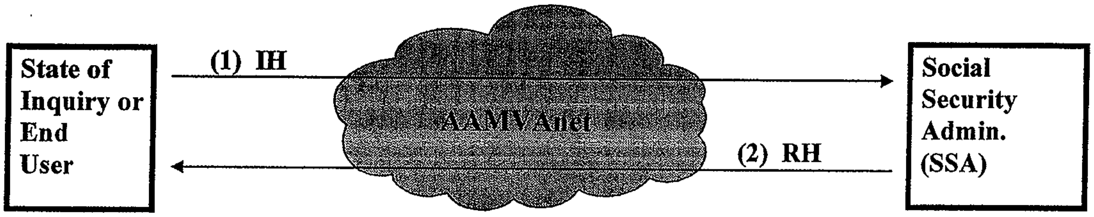
If the driver identification information is verified and the applicant has nothing on his/her record to prevent the issuance/renewal of a license the application/renewal may be processed.

If the information provided by the applicant does not verify, the MVA will utilize jurisdiction specific procedures for handling the applicant.

5.2 Help America Vote Verification Transaction

Purpose: The HAVV transaction is used by an authorized MVA (End User) to request the verification of a name, date of birth and a partial SSN (last four digits) provided by an applicant to aid in the prevention of fraudulent voter registration.

Transaction Message Flow Diagram



1. The State of Inquiry (MVA) formats the request into the AMIE format and forwards it to the SSA via the AAMVAnet network.

2 The SSA receives the request and responds to the State of Inquiry (SOI) with the verification data in the AMIE format.

NOTE:For detailed information on the message formats, the AMIE blocks and the data elements used in HAVV, refer to Appendixes A, B, C and D. For detailed information on interfacing your application to AAMVAnet, refer to the Unified Network Interface Application Developers Reference Manual (available through the AAMVA Operations Department).

5.2.1 'IH' - HAVA Verification Request Message

5.2.1.1 State of Inquiry (SOI) Processing Requirements:

The SOI must provide the following data elements to successfully process the HAVA Verification Request (IH):

- Last four digits of SSN (DDVSLF) Required
- Name (DDVNM4) Required (See the SSA Name Formatting Rules in the Appendix)
- Date of Birth (DDVDOB) Required

In addition the SOI may include the following elements:

- Return as Received (GRREC2) Optional

This data will be validated by SSA, the edits performed are shown in the next section. Jurisdictions should ensure they do not send data that will fail these edits.

5.2.1.2 Social Security Administration (SSA) Processing Requirements:

Upon receiving the HAVA Verification Request (IH), the SSA will validate the contents of the message. If any of the following edit rules fail, the response will have the SSA Verification Response Code (GMSVRC) set to 'S', indicating "Invalid Data".

- The last four digits of the SSN must be a number in the range "0001" to "9999".
- The Date of Birth must be a valid date (though the day of birth is not used).
- The First Name must have:
  - A-Z in position 1.
  - Then in positions 2 through 15: A-Z, a single embedded hyphen, apostrophe or space.
  - Last character must be A-Z, an apostrophe or a space; unless the 14th position is A-Z, then the 15th position can be a hyphen, apostrophe, space or an alphabetic character
  - Consecutive embedded combinations of spaces, hyphen, and/or apostrophe are not permitted.
- Last Name must have:
  - A-Z in position 1.
  - Acceptable characters for position 2 through 20 are A-Z, a single embedded hyphen, apostrophe or space.
  - Last character must be A-Z, an apostrophe or a space; unless the 19th position is A-Z, then the 20th position can be a hyphen, apostrophe, space or an alphabetic character.
  - Consecutive embedded combinations of spaces, hyphen, and/or apostrophe are not permitted.

The name in the message will be in a packed form (see the SSA Name Formatting Rules in the Appendix for details).

Valid messages are then checked against the SSA database using the following elements:

Input	Match Criteria
Last Name	Exact
First name	Exact
Middle Initial	Ignore
Date Of Birth	Month and year must be exact. Ignore day.
Last four digits of the SSN	Exact

5.2.2 'RH' - HAVA Verification Response Message

5.2.2.1 Social Security Administration (SSA) Processing Requirements:

After checking for a record in its database, the SSA will send the HAVA Verification Response (RH) message to the SOI with the Response Code in the MEC block.

The following is a list of the SSA Verification Response Codes (GMSVRC) returned for HAVV and a description of their meaning.

Code	Description
S	Invalid input data
T	Multiple matches – all deceased
V	Multiple matches – all alive
W	Multiple matches – at least one alive (& at least one deceased)
X	Single match – alive
Y	Single match – deceased
Z	No match found
9	System Error. Unable to process at this time

5.2.2.2 State of Inquiry (SOI) Processing Requirements:

The SOI should examine the SSA Verification Response Codes (GMSVRC) on the 'RH' message.

If the information provided by the applicant does not verify, the MVA will utilize jurisdiction-specific procedures for handling the applicant.

Florida Voter Registration System

Guide to FVRS

Version 1.0

September 7, 2005

Florida Department of State
Division of Elections
FVRS Project Office
409 East Gaines Street
Tallahassee, Florida 32399-0250
850.245.6229

Jeb Bush
Governor

Glenda E. Hood
Secretary of State

EXHIBIT
C

Table of Contents

List of Tables

List of Figures

11 VOTER REGISTRATIONS AND APPLICATIONS

When a voter registration application is submitted to FVRS, it is held in an application table until the application has been completely processed. A voter who is already registered may, therefore, have an active registration record, and an unresolved application. This allows the official registration record to be maintained undisturbed while an application is being processed. Application records are linked to their parent voter record by the *FVRS Voter ID Number.* Each application record is further qualified by a sequence number assigned by FVRS upon receipt of an application.

This relationship between a voter record and one or more application records will also be implemented for new registrations where an existing voter record does not previously exist. Under this condition the relevant data elements from the application will be used to populate and create a voter record, generate a unique FVRS Voter ID Number and create an application record related back to the Voter Record.

11.1 REGISTRATION PROCESSING AND DISPOSITION TERMS

For the purpose of clarity, the following terms have a precise meaning in the context of FVRS.

11.1.1 Application Processing Status

An application processing status will be assigned to all voter registration applications submitted to and accepted by FVRS[3]. This designation defines a workflow or processing state and does not define an application's final disposition. An application's processing status may change during its life cycle. The discreet processing statuses and definitions to be managed by FVRS are described below.

Status	Description
Suspended	Voter registration applications can be submitted to FVRS with a suspended status which will instruct FVRS not to apply further validity, verification or eligibility assessment procedures. A suspended application may be submitted by a county data entry operator for the purpose of later retrieval and completion of data entry or for the purpose of routing the application to another county for completion. Suspended applications should be attended to promptly by the assigned county to avoid delay in the registration process.

[3] Only applications which fail basic data validation rules will be rejected by FVRS and not assigned an FVRS ID number.

Status	Description
Pending	A new registration application is pending when it is received by FVRS and the application did not meet the criteria for a Denial or Incomplete disposition, and where the application is still being processed by the Department of State for the purpose of verification of Driver's License Number or Social Security Number conditions.
Closed	An application is closed when a disposition of the application is determined and assigned. The types of valid dispositions that may be assigned to an application are listed and described in Section 11.1.2

11.1.2 Application Dispositions

An application disposition will be assigned to all voter registration applications submitted to and accepted by FVRS for processing. This designation defines the standing of the ***application*** presented for processing and not necessarily the Voter Registration Status (see Section xxx) of the ***registrant***. This distinction is important for applications received as updates for existing FVRS registrants. For instance, an "incomplete" application disposition for an existing eligible voter will not affect the registrant's current voter registration status. The discreet application dispositions and definitions to be managed by FVRS are described below.

Disposition	Description
Denied	Once an application is denied, the voter is provided a notification. The following are the reasons for an application being denied. • Applicant was not 17 years old on the date of the application • The applicant is not a US Citizen A denied voter is not sent an application form with a Denial letter since a new application will not cure the problem (with the exception of the under 17 voter, where time will take care of the problem.)
Incomplete	A voter registration application is complete if it contains the following information necessary to establish eligibility pursuant to s. 97.041: 1. The applicant's name. 2. The applicant's legal residence address. 3. The applicant's date of birth. 4. A mark in the checkbox affirming the applicant is a citizen of the United States. 5.a. The applicant's current and valid Florida driver's license number or , the identification number from a Florida identification card issued under s. 322.051, or b. If the applicant has not been issued a current and valid Florida

Disposition	Description
	driver's license or a Florida identification card, the last four digits of the applicant's social security number. c. In the case where an applicant has not been issued a current and valid Florida driver's license or Florida identification card or social security number, the applicant shall affirm this fact in the manner prescribed in the uniform statewide voter registration application. 6. A mark in the checkbox affirming that the applicant has not been convicted of a felony or that, if convicted, has had his or her civil rights restored. 7. A mark in the checkbox affirming that the applicant has not been adjudicated mentally incapacitated with respect to voting or that, if so adjudicated, has had his or her right to vote restored. 8. Original signature or a digital signature transmitted by the Department of Highway Safety and Motor Vehicles of the applicant swearing or affirming under the penalty for false swearing pursuant to s. 104.011 that the information contained in the registration application is true and subscribing to the oath required by s. 3, Art. VI of the State Constitution and s. 97.051. Notes: 1. *An applicant whose application is denied is sent an incompletion notice listing the reasons for the application not being processed and another application form so that a corrected application can be presented.* 2. *An application to update an existing registration that contains incorrect information or information that can not be verified may acquire an incomplete status. This will allow a notification to be generated, but will NOT alter the voter registration status.* 3. *A voter that has a registration status of Active, Inactive or Pre-Registered cannot be moved to a Denied status. If a voter becomes ineligible, an administrative process must be used to remove the voter.*
Registered	The voter's registration record has been updated with all possible information

11.2 Voter Registration Status

Each voter maintained in FVRS will be assigned a Voter Registration Status which will determine the voter's eligibility to vote. The Voter Registration Status will be updated after an application is processed (application processing status "closed") and an application disposition has been assigned. The discreet voter registration statuses and their definitions to be managed by FVRS are described below.

Status	Description
Active	The voter is properly registered. The voter is eligible to vote in elections.
Inactive	There is one and only one way to acquire an inactive status. Each and every one of the following events must have happened in the correct order: • The voter had an active status • First class mail was returned undelivered from the residence address of record for the voter • An "Address Confirmation Notice" has been sent to the voter • No response was received from the voter for 30 days following the sending of the Final notice 1. *At this point the voter becomes Inactive. The voter is still eligible to vote in elections, and is included in the precinct register.* Any "voter activity" by the voter (which broadly is voting, or written contact from the voter, or signing a petition) will restore the voter to Active Status After two general elections the voter is moved to the "Removed" status.
Removed	The voter is no longer eligible to vote in an election, and will not appear in the precinct register. There are a number of reasons why a voter can be removed: • Failed to attend Admin Hearing • Office errors • Canceled • Deceased • Felon • Moved out of State. Request by voter • Adjudicated Mentally Incompetent • Office Duplicate Registration • Returned Mail, Inactive 2 yrs
Archived	Only voters with Removed Status can become Archived. The only purpose of doing this is to prevent long deceased voters from overwhelming valid voters when doing voter searches.
Denied	The person (citizen or not) was not a registered voter, and their most recent attempt at registration was denied.
Incomplete	The citizen is not a registered voter, and their most recent registration attempt was Incomplete.

Status	Description
Pre-registered	The voter has met all the requirements to be an Active voter but has not yet attained the age of 18. Pre-registered voters that will be 18 years old on or before the election date are included in the precinct register and are eligible to vote in the election, even with Pre-registration status. The voter must be 17 years old to pre-register.
Pending	As soon as a voter receives a FvrsVoterIdNumber, an entry is made in the FVRS Voter table. For new registrations, prior to HSMV and other match processing, the status of the voter will be Pending. This status is only assigned to people making a new voter registration application which have not yet reached disposition.

12 Voter Registration Processing by Counties

The following sections describe the typical steps a county voter registration clerk will execute to submit a registration application for a new voter to FVRS. The processes described in the following sections differ slightly from procedures employed for processing applications from HSMV. Such applications will not have a paper application form and will be transmitted electronically to the FVRS.

Further, the procedures described in this section do not include locally defined workflow or processing steps required by counties. Such locally defined steps may include document preparation or scanning of voter registration applications, but will typically not necessitate interface with FVRS.

Further, the steps outlined in this section assume that a voter registration clerk meets all county security requirements for access to the county voter registration system and the county security administrator has granted appropriate FVRS permissions.

12.1 Processing Registration Forms for Voters Outside of a County's Jurisdiction

FVRS enables any voter registration official to access or update any registration record. This is an important and necessary feature of a statewide system for many reasons including:

- Each voter will be assigned a unique FVRS ID number that will be maintained continuously despite changes in address or voter status. This means that a voter affects a change in legal residence from one county to another through an update to his existing voter registration record. Thus, a voter registration official must be capable of accessing an existing registration record and execute an address update that removes the voter from the jurisdiction of one county and places the voter in another county.

- Any authorized voter registration official shall be capable of simultaneous access to the FVRS from any location with secure communications to FVRS. This offers a previously unavailable level of convenience to the voter for obtaining a common set of services from any voter registration official.
- Voter registration forms may be mistakenly mailed or directed to counties other than the legal residence of a voter. In such cases the jurisdiction receiving the forms shall process the voter registration as described in the sections below and forward the original paper form to the county of jurisdiction. Section 97.053(7) F.S. provides specific direction to voter registration officials under these circumstances.

12.2 General Procedures

While the following sections relate processing steps by voter registration clerks to transactions serviced by FRS, in fact, the county voter registration system in use will shield the clerk from any direct interface with FVRS transactions. The presentation layer of the county voter registration application shall provide all dialogues and data entry forms to be used by the clerk. The county voter registration system will generate the request to FVRS, receive the FVRS response and format the response message within the presentation layer of county system.

The following sections provide a simplified step-by-step description of typical voter registration processes. Variations in these processes are nearly infinite and may be driven by county standards and procedures.

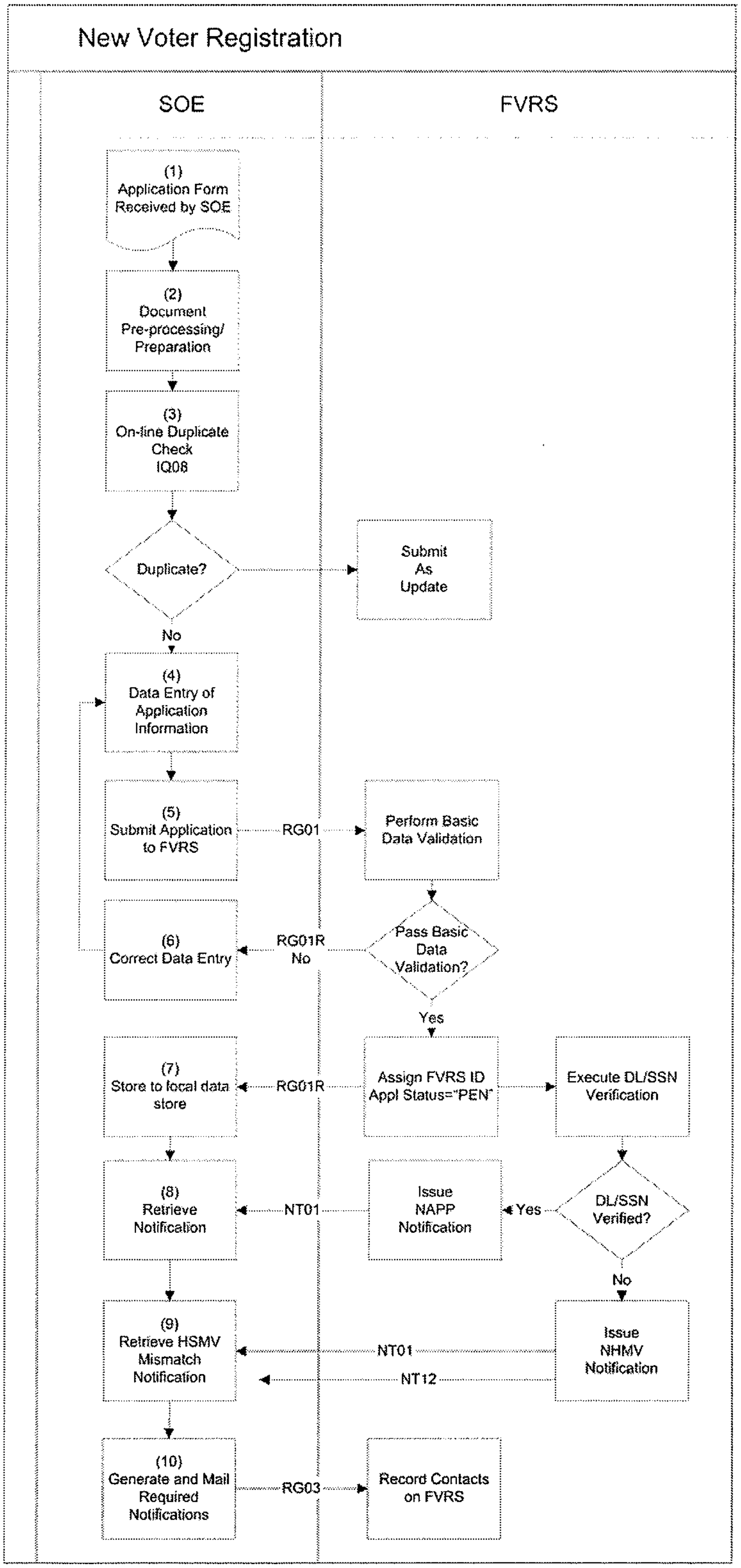
12.3 New Voter Registration

This section will delineate the key steps involved in processing a new voter registration. Most of the steps comprising the new registration process are depicted in Figure 9.

Steps 1 and 2 – Receipt and Pre-processing Registration Forms

The processing of a new voter registration begins with the receipt of a valid voter registration form. This may be a valid State of Florida voter registration form, a Federal postcard form or a National Mail Registration Form. This step is shown in Figure 9 as step 1. Local procedures for opening mail, time-stamping documents or pre-scanning are not prescribed by FVRS, but are left to the discretion of each county. At a minimum, however, each county should have established procedures for document control and pre-screening for valid documents.

Figure 9 Typical New Voter Registration Process


New Voter Registration
SOE
FVRS
(1) Application Form Received by SOE
(2) Document Pre-processing/ Preparation
(3) On-line Duplicate Check IQ08
Duplicate?
Submit As Update
No
(4) Data Entry of Application Information
(5) Submit Application to FVRS
RG01
Perform Basic Data Validation
(6) Correct Data Entry
RG01R
No
Pass Basic Data Validation?
Yes
(7) Store to local data store
RG01R
Assign FVRS ID Appl Status="PEN"
Execute DL/SSN Verification
(8) Retrieve Notification
NT01
Issue NAPP Notification
Yes
DL/SSN Verified?
No
(9) Retrieve HSMV Mismatch Notification
NT01
NT12
Issue NHMV Notification
(10) Generate and Mail Required Notifications
RG03
Record Contacts on FVRS

Step 3 – Check for Possible Duplicate Registration

Before a new voter can be registered with FVRS, a search must be made of the existing registrations on the FVRS data base to insure that the application is indeed a new registration and not an update to an existing one. This is accomplished with the "New Registration" option of the Voter Search transaction (IQ08). Voter identity information such as first name, last name, middle name, and date of birth is submitted to FVRS via the IQ08 input message format. FVRS returns a list of records matching the identity information. Based on the results presented, it is the operator's decision whether the application represents a new voter registration or an update to an existing registration record. This assessment should be completed for each application regardless of the applicant's selection of checked boxes on the application form.

While the IQ08 transaction provides for a specific lookup for New Voters as using the name and birth date of the applicant, it also provides for more exhaustive "General" searches based on a variety of criteria such as address, driver's license number, social security number, etc.

This assessment will become particularly relevant during the critical months after the FVRS becomes operational. During this period most voters may not understand the distinction between a new registration and an update to an existing registration. This may be most evident in a change of address that results in a move between two counties. Prior to FVRS this event would have required the issuance of a new registration, however, after January, 2006 this same action will become an update to an existing registration.

The IQ08 transaction will search both application and voter records (see Section 11). The voter records being searched will include those that are removed ("REM"), administratively ("ADM") deleted and other voter registration statuses. If the existing voter record has a voter registration status other than "INA", "ACT", or "PRE", then, even though there is an existing record, you should supply the FVRS Voter ID Number but use a TransactionType = 'R' for new registration for that record in the RG01 transaction.

An existing *application* can be retrieved using IQ09, and an existing *voter record* can be retrieved using IQ01. County systems can display information from these transactions, to assist with data entry.

Step 4 and 5 – Data Entry of Application and Submission to FVRS

The county's voter registration program accepts the details from the voter registration form and performs all local data validation edits such as valid dates, compliance with mandatory fields and other minimum data requirements. If the voter resides in the current county, the residence address should be validated against data maintained by the county system and the precinct and district information included in the voter registration details. If the voter resides in a different county, then the FVRS will validate the residential address against the street and address data maintained in FVRS (see Section 17). The operator may now submit the application to FVRS through the county voter registration system which will invoke an RG01 transaction.

Steps 6 and 7- Edit and FVRS Voter ID Number Assignment

FVRS will reply to an RG01 transaction with an RG01R response. If the application submitted to FVRS cannot be accepted due to basic data validation errors, invalid security or an inability to validate the message digest, the RG01R will respond without assigning an FVRS ID number and with error message(s) enumerating any errors.

If FVRS can accept the applications then the RG01R response will include an FVRS Voter ID Number. The FVRS Voter ID Number will be provided in RG01R only if the RG01 transaction processed successfully. The county voter registration system should display this number to the operator as an acknowledgement and in case local procedures direct this number to be recorded on external documents such as the original registration form. It is also essential that you use this FVRS Voter Id Number in all subsequent transactions concerning the same application.

If no errors are reported in the FVRS RQ01R reply, the processing for steps 6 and 7 are complete and the voter record is given an application processing status of *pending* (see Section 11.1.1).

FVRS will also issue an NAPP notification, providing the county voter registration with an application acknowledgement. An FVRS IQ09 transaction may be used to retrieve the application processing status. Note that a period of time may elapse before the application completes all FVRS verifications and receives an application disposition (see Section 11.1.2).

How to Process Errors Reported by FVRS

FVRS will apply an evaluation immediately to applications submitted through the RG01 transaction. This level of evaluation will be limited to checks for completeness, compliance with basic data format rules, adherence to security and consistency within application elements. Other business rules requiring further verifications against FVRS data or by other external agencies such as Highway Safety (driver's license) or the Social Security Administration (social security number) will be scheduled automatically by FVRS according to processing agreements with those agencies.

Any errors detected by FVRS upon receipt of the application will be reported in the RG01R reply. These error codes should be interpreted and displayed to the operator.

The operator may then correct the data entry and retry the transaction or take one of the following steps to update the application processing status or the application disposition (see Section 11.1) by processing an RG01 with an appropriate TransactionType.

Action	Explanation
Suspend the Application	The suspended application is held on FVRS with the assigned FVRS Voter ID Number. The application may then be researched, retrieved and completed (see Section 12.7).
Update the application disposition as "incomplete"	An NINC notification will be created. Appropriate communications to the voter will be scheduled by FVRS. An NWFL notification is created for an incomplete notice (RegIncomp) (see Section 19)
Update the application disposition as "denied"	An NDEN notification will be created. Appropriate communications to the voter will be scheduled by FVRS.

Scan and Index the Application Image

Final adjudication of an application by the Department may require manual comparison of the voter registration application against other records to ascertain the accuracy of matching processes. This may be particularly true in the felon matching processes to take one example. Access to an image of the voter registration application may, therefore, be necessary to complete the application processing. Thus, the application image should be scanned and transmitted to FVRS within 24 hours of entry of the application into the system.

The FVRS IM01 transaction may be used to transmit document images to FVRS and link them with the appropriate voter record. For each application there may be two images. One is the complete application image, and the other is a clipped signature.

For suspended applications, an NSUS notification is issued to the targeted county after the images have been received by FVRS. For Suspense applications, no further processing is done.

An application that receives a denied or incomplete disposition is fully processed, and only communications with the voter need to be generated.

New applications that are Pending, Denied or Incomplete update the voter's information on the voter table. Suspense applications remain on the application table and do not affect the voter record. For Pending applications proceed to Step 5, otherwise proceed to Step 8.

Step 9 - HSMV Verifies Driver's License Number and/or Last 4 Digits SSN

Only applications with a status of Pending (i.e., step 7 completed without errors) will be forwarded to HSMV for verification of driver's license numbers or last 4 digits of SSN. HSMV will execute verifications of driver's licenses and will determine one of the following:

- Driver's license is correct
- Driver's license number was not provided, but voter appears to have been issued a driver's license
- Driver's license is incorrect or does not match the name provided on the voter registration application

Where necessary HSMV will forward the necessary information to SSA for verification of social security numbers who will provide the following assessment:

- Invalid Data
- Multi Matches All Deceased
- Multi Matches All Alive
- Multi Matches Mixed
- Single Match Alive
- Single Match Deceased
- No Match Found
- System Error: Unable to Process at this Time

The Department will manually review errors and determine if the voter has made an error in reporting their driver's License Number of last 4 digits of the SSN.

- If an error was made, a NHMV notification is created. The county system then uses NT12 to retrieve information about the DL or SSN4 error. If the county determines the voter registration is in error, an RG01 is processed with a transactiontype of 'I; making the application incomplete. FVRS creates an NINC notification. Proceed to Step 8.
- If the county determines that the registration is correct an RG01 transaction is processed with the HSMVOverride flag set to 'Y'. This progresses the application to Step 6.

FVRS Registration Update

At this point the application is completed and the voter is registered. The voter's status is changed to "ACT".

FVRS creates a Notification to the county SOE for any required communications to the voter. These Notifications typically include pre-registration welcome letters, blank party letters and Voter ID cards. Each document to be sent to the voter will be a notification message.

An NNRG notification is created when a new voter receives an ACT or PRE status. NWFL notifications are created for each of the documents that the voter may receive:

- Blank Party letter
- Pre Registration Letter
- Voter Information Cards

Steps 8 and 9 - County Retrieves Notifications

Notification retrieval is a process execute by the county voter registration system. The purpose of this process is to retrieve notifications from FVRS. Through the notification retrieval process, any changes to the FVRS voter record may be retrieved, and the local database updated. It is only on retrieval of the notification that the county knows whether the new registration attempt has been completed and the disposition assigned to the application and voter registration status.

Step 10 Voter Documents are printed

Contact workflow items are scheduled through the notification process for documents that need to be sent to the voters. When the county prints the documents, a "Registration Contact Add" (RG03) transaction is sent to FVRS.

12.4 Updates to Existing Voter Registration Records

Step 1 - Search for Existing Voters

Before an update can be applied to an existing voter registration record, a search must be made of the existing voters. This is accomplished with the FVRS IQ08 transaction. Voter identity information such as voter id number, name, date of birth, etc. is submitted to FVRS via the IQ08 input message format, and FVRS returns a list of records matching the identity information. Based on the results presented, it is the operator's decision whether the application represents a new voter registration or an update to an existing registration record. This assessment should be completed for each application regardless of the applicant's selection of checked boxes on the application form.

This assessment will become particularly relevant during the critical months after the FVRS becomes operational. During this period most voters may not understand the distinction between a new registration and an update to an existing registration. This may be most evident in a change of address that results in a move between two counties. Prior to FVRS this event would have required the issuance of a new registration, however, after January, 2006 this same action will become an update to an existing registration.

The IQ08 transaction will search all voter records (see Section 11). The voter records being searched will include those that are removed ("REM"), archived ("ADM") deleted, Pending ("PEN") and other voter registration statuses. If you are able to locate an existing record for the person being processed, you should supply the FVRS Voter ID Number for that record in the RG01 transaction.

SSA's HAVA Verification

Peter Monaghan
Social Security Administration
February 6, 2006

EXHIBIT
D

tabbies

Legislation Verification Highlights

- MVA responsible for verifying information and working with SSA
- Drivers License number primary source
- If no DL/ID, election agency collects last four digits of SSN
- SSA must develop process to provide "last four digit" verification
- SSA to be reimbursed for funds expended

System Development Highlights

- October 2002 – President signs HAVA
- November 2002 – SSA begins internal work
- February 2003 – SSA & AAMVA reach conceptual agreement on telecommunications
- May 2003 – SSA joins NASS-NASED-AAMVA Task Force
- January 2004 – "Joint Communiqué" sent to MVAs, Election Agencies
- August 2004 – SSA's systems development complete
- October 2004 – First live use of system (Iowa)

Agreements

- Election Agency - MVA
- MVA - AAMVA
- AAMVA - SSA: telecommunications and billing
- SSA & MVA: Privacy, process and reimbursement

SSA/MVA Agreement Process

- SSA developed “model” agreement
- Discussions between SSA Regional Office and MVA
- “Final” agreement reviewed by SSA attorneys
- Signed by MVA and Regional Commissioner
- Data transmission can begin

Participation to Date

55 total jurisdictions:

- 8 exempt
- 23 signed agreements
  - 19 implemented
  - 4 in testing
- 8 final review of agreement
- 13 agreements underway
- 3 no current SSA activity

Implementation Process

- Agreement discussions SSA/MVA
- Systems testing with “dummy” data
- Agreement finalized
- Exchange of “live” data can begin

Verification Routine

- Election agency collects name, DOB and last four digits of SSN
- Transmitted to MVA
- MVA transmits to SSA via AAMVA
- SSA does exact search of name, month/year of birth and "last four"
- Real-time reply sent to MVA via AAMVA

Verification Replies

Response Code	Definition
S	Invalid Data
T	Multi Matches All Deceased
V	Multi Matches All Alive
W	Multi Matches Mixed
X	Single Match Alive
Y	Single Match Deceased
Z	No Match Found
9	System Error

Results to Date

143,000 queries

Response	Percent
Invalid Data	.001%
Multi Matches All Deceased	0
Multi Matches All Alive	.014%
Multi Matches Mixed	.001%
Single Match Alive	69%
Single Match Deceased	2.5%
No Match Found	28.5%

Reimbursement

- SSA legally precluded from using trust funds for non-SSA work
- HAVA specifically provided for reimbursement
- Development, maintenance and ongoing SSA & AAMVA costs included
- Billing through AAMVA

Contact Information

Pete Monaghan:

- 410-966-9972
- Pete.Monaghan@SSA.gov

SSA's HAVA Verification

Peter Monaghan

Social Security Administration

February 12, 2007

tabbies

EXHIBIT

E

Participation to Date

55 total jurisdictions:

- 7 exempt
- 42 signed agreements
  - 32 implemented
  - 10 in testing or not yet using
- 3 agreements underway
- 3 other

Agreements

- Election Agency - MVA
- MVA - AAMVA
- AAMVA - SSA: telecommunications and billing
- SSA & MVA: Privacy, process and reimbursement

Reimbursement

- SSA legally precluded from using trust funds for non-SSA work
- HAVA specifically provided for reimbursement
- Development, maintenance and ongoing SSA & AAMVA costs included
- Billing through AAMVA

Development Highlights

- NASS sponsored workgroup:
  - NASED, AAMVA, SSA, several States
- Desired outcomes:
  - Prevent fraudulent voting
  - Avoid incorrect prevention of registration
  - Provide maximum information
- Group developed requirements based on:
  - Desired outcomes
  - Limitations of information

Design Considerations

- Limited input information
- Balance inclusion v. exclusion
- True match unknowable:
  - Each "last four" equals 40,000 SSNs
- Provide maximum output information to election agency

Verification Routine

- Election agency collects name, DOB and last four digits of SSN
- Transmitted to MVA
- MVA transmits to SSA via AAMVA
- SSA does exact search of name, month/year of birth and "last four"
- Real-time reply sent to MVA via AAMVA

Verification Replies

Response Code	Definition
S	Invalid Data
T	Multi Matches All Deceased
V	Multi Matches All Alive
W	Multi Matches Mixed
X	Single Match Alive
Y	Single Match Deceased
Z	No Match Found
9	System Error

Results to Date

2.6 million queries

Response	Percent
Single Match - Alive	53.3%
Single Match - Deceased	.3%
No Match Found	46.2%
All Other	.3%

Results

- False positives cannot be identified
- Match rate?
  - Last name changes not reported to SSA
  - Proper first name historically not required
  - Exact month/year of birth required
- SSA data may be outdated, incorrect

Contact Information

Pete Monaghan:

- 410-966-9972
- Pete.Monaghan@SSA.gov

9/27/2004

BOARD OF ELECTION & DEPARTMENT OF MOTOR VECHICLE IDENTIFICATION VERIFICATION REPORT

BOROUGH	BOE DATA ENTRY ERROR	VOTER ERROR	DMV ?	TOTAL PER BORO
MANHATTAN	1009	228		1237
BROOKLYN	885	171	2	1058
QUEENS	670	140	5	815
BRONX	250	48		298
STATEN ISLAND	137	22	1	160
TOTAL COUNTYWIDE	2951	609	8	3568

BOE Data Entry Errors – Numbers were data entered wrong.
Social Security numbers were put in as the drivers I.D. #.
Telephone numbers were put in as the drivers I.D. #.
Voter serial numbers were put in as the drivers I.D. #.
Zip code numbers were put in as the drivers I.D. #.

Voter Errors - Social Security numbers were put in as the drivers I.D. #.
Out of state driver I.D. # (only the number, not a copy of their card).
Number did not match DMV I.D. # .

DMV Errors – Voter on Queens DMV list, but moved to Staten Island.
DMV I.D. # matched the DMV voter registration, same name, same date of birth, but on DMV Error Report.

EXHIBIT
F

EXHIBIT G

FLORIDA VOTER REGISTRATION APPLICATION

YOU CAN USE THIS FORM TO: REGISTER TO VOTE IN THE STATE OF FLORIDA • CHANGE NAME OR ADDRESS • REPLACE YOUR DEFACED, LOST, OR STOLEN VOTER INFORMATION CARD • REGISTER WITH A POLITICAL PARTY OR CHANGE PARTY AFFILIATION • UPDATE YOUR SIGNATURE

To Register, You Must:

- Be a citizen of the United States of America. **(BOX #2)**
- Be a Florida resident. **(BOX #8)**
- Be 18 years old (you may pre-register if you are 17). **(BOX #5)**
- Not now be adjudicated mentally incapacitated with respect to voting in Florida or any other state. **(BOX #4)**
- Not have been convicted of a felony in Florida, or any other state, without your civil rights having been restored. **(BOX #3)**
- Provide your current and valid Florida driver's license number or Florida identification card number. If you do not have a current and valid Florida driver's license or Florida identification card, you must provide the last four digits of your Social Security number. If you do not have a FL DL#, FL ID card#, or SSN, write "NONE" in the box. **(BOX #6)**
- Complete all information in the black boxes on the application. **(BOXES #2,3,4,5,6,7,8 & 16)**

Deadline Information:
If this is a new registration application, the date the completed application is postmarked or hand delivered to a driver's license office, a voter registration agency, an armed forces recruitment office, the Division of Elections, or the office of any supervisor of elections in the state will be your registration date. If this is a new Florida application, you must be registered for at least 29 days before you can vote in an election. If your application is complete and you are qualified as a voter, a voter information card will be mailed to you.

Party Affiliation (BOX #12):
If you wish to register with a major political party, place an **"X"** in the box preceding the listed party with which you wish to affiliate. If you wish to register with a minor political party, place an **"X"** in the box preceding "Other, Minor Party" and print the name of the party with which you wish to affiliate. A list of the minor political parties is on the website for the Division of Elections: **http://election.dos.state.fl.us/online/parties.shtml** If you wish to register without party affiliation, place an **"X"** in the box preceding "No Party Affiliation".

Florida is a closed primary state. If you wish to register to vote in a partisan primary election, you must be a registered voter in the party for which the primary is being held. All registered voters, regardless of party affiliation, can vote on issues and non-partisan candidates.

Notice:
The office at which you register, or your decision not to register, your SSN, your FL DL# and your FL ID card# will remain confidential and will be used only for voter registration purposes.

Note: If the information on this application is not true, the applicant can be convicted of a felony of the third degree and fined up to $5,000 and/or imprisoned for up to five years.

Questions:
Contact the office of your county supervisor of elections for additional information. Contact information is on the website for the Division of Elections: **http://election.dos.state.fl.us/county/index.shtml**

Informacion en Espanol:
Sirvase llamar a la oficina del supervisor de elecciones de su condado si le interesa obtener este formulario en Español.

PLEASE COMPLETE THE APPLICATION BELOW. PLEASE PRINT USING A BLACK BALL POINT PEN.

1) Black boxes must be completed on the application below for registration to be valid. **2)** Return this completed application to the office of your supervisor of elections. **3)** If you are a first-time voter in this state applying by mail to register to vote and you have not been issued a FL DL#, FL ID#, or SSN, include a copy of your ID with the application. **4)** Mail with first class stamp.

FLORIDA VOTER REGISTRATION APPLICATION

REVISED 1/06

1 Check boxes that apply: □ New Registration □ Address Change □ Party Change □ Name Change □ Card Replacement □ Signature Update | OFFICIAL USE ONLY: DS DE 39 1/06

REQUIRED

2 Are you a citizen of the United States of America? Yes? □ No? □ (If NO, you cannot register to vote)

3 □ I affirm I am not a convicted felon, or if I am, my rights relating to voting have been restored.

4 □ I affirm I have not been adjudicated mentally incapacitated with respect to voting or, if I have, my competency has been restored.

IF YOU ANSWERED NO TO QUESTION 2, OR IF YOU ARE UNABLE TO AFFIRM THE STATEMENTS IN BOXES 3 AND 4, YOU ARE INELIGIBLE TO REGISTER TO VOTE. DO NOT COMPLETE THIS APPLICATION.

5 Date of Birth (MM/DD/YYYY) / /

6 If you have a current and valid FL DL# or FL ID card#, you must provide the number in this box. If you do not have either, provide the last 4 digits of your SSN. If you have not been issued a FL DL#, FL ID card#, or SSN, write "NONE":

7 Last Name | Suffix (circle) Jr. Sr. II III IV | First Name | Middle Name/Initial

8 Address Where You Live (Legal Residence) DO NOT GIVE P.O. BOX. | Apt/Lot/Unit | City | County of Legal Residence | State | Zip Code

9 Mailing Address If Different from Above | Apt/Lot/Unit | City | Country | State | Zip Code

10 Address Last Registered to Vote | Apt/Lot/Unit | City | County | State | Zip Code

11 Former Name if Making Name Change | Day Phone Number

12 Party Affiliation (Check only one) □ Democratic Party □ Republican Party □ Other, Minor Party (print party name): □ No Party Affiliation

13 Race/Ethnicity (Check only one) □ American Indian/Alaskan Native □ Asian/Pacific Islander □ Black, not Hispanic □ Hispanic □ White, not Hispanic

14 Sex □ M □ F | Do you need voting assistance at the polls? □ Yes □ No | Are you interested in being a poll worker? □ Yes □ No | State or Country of Birth

15 Are You: □ Active Duty Military/Merchant Marine □ Dependent of Active Duty Military/Merchant Marine □ U.S. Citizen Currently Residing Outside the U.S.

REQUIRED

16 **OATH:** I do solemnly swear (or affirm) that I will protect and defend the Constitution of the United States and the Constitution of the State of Florida, that I am qualified to register as an elector under the Constitution and laws of the State of Florida, and that all information provided in this application is true.

SIGNATURE: Sign or mark on line in box below. (Invalid without signature or mark of applicant.)

X **Date:**

SPECIAL IDENTIFICATION REQUIREMENTS

If you are registering by mail, you have never voted in Florida, and you have not been issued a Florida driver's license, Florida identification card, or Social Security number, you will be required to provide additional identification prior to voting the first time. To ensure that you will not have problems when you go to vote, you should provide a copy of the required identification listed below at the time you mail your voter registration application.

You may provide a copy of one of the following photo identifications (ID) that includes your name and picture:

- U.S. Passport • Employee Badge or ID • Buyers Club ID • Debit/Credit Card • Military ID
- Student ID • Retirement Center ID • Neighborhood Association ID • Public Assistance ID

Or, you may provide a copy of one of the following documents that contains your name and current residence address:

- Utility Bill • Bank Statement • Government Check • Paycheck • Other Government Document

Or, if you are one of the following persons, you are exempt from having to provide a copy of an ID at this time. These exemptions are:

- Persons 65 years of age or older • Persons with a temporary or permanent physical disability
- Members of the active uniformed service or merchant marine who, by reason of such active duty, are absent from the county
- Spouse or dependent of an active uniformed service member or merchant marine who, by reason of the active duty or service of the member, is absent from the county
- Persons currently residing outside the U.S. who are eligible to vote in Florida

All voters are required to provide ID containing photo and signature at the time of voting in the polling place. Without proper identification, a voter can only vote a provisional ballot.

DO NOT SEND ORIGINAL IDENTIFICATION DOCUMENTS TO THE SUPERVISOR OF ELECTIONS.

Association for Computing Machinery
Advancing Computing as a Science & Profession

Statewide Databases of Registered Voters:

Study Of Accuracy, Privacy, Usability, Security, and Reliability Issues commissioned by the U.S. Public Policy Committee of the Association for Computing Machinery

February 2006

Preface

The Association for Computing Machinery (ACM) is an educational and scientific society uniting the world's computing educators, researchers and professionals to inspire dialogue, share resources and address the field's challenges. ACM strengthens the profession's collective voice through strong leadership, promotion of the highest standards, and recognition of technical excellence. As such, ACM cares deeply about the dependability and reliability of computing technology. Voter registration systems encompass not only the databases that house voter information, but also an entire information technology infrastructure that must be carefully managed by election officials. The U.S. Public Policy Committee of the ACM (USACM) commissioned this study to provide objective technical information and expert recommendations to state and local election officials, policy makers, and the public about these systems.

The USACM serves as the focal point for ACM's interaction with U.S. government organizations, the computing community, and the U.S. public in all matters of U.S. public policy related to information technology.

Supported by ACM's Washington, D.C., Office of Public Policy, USACM responds to requests for information and technical expertise from U.S. government agencies and departments, seeks to influence relevant U.S. government policies on behalf of the computing community and the public, and provides information to ACM on relevant U.S. government activities. USACM also identifies potentially significant technical and public policy issues and brings them to the attention of ACM and the public.

More information about ACM may be found on the World Wide Web at http://www.acm.org, and information on USACM may be found at http://www.acm.org/usacm.

Table of Contents

"An adequate and effective registration will go far toward assuring honesty and fairness in the conduct of elections. Upon the honest and faithful maintenance of the registration books depends the purity of the ballot box. And upon the purity of the ballot box depends the success or failure of our democratic form of government."

-- *Registration of Voters in Louisiana*, Alden L. Powell and Emmett Asseff, Bureau of Government Research, Louisiana State University, 1951

Executive Summary

The voter registration process may seem simple to most voters. They give their names, addresses, birth date, and in some cases party affiliations to election officials with the expectation that they will be able to vote on Election Day. In reality, election officials must oversee a complex system managing this process. They must ensure that the voters' information is accurately recorded and maintained, that the system is transparent while voter information is kept private and secure from unauthorized access, and that poll workers can access this information on Election Day to determine whether or not any given voter is eligible. A well-managed voter registration system is vital for ensuring public confidence in elections.

State and local governments have managed voter registration using different approaches among different jurisdictions. In 2002, Congress sought to make these disparate efforts more uniform by passing the Help America Vote Act, which required that each state have a computerized statewide voter registration database. In implementing this mandate, state and local governments still have differing approaches, but it is clear that information technology underpins each of their efforts. While technology will help election officials manage this complex system, it also creates new risks that must be addressed.

This study focuses on five areas that election officials should address when creating statewide voter registration databases (VRDs): accuracy, privacy, usability, security, and reliability. Each chapter contains detailed discussions and recommendations. The following are some of the overarching goals for VRDs and selected recommendations for achieving them.

1. **The policies and practices of entire voting registration systems, including those that govern VRDs, should be transparent both internally and externally.**

VRDs control access to voting; therefore, they have a direct impact on the fairness of elections, as well as the public's perception of fairness. It must be possible to convince voters, political parties, politicians, academics, the press, and others that VRDs are correct and are operating appropriately. Internal procedures and interfaces also must be clear to election workers in order to minimize errors. Transparency can be provided by allowing voters to verify their voter registration status and data; publicly disclosing outside data sources that officials use for verification; indefinitely keeping a secure write-once VRD archive in electronic form to allow audits of previous elections; and using independent experts to audit and review VRD security policies. Other goals such as accountability, audits, and notification also support transparency and are discussed below.

2. Accountability should be apparent throughout each VRD.

It should be clear who is proposing, making, or approving changes to the data, the system, or its policies. Security policies are an important tool for ensuring accountability. For example, access control policies can be structured to restrict actions of certain groups or individual users of the system. Further, users' actions can be logged using audit trails (discussed below). Accountability also should extend to external uses of VRD data. For example, state and local officials should require recipients of data from VRDs to sign use agreements consistent with the government's official policies and procedures.

3. Audit trails should be employed throughout the VRD.

VRDs that can be independently verified, checked, and proven to be fair will increase voter confidence and help avoid litigation. Audit trails are important for independent verification, which, in turn, makes the system more transparent and provides a mechanism for accountability. They should include records of data changes, configuration changes, security policy changes, and database design changes. The trails may be independent records for each part of the VRD, but they should include both who made the change and who approved the change.

4. Privacy values should be a fundamental part of the VRD, not an afterthought.

Privacy policies for voter registration activities should be based on Fair Information Practices (FIPs), which are a set of principles for addressing concerns about information privacy. FIPs typically address collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability. There are many ways to implement good privacy policies. For example, we recommend that government both limit collection to only the data required for proper registration and explain why each piece of personal information is necessary. Further, privacy policies should be published and widely distributed, and the public should be given an opportunity to comment on any changes.

5. Registration systems should have strong notification policies.

Voters should be informed about their status, election information, privacy policies of the government, and security issues. As with audit trails, notification procedures can improve transparency; however, they are not always widely embraced. A recent survey found that approximately two-thirds of surveyed states do not notify voters who have been purged from election rolls. Voters should be notified by mail about their polling places, any changes that may affect their ability to vote, or any security breaches that expose private data.

6. **Election officials should rigorously test the usability, security and reliability of VRDs while they are being designed and while they are in use.**

Testing is a critical tool that can reveal that "real-world" poll workers find interfaces confusing and unusable, expose security flaws in the system, or that the system is likely to fail under the stress of Election Day. All of these issues, if caught before they are problems through testing will reduce voter fraud and the disenfranchisement of legitimate voters. We recommend many different ways to test various aspects of VRDs throughout the report. Examples include, evaluation of VRD interfaces by laypersons and experts for consistency, feedback, and error handling; testing interfaces with real-world users and conditions, including extreme or sub-optimal conditions such as high processor load or network congestion; and allowing thorough, independent evaluations of the security and reliability of the VRD.

7. **Election officials should develop strategies for coping with potential Election Day failures of electronic registration databases.**

VRDs are complex systems. It is likely that one or more aspects of the technology will fail at some point. Different strategies can be employed to adjust for various failures. For example, Election Day verifications can be done via any of the following: paper systems, personal computers or hand-held devices with DVD-ROMs or other methods of holding static copies of the voter list, or via personal computers or hand-held devices connected by electronic communication links to central VRDs. Regardless of the method used, a fallback process should be devised to deal with a VRD failure. When appropriate, these processes should operate in tandem with provisional balloting and other measures designed to protect the voters' right to vote.

8. **Election officials should develop special procedures and protections to handle large-scale merges with and purges of the VRD.**

One of HAVA's main requirements is that VRDs be coordinated with other state databases (such as motor vehicle records). Ensuring that voter records reflect up-to-date information from other databases can improve the accuracy of VRD, but coordination can introduce errors from the same databases, thereby undermining accuracy. Because large-scale merges and purges can render voters ineligible, the action should only be performed by a senior election official with procedures that force some sort of manual review of the changes. Further, if large-scale purges occur, they should be done well in advance of any election, and anyone purged from the database should receive notification so that any errors can be corrected.

Conclusion. State and local election officials face an ongoing and challenging task in creating and implementing statewide voter registration databases. We hope that the discussion and recommendations in this report will help inform officials and the public on how to meet these challenges.

In issuing this report, we recognize that many states have been working diligently toward meeting the federal requirement to have an operational statewide VRD. Both because many states will not meet this deadline, and because there will be ongoing maintenance and changes to any such system, state and local governments will also face the issues identified in this report well beyond the federal deadline. For this reason, we offer our continued guidance to officials who may wish to discuss any of the topics raised in this report.

2. Accuracy

Maintaining the accuracy of VRDs requires balancing two opposing concerns. The first concern is that a VRD needs to be inclusive to avoid disenfranchising legitimate voters. The names of all people who have registered and are duly eligible to vote must be included in the VRD; any omissions will exclude eligible voters from voting. The second, somewhat contrary concern is that the VRD must not be overly inclusive. To prevent fraud, only legally registered persons should be listed in the VRD as eligible to vote. We will address both of these concerns.

Not only must VRDs be accurate, the public must also believe that they are accurate. Because VRDs control access to voting, transparency is critical. It must be possible to convince those with interests in elections—including voters, political parties, politicians, academics, and the press—of the correctness of the VRDs. To provide transparency, policies should minimize the possibility of error and facilitate the correction of errors. Election officials must also take responsibility for ensuring adherence to these policies.

Data Entry and Errors. Most errors in individual database records occur during data entry. Errors include misspelling of names and addresses, incorrect recording of unique IDs, misidentification of people to whom access to the system should be allowed or denied, and misdirecting voters to the wrong polling place.

Data is entered into the VRD using one of two methods: manual entry or via automatic scanning devices. An automatic scanning device is a machine that looks like a copier and is used to scan a document into a computer system. Once the document is scanned in, software that can recognize characters transfers the data from the printed form into the VRD, while providing a clerk with the opportunity to correct mistakes. For either manual entry or automatic scanning, a well-designed user interface for the clerk will reduce errors. (Chapter 4 on usability contains further discussion of user interfaces.)

While quality control systems and appropriate supervision of data entry may reduce data entry errors, some errors will inevitably occur. Problems can arise because of variations of name spellings (Stevens or Stephens), first and last names that use accent marks or more than one capital letter (McMullen), and names that have no vowels (Ng). Incorrect or incomplete spellings of street names are additional potential sources of errors. Changes that are primarily entered in other state databases—such as changes in marital status and court approved name changes—also compound the challenge to accuracy.

Voter Verification and Notice. To minimize the impact of errors in the VRD, voters should be provided with (1) opportunities and methods to view and verify their data, and (2) notices about changes to their records. For example, the system might provide an Internet website or automated telephone service where voters can examine parts of their records, check their registration status, and determine their assigned polling places.

Whenever a voter or potential voter is determined to be ineligible to vote, the reason and source of information for the determination of ineligibility should be included in the VRD. This information should be retained so that someone who has been inappropriately labeled as ineligible can easily challenge the decision and demonstrate that an error has occurred.

Finally, election officials should mail each registered voter in the VRD a postcard with his or her registration information and information necessary for voting, such as polling place location or instruction for voting by mail. Voters also should be notified when their registration status changes. A voter removed from the rolls or reassigned to a new polling place should be notified by mail of the change and be provided an opportunity to seek correction if the change is an error. A voter recorded as having moved should be notified by mail sent to both the new address and the old address (similar to the method the United States Postal Service uses with respect to change of address forms).

To help correct errors in voting records, contact information for the person or office responsible for complaints and questions should be provided to voters. Further, voters and system administrators should understand how complaints and errors are addressed, and voters should receive feedback explaining the reasons for a final determination.

One recent survey found that approximately two-thirds of surveyed states do not notify voters who have been purged from the election rolls.[13] Advance notice, which can be facilitated by the VRD, would provide voters with an opportunity to identify mistakes prior to an election. Care must be taken in designing such systems so that violations of privacy and security do not occur.

Notification processes are not always foolproof. For example, in 2004, 8,800 Maricopa County, Arizona, residents received election notification cards listing the wrong polling places in the wrong cities.[14]

To help minimize the impact of incorrect notification, we recommend that public notice be provided well in advance of an election. That notice should include the polling place's geographic location and official name (school, church, library name), a description of the exterior of the polling place to assist voters in locating the entrance, times of poll operation, residential boundary lines, and corresponding zip codes.

Some states allow voters to verify that they are registered through an Internet web site or by phone. For states that use Internet verification the user interface should protect voters' privacy by requiring the voter to provide his or her name and address and limiting the response to "yes, you are registered to vote and here is where you go" or "no, you are not registered to vote." The response should not include personally identifiable information about the potential voter.

Some provision needs to be made to deal with corrections on Election Day because not all errors can be corrected in advance. Poll workers are likely to be preoccupied with running an election and should not be allowed to make changes to the VRD. Under the right circumstances, after extensive testing for accuracy and usability, it might be possible to allow poll workers to send electronic reports of needed changes to election workers. If such a system is implemented, the updates would need to satisfy the auditing and authorization requirements discussed elsewhere in this report.

A simple alternative is to provide paper forms that are filled out at the polling place and submitted to election workers after the close of the election.

Generating the List of Registered Voters. A printed voter registration list for those precincts served by a polling place is typically used to verify registered voters. While

[13] Electionline.org, op. cit.

[14] Dennis Wagner, 2004, "8,800 Voting Cards Have Wrong Poll Address," *The Arizona Republic*, October 27, p. B5.

these printed lists are convenient and easy to control, sometimes the wrong list is provided to a polling place. To minimize the chance of the delivery of an incorrect list, we recommend that automated generation of polling place lists be used as much as possible and that the lists be carefully checked by at least two people. Local officials can conduct these checks, but they need to be made far enough in advance of elections to allow time for corrections.

Incorrect voter lists could be delivered to polling places independent of whether the data are provided on paper, DVD-ROMs, in a PC, or in a handheld device. In all of these cases, a computer operator might provide incorrect directions to the computer, resulting in the wrong electronic list going to the polling place. As with paper printouts, we recommend that electronic versions of voter lists be checked by at least two people well in advance of elections to allow time for corrections.

Information Deletion and Retention. In addition to being a list of currently registered voters, a VRD is a comprehensive set of records reflecting voter registration activity and administration. Consequently, we recommend that after records appear to be no longer relevant, they be retained in the VRD at least for the next two Federal elections or for the statutorily-mandated minimum of twenty-two months.[15] The retained record should include a dated annotation stating that the voter is not eligible to vote, along with the reason for ineligibility. Thus, a VRD might contain information about those who have applied, been approved, been questioned, died, moved, or been denied the right to vote, as well as those who currently are eligible to vote.

When records were stored on paper, retaining old records imposed a non-trivial administrative burden. Electronic databases have made the cost of retention negligible, so old information can be retained relatively easily and inexpensively. When information is sufficiently old, it should be moved from the VRD into an offline archival database that is never purged. Retention of such information will enhance transparency and facilitate the correction of errors such as those that can occur when voters are thought to have died, moved, been convicted of a felony, or otherwise determined not to be eligible to participate in a public election.

Other Databases. HAVA requires that states authenticate each potential voter by cross-checking with other state databases—in particular, databases of driver's licenses.[16] If a potential voter does not have a state driver's license, then the last four digits of the voter's Social Security number must be used for authentication.

Because other databases can be inaccurate as a result of ambiguous or incorrectly entered data or computer-related problems, wholly automated procedures are risky. Consequently, we recommend that other databases not be used to enroll or de-enroll voters automatically. External databases could be used for initial screening, but an appropriate election official should perform any final determination of voter eligibility or

[15] The Civil Rights Act of 1960 requires that every officer of elections retain for 22 months registration and other voting records and papers for federal elections. 42 U.S.C. § 1974.

[16] HAVA provides for coordination of voters lists with other state agency databases (42 U.S.C. § 15483(a)(1)(A)(iv)) and requires that registration applications include either a current and valid driver's license or the last 4 digits of the applicant's Social Security number (42 U.S.C. § 15483(a)(5)).

ineligibility. We suggest that every change, addition, or deletion to the VRD require explicit approval by an individual authorized to make that change. We discuss how this might be done in Chapter 5 on security.

Errors can arise because of court-approved changes in legal name that conflict with existing birth records, motor vehicle records, or other state records. Name similarities also can create problems. For example, a death record database may show that Mr. John Smith who lives at 254 Vine St. has died. There may be a Mr. John Smith, Jr. living at the same address who is eligible to vote. If the death record database is applied with no cross checking, John Smith Jr. may learn on Election Day that he has been denied his right to vote.

Databases also can be inaccurate or unreliable because of computer viruses, programming errors, and system failures. For example, in 2003 the Maryland Motor Vehicle Administration (MVA) offices were attacked by a computer worm.[17] The worm shut down the MVA's computers and telecommunication systems, cutting them off from all forms of remote communication and disrupting operations in all 23 MVA offices located throughout the state. A second event occurred on January 20, 2004, when the MVA could not process work on the mainframe computer for about an hour after opening. The problem was characterized as a computer glitch.[18]

A further risk to the accuracy of databases is insider fraud, involving either the VRD itself or external databases, such as driver's license databases, that are used to authenticate voters.[19] Therefore, election officials should carefully consider if the accuracy and security of external databases is sufficient to meet voter registration needs. Risks associated with insider fraud are discussed further in Chapter 5 on security.

Avoid Large-Scale Merges and Purges. Computers make it easy to automate sweeping batch updates to a VRD; at the same time, errors can be magnified by the use of automation. In the context of VRDs, a batch update is a group of updates received from what is believed to be an authorized source (e.g., a local county). Because many voter records could be affected by a single batch transaction, a greater level of authority should be required to perform a batch update than is required to make individual changes. As is the case with all updates, election officials should develop policies and procedures to ensure the accuracy of large batch updates to the VRD. For example, a policy might prohibit batch updates affecting more than a maximum number of voters or jurisdictions (essentially requiring that large changes be broken down into multiple smaller batches that can be reviewed more effectively), or a policy might require individualized review and approval of each voter record that is affected. A policy might specify that batch updates be reviewed by several people or mandate that audits of a statistically-significant

[17] Christian Davenport and Hamil R. Harris, 2003, "MD's MVA Offices Forced to Shut Down," *Washington Post*, August 13, p. A09.

[18] "Glitch at MVA Branch Offices Delays Some Transactions for an Hour," 2004, *The Baltimore Sun*, January 21, p. B6.

[19] For example, a Maryland MVA employee was charged with conspiring with others to sell more than 150 state identification cards. See Eric Rich, 2005, "MD, MVA Employee Charged in ID Card Sales," *Washington Post*, April 23, p. B03. For a collection of stories of security problems of motor vehicle records, see Center for Democracy and Technology, *Tracking Security at State Motor Vehicle Offices*, available online at http://www.cdt.org/privacy/030131motorvehicle.shtml.

random sample of records in the batch be performed before approving the batch update.

Given the inaccuracies that exist in many governmental databases, large-scale automated merges between databases increase the risk of errors in a VRD.[20] Consequences of inaccuracies in other databases could result in the widespread disenfranchisement of eligible voters, the inclusion of ineligible voters in a VRD, or both.

We recommend special caution in deploying large-scale purges of VRDs. The move to a statewide VRD may make it tempting to attempt to automatically eliminate duplicates by comparing lists of eligible voters across counties, something that previously could not be done. However, automatic purges of duplicate entries could disenfranchise large numbers of legitimate voters. If large-scale purges occur, they should be done well in advance of any election, and all people whose names are purged from the VRD should receive notification in sufficient time for them to be able to correct any errors arising from the purge.

Accountability. Clearly defined accountability for all changes to the database is a fundamental requirement for helping instill voter confidence in VRDs. Voters, politicians, election officials, the press, and others should be able to determine who is responsible for changes to the VRD.

These changes include, changes to the data such as adding new voters, purging voter records, changing addresses, names, etc.; changes to the software configuration such as incorporating new software releases into the VRD; changes to the security policy and access rights; or changes to the database design. Any of these changes can adversely affect the data, so in order to provide the desired accountability there must be a record of each change, when it occurred, and who approved the change.

Audit Trail. The record of the changes to the VRD is called an *audit trail*. In order to ensure accuracy and transparency, VRDs must be auditable. VRDs that can be independently verified, checked, and proven to be fair will increase voter confidence and help avoid litigation.

The audit trail should include the record of all possible changes mentioned, namely, data changes, configuration changes, security policy changes, and database design changes. Although we call this an audit trail, it is not a single entity. The records of configuration, policy and design changes, including who approved them, can be kept in computer files or on paper as long as they are auditable by a third party. The record of changes to the data, because there will be many of them, must be kept in computer files to facilitate auditing.

In DBMS applications, there are typically two files generated because of a change to the database. The *transaction log* records in a file the data values before and after the change occurred, as well as the time of the change. The *audit log* records information about the user ID of the person who made the change. The transaction log is used to provide backup should a system failure occur.

The content of audit logs varies among DBMSs. In some, it is possible to configure the system so that the audit log tracks changes to the security of the system (the

[20] In 1988, Congress enacted the Computer Matching and Privacy Protection Act to address some of the unfairness and inaccuracies arising from federal government use of computer matching techniques. See Public Law 100-503, 102 Stat. 2507 (codified at 5 U.S.C. §552a).

permissions given to particular users), changes to the data, and changes to the database design. For the purposes of the VRD auditing requirements, this is not sufficient. The VRD should record not only which user made the change, but also the identification of the person who authorized the change. Therefore, it may not be possible to rely on the commercial DBMS's auditing capabilities alone for the audit trail that a VRD requires. VRD implementers will need to augment the application code of the commercial database audit log to provide a complete audit trail.

Well-maintained audit trails are critical because they may allow reconstruction of the circumstances of a system failure, thereby facilitating future improvements to access policies and possibly to the database itself.

Approval Mechanism. Given that there is an audit trail that records whose approval was given for each change, state or local officials must set policies on who is actually authorized to make changes. Access control polices are discussed in more detail in Chapter 5 on security. We assume that the person with ultimate authority to make the changes is an election official, and we recommend that the responsibilities and authorities of such election officials be clearly defined and publicly available.

For system changes, we recommend that there be a formal change control process that states how changes to the system configuration, security policy, and database design are reviewed, approved, and recorded.

Summary reports or excerpts from audit trails should be provided to supervisors and made available to external auditors. These reports should be inspected frequently for unusual or suspicious activities such as access from unexpected Internet Protocol (commonly referred to as "IP") addresses or at unusual times of day, surges in the number of accesses by a single user, and other anomalous activity.

Conclusion. Well-designed accuracy features must be accompanied by appropriate training and resources. Even the best designed VRD will be of little value if officials do not monitor and verify that only authorized changes are made to the VRD. Log files that are never read and system quality control processes that are not supervised will not ensure database accuracy. Since accuracy should be viewed as an ongoing responsibility, election officials should assign specific staff to oversee these continuing activities.

COMPUTERS AND BIOMEDICAL RESEARCH 25, 486–509 (1992)

Tolerating Spelling Errors during Patient Validation

CAROL FRIEDMAN* AND ROBERT SIDELI†

*Queens College of the City University of New York, Flushing, New York 11367-0904;
and †Center for Medical Informatics, Columbia-Presbyterian Medical Center
New York, NewYork 10032

Received July 9, 1991

Misspellings, typographical errors, and variant name forms present a considerable problem for a Clinical Information System when validating patient data. Algorithms to correct these types of errors are being used, but they are based either on a study of frequent types of errors associated with general words in an English text rather than types of errors associated with the spelling of names, or on errors that are phonologically based. This paper investigates the types of errors that are specifically associated with the spelling of patient names, and proposes an algorithm that effectively handles such types of errors. This paper also studies the effectiveness of several relaxation techniques and compares them with the one that is being proposed. © 1992 Academic Press, Inc.

1. INTRODUCTION

In a Clinical Information System (CIS), it is crucial to take measures to ensure that the correct association is being made between the clinical data and the patient. Such a system usually includes a patient registry component, which associates patient identification information with unique numbers which we call patient identification numbers (PID). Once the patient is assigned a PID, whenever he or she has a clinical encounter (including future readmissions and clinic visits), the clinical data from that encounter is recorded using the patient's PID, along with the patient's name and possibly other additional identifying information, such as sex and date of birth. Typically, validation is achieved by matching the name in the encounter record with the name in the registry database record associated with the PID given in the encounter. If the names do not match, there is a possible identification error, and the clinical data is rejected by the system.

While matching name pairs seems like a simple task, this procedure is complicated by the phenomenon that very often the names of patients are not spelled exactly the same on each clinical encounter. In a preliminary study of spelling discrepancies in patient names, we found that there were 27,000 (27%) name mismatches out of 100,000 patient records in the Radiology database. The results of two subsequent studies were consistent with these findings, and we found that the rate of name mismatches ranged from 23.1 to 36.5% for the two other

0010-4809/92 $5.00

departments (Pathology and Medical Records discharge summaries) studied. This situation occurred because of spelling errors, spelling variations, and typographical errors made by both healthcare personnel and patients.

Columbia Presbyterian Medical Center (CPMC) has a centralized Clinical Information System which obtains departmental patient data by uploading the data from the departmental databases. During the upload process the data is validated and transformed into a form which is consistent with the centralized database so that the clinical information is made accessible online to authorized healthcare workers throughout the facility. A component of the system includes a patient registry, which contains demographic patient data and associates unique PIDs with each patient. Presently, the departmental systems at CPMC are loosely coupled to the CIS; for the most part they were developed independently of the centralized system and do not interface with the registry database to check that the patient's PID and name match the patient registry. The validity of clinical data from these systems must therefore be checked by the CIS during the upload process. If the identification validation process occurred earlier by the departmental systems, the name mismatch problem would still occur but would be detected at an earlier stage by the individual departments, and the correction task would be relegated to them. In this situation, the correction of the errors would be handled in a nonuniform manner and standards would be difficult to enforce. Although validating clinical data during the upload process is also an important issue, in this paper we discuss only the issue of validating the patient's identity.

A simple identification validation approach we initially tried to use was to reject data from those reports where there are name mismatches between the registry and the report records. However, many (but not all) of the name mismatches were due to minor errors or variations in spelling the name of the same patient. Because this situation occurred so frequently, this procedure caused too many patient records to be erroneously excluded from the CIS system. The consequence of excluding the records was very detrimental to the system because a noticeable amount of clinical data was made unavailable. It was quickly evident that this simple approach had to be replaced by a more sophisticated validation procedure that tolerated so-called *minor* discrepancies while rejecting errors that were likely to be identification errors.

In this paper we present an automated spelling relaxation algorithm that effectively handles name mismatches that occur during patient validation. Although this technique is applied in a healthcare setting, it is also applicable to information systems where a person is assigned a unique identifier number within the system. The method is based on computing a similarity measure for a mismatched name pair. A pair with a similarity measure greater than a certain threshold is accepted, whereas a pair with a similarity below the threshold is rejected. This paper also discusses the results of two related studies: one study discusses the strengths and limitations of other automated procedures which tolerate spelling errors, and compares those methods with the one we have developed. The second study analyzes the types of name mismatches that generally occur in clinical visits. We felt this was essential because the spelling toleration algorithms used for correcting misspellings of names are usually based on a study by Damerau (*6*) of common types of spelling errors that occur in words. Our preliminary observations on a training set of data led us to believe that the common types of spelling errors for names are not the same as those for words in general.

The training set of data consisted of 1000 patient records from the Department of Radiology. The test set consisted of 14,793 patient reports from the Department of Pathology and 10,000 reports from the Medical Records discharge summary. The PIDs associated with the reports were used to extract the corresponding names from the CIS patient registry. The number of PIDs that were actually found in the patient registry was used to determine the rejection rate. Each name pair, consisting of the name from the departmental report and the name from the registry database, was subjected to several comparison methods, the findings of which are presented in this paper. Additional identification information, such as date of birth and sex, are available in the registry database and some of the departmental databases, but analysis of the training set showed that the quality of this data was too poor to be used, and therefore only name comparisons were studied. For the second study, a detailed manual analysis of 560 randomly chosen name pair mismatches was performed in order to categorize the common types of errors and to determine their frequency.

Section 2 presents an analysis of spelling errors that typically occur in words and compares them with the results of our study, which analyzes the types of errors that occur with names. It also discusses the different circumstances associated with general spelling errors from those associated with validating the identification of a patient record in a departmental report. Section 3 describes algorithms that are commonly used to correct or tolerate name misspellings, and Section 4 presents a detailed description of the method we have developed. Section 5 reports on the results of the study we performed comparing several different error relaxation algorithms, and discusses the effectiveness, efficiency, and limitations of the various methods.

2. An Analysis of Spelling Errors

2.1 Spelling Errors in Words

In a review by Peterson (*19*) of spelling checkers and correctors, the techniques discussed were all aimed at correcting four of the most common types of spelling errors. The determination of these errors was based on an investigation of types of spelling errors by Damerau (*6*), who found that over 80% of spelling errors fell into one of four categories:

1. One extra letter (insertion)
2. One missing letter (deletion)
3. One wrong letter (substitution)
4. Transposition of two adjacent letters (transposition)

The errors noted above are all cases of *single errors*. Two more recent spelling correcting techniques (*18, 3*) were subsequently presented which also are based solely on detecting and correcting the types of spelling errors noted by Damerau. In particular, the technique proposed by Bickel (*3*) is directed toward the problem of automatically correcting misspelled names.

Since the spelling correction techniques cited above are all aimed at correcting the types of common errors noted by Damerau, we felt that a preliminary study was necessary to determine whether or not the types of spelling errors that occur in words of texts associated with standard English and the types of spelling errors that occur with patient names in clinical encounters (at CPMC) are the same. If there is a significant difference between the types of errors in the different environments, then there is a built-in inherent limitation on the effectiveness of correction techniques based on common errors in English words when applied to correcting names in the clinical environment.

2.2. Spelling Errors in Names

In this section we discuss the findings of a study we performed to analyze the type of spelling discrepancies in patient names that typically occur at CPMC, a large-scale healthcare facility. A set of 560 randomly chosen name pair mismatches that were not considered identification errors were manually analyzed. The name obtained from the registry database was considered the correct name, and the one obtained from the departmental report was considered the one with an error in spelling. In our study of mismatches occurring with names, we found that there were different types of spelling errors associated with names than with text words in general. The types of spelling errors were categorized and their frequencies noted. The errors were categorized by manual inspection of a list containing pairs of mismatched names. We found that the most frequent types of errors in names are as follows (the types of errors are shown below in order of their frequency):

1. Insertion of additional names, initials, and titles (36.4%)
   Smith, Mary; Smith, Mary Ann
   Smith, John; Smith, John Jr.
2. Several letters of the name are different due to nicknames and slight spelling variations (13.9%)
   Nicholas; Nick; Nickie; Nicky
3. One letter is different (13.7%)
   Nicholas; Nickolas
4. One letter added or deleted (12.9%)
   Gomnez, Gomez
5. Differences due to punctuation marks and number of blanks (11.8%)
   O'Connor; O Connor; OConnor
6. Different last name for female patients (7.8%). This typically occurs because the patient's name is changed when she gets married
   Gomez, Ann; Vega, Ann

TABLE 1

CATEGORIES OF NAME MISMATCHES

Type of spelling error	Frequency	Percentage
1. Extra name or title	215	36.4
2. Several letters different	82	13.9
3. 1 Letter different	81	13.7
4. 1 Letter added or deleted	76	12.9
5. Punctuation and blanks	70	11.8
6. Different last name for female patient	46	7.8
7. Permutation of parts of complete name	8	1.4
8. Different first name	8	1.4
9. Permutation of 1 letter	5	0.8
Total	591	

7. Parts of the name are permuted (1.4%)
   Gomez, Ann; Ann, Gomez
8. Different first name (1.4%)
   Smith, Helen; Smith, Ellen
9. Permutation of 1 letter (0.8%).
   Robrets, Bill; Roberts, Bill

Table 1 shows a summary of the different categories and their frequencies. Notice that the frequency rates are based on the total number of errors found instead of on the total number of mismatches. There are more errors than mismatches because occasionally more than one error is associated with a mismatched pair.

According to the results of our study, the *single letter* types of spelling errors occurring with names (types 3, 4, 5, and 9) account for only 39.2% of the spelling errors, whereas, according to the study by Damerau (*6*), these same types of errors account for 80% of the spelling errors in words. Although there is no special category in Damerau's study that is equivalent to the type of error associated with punctuation marks and blanks (i.e., type 5), this type of error is mainly a single letter type of error, and therefore we include it in a group of single letter errors. According to our study, multiple letter types of spelling errors in names (types 1, 2, 6, 7, and 8) total 60.9% of the errors. These findings show that there are significant differences between the two applications, and that an algorithm designed for handling single letter types of spelling errors would be effective only in approximately 39.2% of the name mismatches.

It should be noted that this study is based on the environment of CPMC. Several factors are present at CPMC that may not be typical of other healthcare facilities, and therefore the findings of our study may be valid only in similar environments. In CPMC, one factor that probably increases the frequency of spelling errors in patient names is due to a patient population consisting of a broad range of ethnic diversities. It is generally more difficult for a native speaker to spell a name correctly which is of foreign origin. For example, it is often difficult for native speakers to:

- Differentiate between first and last names—*Yung Hsien Sun; Hsien Yung Sun*
- Spell long unfamiliar names or detect errors in their spelling—*Panagiotakopoulis,* and *Kyzwieslowski*
- Be consistent in spelling because foreign pronunciation varies—*Fen, Phen, Ven*

Another factor that probably effects the types and frequencies of spelling errors is that CPMC is a large-scale busy facility with a transient patient population and therefore the names of patients are generally unknown to the departmental personnel that handle the record keeping.

2.3. Spelling Errors in Text versus Validation Spelling Errors

The problem we are addressing, the mismatching of two names during validation of a patient's report, is similar to the problem of detecting and correcting general spelling errors that occur in English text, but there are important differences. In the latter situation, there is a reference set consisting of correct names or words: a spelling error is detected when a name or word does not match *any* name or word in the reference set. In that case, a correction consists of finding a name(s) that is *close* to the one given. A very similar situation occurs in the clinical environment during patient admission in order to find the PID of a previously admitted patient by using the patient's name to retrieve a list of patient names in the registry that are close to the given name. Once these names are found, additional demographic information is used to identify the patient. This type of procedure would also be used to avoid giving a patient a redundant PID. When admitting a seemingly new patient, a registry name check can be made to verify that the patient is not already registered; if he/she is, the patient already was assigned a PID which should be used. In these cases, the reference set is the set of patient names in the system rather than a set of words.

The above process is typically performed during text editing or patient admission. It is usually a semiautomatic process because a list of possible corrections is suggested to the user who considers which one to choose, if any. The algorithm we are proposing could be applied to this situation, but it would be very inefficient because it is based on computing a similarity measure between a pair of names. Using that technique, the patient name must be compared with each member in the reference set and a similarity measure must be computed for each name pair. Only then will the algorithm be able to suggest names which, when paired with the given name, have the highest similarity measures to the given name. This necessitates that the entire name space be searched, and a string matching algorithm be applied to each name in the space. For this situation, other algorithms, such as the Russell Soundex Code or algorithms containing other hash-coding schemes, are more efficient because in these, the search space is considerably narrowed. When using a hash-coding scheme, a code is computed for each name independently of other names in the name space, and only those names associated with the same code are considered *similar* to the given name.

In the case of name validation the situation is considerably different because the detection of an error occurs when there is a mismatch between the name (given along with the PID) of the person in the report and the name associated with the matching PID in the registry database. Because there is only one name pair involved in this situation, a string comparison algorithm is not inefficient since the pattern-matching procedure has to be performed only once. Tolerating the error is a completely automatic procedure because mismatches occur during uploads, in which case there are no users to interface with; the errors must be automatically tolerated or rejected. In this situation, the toleration of a mismatch signifies that the mismatch is *not likely* to be an identification error, and that it is *probably* correct to associate clinical data from the encounter with the PID of the patient in the registry database. However, whenever this situation occurs there is still a small possibility that the mismatch is really due to an identification error. Therefore, whenever there is a name mismatch that is tolerated, information is recorded along with the clinical record signifying that a name mismatch has occurred that was tolerated by the CIS. Furthermore, the actual name occurring in the departmental record is included so that in addition to the warning message which is generated during result reporting, a clinical user may see the actual name associated with the report. Even more importantly, because clinical information is also used by the CIS for decision support purposes, any clinical advice or alert which is generated by the decision support component that involves clinical data associated with a name mismatch that has been tolerated will also be noted.

Clinical information from mismatches which are considered likely to be actual identification errors are never uploaded to the central database, but are sent back to the appropriate department to be corrected. In this scenario, true rejections are invaluable because they have prevented violations of the integrity of the patient database. However, false rejections are very costly, not only because they are an extra burden to departmental personnel, but also because the clinical data associated with the report is not available for general online access until the discrepancy is manually corrected by the department and the correction uploaded once again to the CIS. False acceptances are also very costly, because they associate clinical data with the wrong patient, which is unacceptable in the clinical environment.

3. Related Work

3.1 Soundex Based Methods

The Russell Soundex method (*15, 21*) is one of the best known methods which performs phonetic reductions of names. In this technique, each name reduces to a code. The underlying principle is that names that sound alike should reduce to the same code. This technique is specifically aimed at situations where a person's name is given verbally instead of in written form. In those cases, typical spelling errors consist of someone mistakenly substituting letters that sound alike for the correct letters of the name.

The Soundex algorithm is efficient timewise, particularly for those cases where names are used to retrieve records from a database. If a numeric code is computed for each name in the database, then a secondary index may be maintained that relates a set of names to a particular code. When a given name cannot be found in the database, names in the database which hash to the same code may be supplied by the database system, and using other demographic information, an attempt can be made to find the correct name from the list of potential candidates.

This method is also useful for tolerating spelling errors when validating patients, because two names that hash to the same code are considered sound alikes, and those type of mismatches are tolerated.

The Soundex code for a name as given by Knuth (*15*) consists of the initial letter of the surname plus three digits derived from the remaining letters of the surname. The number of digits used may be varied to four or five, but usually it is not greater than five. The coding scheme is as follows:

- All vowels, and the letters *H*, *W*, and *Y* are dropped
- The following letters compute to the following digits:
  —1. B, F, P, V
  —2. C, G, J, K, Q, S, X, Y, Z
  —3. D, T
  —4. L
  —5. M, N
  —6. R
- All consecutive repeating digits are ignored
- If there are less than three digits, add trailing zeros so that the code consists of a letter and exactly three digits.

Thus, using this method, a maximum of 26 × 7 × 7 × 7 (8918) different codes can be obtained. This means that many names may correspond to one particular code. Furthermore, since the first letter of the name is always the first digit of the code, more weight is given to the first letter of the name than to the other letters. In particular, the first letters of both names in the name pair *must* match or their Soundex codes will be different.

The Soundex method is based on similarities in sound between groups of letters in the English alphabet, and therefore it is basically applicable to words or names corresponding to English. To maximize its effectiveness, the assignment of digits for the different letters of the alphabet has to be adapted for different languages. However, not much can be done to adapt the method if names in the registry stem from different languages because the nationalities of the patient population are very diverse.

The effectiveness of the Soundex code in tolerating spelling errors in medical words is reported in a study by Joseph and Wong (*12*) who examined the method with regard to different types of errors. They also analyzed differences between the medical vocabulary and general English vocabulary and noted that a large number of medical words were derived from Latin or Greek roots. They experimented with several variations of the Soundex method and found that the rate of words matched ranged from 58.7 to 65.2%. They found that Soundex was not tolerant enough for finding a set of similar words, and that a considerable number of failures resulted because the error occurred on the first letter of the word. Two other studies by Greenfield (*11*) and Goehring (*10*) also found that the Soundex method had a low tolerance for spelling errors. In the latter study, the patient population was distributed among several nationalities, the majority of which were European. Our study, which is discussed in Section 5, found that the rate of matching name pairs using the Soundex method during validation was much better than the above three studies reported (i.e., the rate ranged from 88.5 to 93.4%), but the false rejection rate was still too high for our needs. This is discussed further in Section 5, which contains the results of the study which compares various techniques.

3.2. Variations of the Soundex Method

Other techniques have been developed using variations of the Soundex method. A method, called the Koeln Phonetic algorithm (*20, 8*) uses the full length of the name to generate the code. A partial study of error tolerance of that method was reported in (*10*). The study looked only at the two error types which correspond to deletion of a character and transposition of two adjacent characters. For a name consisting of nine characters, the probability of accepting the same name in spite of a deletion error using the Koeln Phonetic method, is only 1%, and that of tolerating a transposition error is 20%. The probability of generating the same code for those types of errors decreased dramatically as the length of the name increased. These statistics demonstrate that the Koeln Phoenetic code would be highly ineffective for name validation.

Three other variations of the Soundex code are the Davidson's Consonent Code (*7*), the MEGADATS-2 code (*14*), and the match code (*10*). The first two codes are formed based on varations of the Soundex method, and they basically have the same strengths and weaknesses of the Soundex method with respect to tolerating spelling errors. Greenfield (*11*) discusses the Davidson method in more detail. The match code method is often used in MUMPS applications. In this method, a certain number of letters are selected from both the last and first name; these letters form a key used to identify the person. According to the study performed by Goehring (*10*) there is minimal error tolerance using the match code method.

3.3. Other Spelling Correction Techniques

A hash code method proposed by Mor and Fraenkel (*18*) is based on detecting and correcting the most frequent types of spelling errors that occur in words. This method does not use phonetic similarities, but is based on using combinations of deletion, exchange, and rotation operators to create the set of all possible single error misspellings from a reference set of words. This method would be impractical when applied to the patient population because the number of different patients in the registry database is very large, and the length of the average patient's name is at least 9 characters. This means for the average patient, there will be approximately 511 single letter misspellings generated. An even greater disadvantage of this method is that its effectiveness is very limited, because the majority of name misspellings are not the single letter types of errors that this method corrects for.

Another method proposed by Bickel (*3*) is specifically geared toward correcting misspelled names. This algorithm computes a likeness value for a pair of names based on the frequency of letters common to both names. The likeness value is created as follows: if a letter is present in both names, the weight of the letter is added to the likeness value. Each letter of the alphabet is assigned a weight depending on its frequency. For example, the weight of *z* is 9, whereas the weight of *s* is 3 because letters that are less frequent are given greater weights. This is based on the assumption that a letter that occurs less frequently carries more information. Each letter of the alphabet contributes to the likeness value only once, and therefore letters occurring more than once are ignored. The positions of the letters in the name are also ignored because transposition, deletion, and insertion type errors cause characters to be shifted. Names with the highest likeness values are potential corrections.

The effectiveness of this method was evaluated by Bickel using simulated runs where three different types of errors were automatically generated from a database of names and the algorithm was tested using the generated misspellings. The generated misspellings consisted of single deletions, single insertions, and exchanges of two letters. For these cases, the algorithm was found to be more effective than the Soundex method.

We did not use this algorithm for two reasons: it would not be effective for the most common types of error we were encountering (due to an extra name, initial, or title), and more importantly, it was not clear how the likeness measure could be utilized as a criterion for acceptance or rejection of a mismatch in a validation situation where there is only one pair of names. In this algorithm the likeness measure is meant to be compared with other likeness measures computed from all the names in the database, and names with the closest likeness measures are considered candidate matches. In particular, a single measure has no meaning by itself.

3.4. Molecular Sequence Comparison Algorithms

Another problem that shares many similarities with that of spelling relaxation is found in the area of automated nucleic acids and protein sequencing. In this area, pattern matching techniques are also very important. One typical procedure compares two different sequences of proteins and finds similar subsequences within the larger sequences; another procedure identifies minimal changes needed to transform one sequence into the other. Several papers discussing this problem are (*25*, *17*, *23*, *4*, *24*). Although this area has much in common with the one we are addressing, we found there are significant differences between the two, and therefore it was not practical to use the same algorithms.

One major difference is that permutations are not considered between two molecular sequences whereas permutations (or transpositions) are considered a basic problem in the misspelling of names. The effect of a permutation can be simulated by deleting elements of one sequence and inserting them into the other. However that is not the same as a single operation where a permutation is considered a primitive operation rather than a combination of two operations. Another significant difference is that there can be arbitrarily large gaps between elements as well as large gaps between similar regions in molecular sequencing. This is not practical in spelling relaxation or correction.

Although we do not use any of the sequencing algorithms per se, we use a pattern-matching algorithm which is based on similar principles because it utilizes dynamic programming techniques. This is discussed in Section 4.3.

4. Description of the Longest Common Substring Method

4.1. Heuristics

Before presenting the error toleration method we developed, it is important to mention several conditions which must be present in order for this algorithm to be appropriate:

- A registry database that uniquely associates a patient with an identification number must exist; once a patient is assigned such a number, it should always be used to identify the patient to the CIS.
- The information obtained from the clinical encounter includes the patient's PID and name; optionally the sex and date of birth may also be included.
- The name space (i.e., the set of all possible names) is exceptionally large (let n equal the size of the name space).

Suppose the above three conditions were true, and we were to match a pair of names, where the first name corresponds to the name associated with a valid PID chosen randomly from the set of all valid PIDs, and the other name consists of a name randomly chosen from the set of all possible names; the probability of a complete match between the pair of names would be very small. Let us suppose that Name X, corresponding to a valid PID, is chosen. The probability of randomly choosing X from the set of all possible names is $1/n$. Although we do not know the exact number of all possible names, we do know from our patient registry that there are at least 2000 unique patient names. We can use this number as a lower bound on the size of the name space, but we know its size is considerably larger because we must include in our set names from all the different regions in the United States and also from all the various countries in the world. The upper bound on the probability of a match is therefore 1/2000 or 0.05%, although realistically the probability of a match is substantially smaller than 0.05%. If we relaxed the exact match condition to consider a *partial match* consisting of one or more common substrings of the pair, then the probability of a *partial match* would be larger, but would vary depending on the size of the common substrings; if we required that the total portion that match be relatively large (for example, 97% of each name), the probability of a 97% match between the names would still be very small; if the common portion that matched were allowed to be very small (i.e., one letter only), the probability of a partial match would be considerably larger. The calculation of the exact probabilities for the different variations is complex and is not discussed further in this article, except for one example. Suppose we were to assume, that the name obtained by randomly choosing a PID contains nine unique letters of the alphabet, and that each letter of the alphabet is equally likely. Then the probability that only one letter of that name matches a second name randomly chosen from the set of all names would be

$$\sum_{i=1}^{m} (1/26 * p(i)),$$

where $p(i)$ represents the probability that the second name contains at least i letters, and m is the maximum length of the names in the name space.

In the case of patient validation, the situation is different from the one described above because the name pair is not chosen at random, but consists of the patient's name recorded at the clinical encounter and the name (in the registry) associated with the matching PID of the clinical encounter. Therefore if the two names are not identical but are similar, the following two possibilities exist:

- The two names correspond to the names of two different patients, and there is an identification error.
- The two names refer to the same patient; there is no identification error, only an occurrence of a variant form of the name.

The first possibility is highly unlikely if the names are very similar because as we discussed above, the probability of a match between two random names is extremely low. Therefore the second possibility, the occurrence of a name variation of the patient, is much more probable. This observation helps explain the judgment of experienced healthcare workers when validating a patient's identity. If most of the two names are very similar, the healthcare worker is likely to infer that they correspond to the same patient, even though they do not match exactly. This inference is typically made intuitively, because the likelihood of two similar names referring to two different patients is so small since the set of possible names to choose from is so large. The same reasoning is also appropriate with a Soundex type algorithm, which is a phonetic based hashing technique. The probability that two names chosen at random will have the same code is very unlikely (i.e., 0.01% because there are only 8918 different codes), and therefore, if two name pairs hash to the same code during patient validation, they are likely to consist of variations of the names of the same person.

4.2. A Discussion of the LCS Method

The technique which we developed has been adapted from an algorithm written by Baskin and Selfridge (2). We call it the LCS (Longest Common Substring) method because it is based on the notion of a likeness measure between two strings. The likeness measure is obtained using a procedure which iteratively finds and removes the longest common substring between two strings. The likeness measure is based on the total length of the common portions of the name pairs compared to the length of the actual names. This measure can be calculated in one of several ways, which are discussed further below. Because the LCS method removes the common substring from each name of the pair and repeats the matching process, it is effective in handling a variety of different types of errors, such as minor typos, small variations, and name permutations, and is not limited to single letter types of errors.

We will demonstrate the technique by providing two examples. In the first example, we use the name pair *Smith Mary* and *Mary Smith*, which corresponds to a permutation of part of a person's name.

1. The LCS is *Smith,* which has a length of 5. *Smith* is removed from both names, leaving a new pair of strings *Mary* and *Mary.*
2. The above process is repeated and an LCS *Mary* of length 4 is found. *Mary* is removed leaving two empty strings.
3. The iteration process ends because there are no more string portions left to be matched.

In this case, the match consists of a 100% match between both strings; this occurs when two names in the name pair are spelled exactly the same, but one of the pairs is permuted.

A second example is demonstrated where both a multiple letter error and a single letter typographic error exist. The name pair is *Nicholas Harrington* and *Nicky Harrinton:*

1. An LCS *Harrin* of length 6 is found and removed from both names, leaving a pair of strings *Nicholas gton* and *Nicky ton.*
2. An LCS *ton* is found (although there are two different LCSs of length 3, arbitrarily the last one is chosen first) and removed leaving the pair *Nicholas g* and *Nicky.*
3. An LCS *Nic* is found and removed leaving *holas g* and *y.*
4. There are no more common substrings, and the iteration procedure ends.

The common portion of the two strings has a total length of 11. The likeness measure can be computed by dividing the length of the common portion by a number consisting of one of the following:

1. The average length of the two strings.
2. The length of the smallest of the two strings.
3. The length of the largest of the two strings.
4. The length of the name in the registry database.

The second method is more permissive than the others, and therefore tends to maximize the likeness measure. It does not penalize cases where the departmental report contains an incomplete name or a nickname because it uses the length of the shortest name string as the denominator in the formula computing the similarity measure. In our trial runs at CPMC, we found the shortest of the two strings to be the best number to use because departmental reports frequently contains partial names. Using that number, the likeness measure of our example is 11/15, which is 73%.

In order to facilitate experimentation with different variations of the LCS technique, the corresponding matching procedure was written using several different parameters, each representing a different aspect of the matching process. Parameterization of the matching procedure also has the added advantage of being capable of ranging from more tolerant to less tolerant depending on requirements determined by the environment and users. The following are the parameters whose values may be changed as desired:

A likeness threshold. This allows the user to set the tolerance level for the matching algorithm. A pair of names are considered acceptable if the likeness measure is greater than or equal to the given threshold. Thus, the matching procedure would be more tolerant of errors when using a threshold of 40% than when using a threshold of 90%. We chose a threshold of 40% based on our study and the demands of the environment.

A minimum length threshold. This number determines the minimum length permitted for the longest common substring; common substrings less than this threshold are not considered in the matching process. We found that a minimum length of three characters was effective in our environment. A common substring of only one character would allow a partial match between a pair of names to be accepted where the letters of the two strings are exactly the same, but the order of the letters are completely different. For example, if the minimum length threshold were one, the strings *Ethan, Thane,* and *Hanet* would all be considered acceptable matches (i.e., the similarity measure between any pair would be 100%), and yet, these names do not appear similar in any way. Since we observed that pairs of names containing chunks of common portions appeared more similar than pairs of names containing very small common portions consisting only of one or two characters, we felt this factor should be reflected in the similarity measure.

A repetition threshold. This number limits the number of times the process of finding and removing the common substring may be repeated. This limits the number of independent errors that may occur in one name pair. We found 3 to be an effective limit for our environment.

4.3. Finding the Longest Common Substring

Pattern matching procedures can be very time-consuming. Much literature has been written about improving the efficiency of string matching algorithms (*1, 13, 22*). One problem that is frequently presented is the problem of finding the occurrence of one string *A* in another string *B*. Text editors and word processors generally have the capability of performing this search function. Two algorithms, the KMP algorithm (*16, 1*), and the Boyer–Moore algorithm (*5*), find an occurrence of one string in the other in an amount of time which is proportional to the length of the longest string. These algorithms are particularly effective when that string has repeating patterns. A straightforward procedure would perform this task less efficiently, in an amount of time which is equal to the product of the length of *A* and the length of *B*. However, finding the longest common substring of two strings is a somewhat different problem, which is more difficult and time-consuming than the one consisting of finding whether one string is contained in another.

Finding the longest common substring of two strings is also similar to the problem of molecular sequence comparison where the DNA sequence of a protein which is contained in a very large database is compared with a newly sequenced protein to find regions of similarity between the two proteins. The case of sequence comparison is more general than the one we are considering because the elements of the sequence do not have to be contiguous, whereas we are requiring that the elements of the common subsequence be adjacent in each string. Another difference is that sequence comparison permits substitutions within the similar subsequence, although substitutions are penalized more than insertions and deletions. The algorithm we describe below, which finds the common subsequence of two strings, is similar to some of the algorithms that are used to automatically compute the similarity (or distance) between two protein sequences. A survey and discussion of sequence comparison methods are described in detail in (*24*). Many of the algorithms in (*24*) are based on the computation of a similarity matrix $S(i,j)$ or a distance matrix $D(i,j)$, both of which are based on dynamic programming techniques. Our approach is similar because we utilize an array where each element contains the length of the common suffix of the name pair. However, the algorithm we use is simpler because our problem is more restrictive.

We use a dynamic programming algorithm to perform the pattern-matching task efficiently. Dynamic programming techniques solve a problem by solving similar subproblems and storing their results in a table of solutions. Larger problems are solved based on a formula which uses the results of the smaller problems. This technique is efficient timewise because the smaller problems never have to be recalculated because their results are saved. However, additional storage is required for the table. A discussion of dynamic programming techniques and some applications of the technique is in (*1*).

The technique we developed finds the longest common substring of two strings *A* and *B* by utilizing an array *L*, whose elements consist of the lengths

TABLE 2

Array of Common Suffixes

	2 (*C*)	2 (*e*)	3 (*r*)	4 (*n*)	5 (*e*)
1 (*C*)	1	0	0	0	0
2 (*h*)	0	0	0	0	0
3 (*e*)	0	1	0	0	1
4 (*r*)	0	0	2	0	0
5 (*n*)	0	0	0	3	0
6 (*y*)	0	0	0	0	0

of the common suffixes of substrings of A and B. The array L is defined as follows:

- Let n be the length of string A, and m be the length of string B.
- Let $L(i,j)$ represent the length of the common suffix of substrings of A and B such that i represents the substring A_i of A (i.e., this substring consists of characters 1 through i of A.), and j represents the substring of B_j of B. For example, if A is *Cherny* and B is *Cerne*, L (2,3) would represent the common ending of substrings *Ch* and *Cer* of A and B, respectively; its value would be 0 because those two substrings have no common ending. However, $L(5,4)$ would equal 3 because substrings *Chern* and *Cern* have a common suffix *ern* which has a length of 3.
- $L(i,j)$ is computed using the following rules:
  1. Initialize $L(i,j)$ to 0 for all i, where $0 \leq i \leq n$, and for all j, where $0 \leq j \leq m$.
  2. For all i, where $1 \leq i \leq n$, and for all j, where $1 \leq j \leq m$, if $A_i = B_j$ then $L(i,j) = L(i-1, j-1) + 1$; otherwise $L(i,j) = 0$.

Table 2 shows the complete array of common suffixes for the strings *Cherny* and *Cerne*. Notice that $L(2,1) = 0$ because *Ch* and *C* do not have a common suffix. However, $L(3,2) = L(2,1) + 1 = 1$ because the third letter of *Cherny* and the second letter of *Cerne* are both *e*. Likewise $L(4,3) = L(3,2) + 1 = 2$ and $L(5,4) = L(4,3) + 1 = 3$. However, $L(6,5) = 0$ because the sixth letter of *Cherny* is *y* which does not match *e*, the fifth letter of *Cerne*.

When computing $L(i,j)$ the maximum value of the elements of L is saved, along with the corresponding values of i and j. When L is completely filled in, the maximum value of the array is the length of the longest common subsequence, and the corresponding values of i and j represent the ending positions of the common subsequence in A and B, respectively.

The above method of filling in the array is based on the fact that if the last k characters of A_i and B_j match, and the $(i + 1)$st character of A matches the $(j + 1)$st character of B, then the last $k + 1$ characters of A_{i+1} and B_{j+1} also match. In contrast, if the $(i + 1)$st character of A and the $(j + 1)$st character of

B do not match, then there is no common suffix for that pair of substrings, and *L*(*i*,*j*) is 0.

The above pattern-matching algorithm is efficient timewise because it computes the longest common substring of *A* and *B* in an amount of time which is proportional to the product of *n* and *m* (i.e., the product of the lengths of the two strings), whereas a straightforward algorithm would take an amount of time which is proportional to the product of *n* squared and *m*.

4.4. *Other Aspects of the LCS Method: The Length of the Name Strings*

The LCS method should not be used for names that are short (i.e., less than six characters) unless the minimum length threshold is reduced to 1, because the results would be dependent on the position of the spelling error. For example, we will assume that the minimum length threshold is three, and two 5-letter strings are being matched. If there is only a single letter mismatch at the beginning (or end) of the word pair, the similarity measure would be 80%, because four out of five contiguous characters would match. But if a single letter mismatch occurred in the middle of the word, the similarity measure would be 0% match because there would be no common substrings of length 3 or more on either side of the mismatch. However, the lengths of patients' names (i.e., the entire name) are generally longer than nine characters, and therefore this is not a problem.

5. The Comparison Study

In order to analyze the effectiveness of the LCS method, we compared it to the Soundex method. Our analysis of the other methods determined that they would not be effective for validation purposes. We choose a test set of data, consisting of 14,793 unique pathology encounters and 10,000 unique patient entries in the Medical Records discharge summary. The PIDs of the patient records were used to extract the *accepted* names for the patients from the patient registry. Name pairs were formed using the patients' names recorded in the patient records and the names in the registry with the corresponding PIDs. An exact string match of the name pairs was performed, and the output of the name pairs that did not match was subjected to a manual analysis in order to determine the true rate of mismatches. The results of the manual analysis served as the *gold-standard* test in our study. The name pairs that matched exactly were not considered further in this study, and therefore we have ignored the possibility that a matching pair of names could actually correspond to two different patients.

The manual analysis of mismatched pairs was performed by one of the authors of this paper (R.S.), who was very familiar with the hospital environment and the ethnicity and history of the patient population, and therefore used background knowledge when comparing name pairs. The outcome of this analysis determined which mismatching name pairs appeared to represent name variants rather than true identification errors and therefore should be accepted rather than rejected.

The above study was not optimal for several reasons. It would be preferable for an independent auditor, also familiar with the hospital environment, to perform the manual analysis to remove possible bias from the analysis. However, the manual analysis was performed blind to the results of the other methods. The author who performed the analysis was involved in one aspect of the study, which analyzed different name matching techniques. The other author (C.F.) devised the LCS algorithm. Even if all potential sources of bias could be removed from the manual analysis, there would still be a problem with using this method as a *gold-standard*. At best, this method replicates the analysis procedure performed by a healthcare provider associated with the patient when the provider is presented with a name discrepancy. This is not a true test of the patient's identity. A true test would more definitively validate the patients identity by using additional factors other than the patient's name. At best this would involve contacting each patient, but this would be too impractical, time-consuming, and costly to carry out. Another study could compare other demographic attributes that are stored in both the departmental system and the central registry. This study is also time-consuming and costly, but should yield the prior probability of the odds that two patient names represent the same individual.

In addition to the manual analysis, the same two sets of data were used to compare the results of several different methods: the exact match method, the Soundex method, the LCS method using a likeness threshold of 0.40, and the LCS method using a likeness threshold of 0.60. In the two cases where the LCS method was used, both the minimum length threshold and the repetition threshold were 3 in each study. Table 3 summarizes the results for Pathology and Medical Records discharge summaries. These figures were based on the number of PIDs actually found in the registry. For 14,973 patient records from Pathology, 14,587 (97.4%) had matching PIDs in the registry, and for the 10,000 patient records which were discharge summarizes, 9,596 (96.0%) records were found to have matching PIDs.

Pathology and discharge summary data were chosen because they represent two different sources of data entry. The patient information in the pathology database is entered by in-house medical transcriptionists in different divisions of the Pathology Department, using requisitions that have patient information printed on them by using embossed identification cards. Our study shows that the pathology database contains fewer identification errors than the discharge summary database. In pathology, 2.6% of the patient records are rejected because no matching PIDs were found, and 1.9% of the records are rejected using the *gold-standard* (manual analysis) because of serious name mismatches. The discharge summarizes are dictated by in-house officers, are transcribed by typists at a remote transcription service, and are then uploaded to the Medical Records System. The discharge summaries understandably have a higher rejection rate (3.7%).

Obviously, the exact string method has a sensitivity of 100% for both types

TABLE 3

COMPARISON OF METHODS

Method	Pathology mismatches	%	Discharge summary mismatches	%
Manual analysis	277	1.9	352	3.7
Exact match	3375	23.1	3502	36.5
1. Sensitivity		100.0		100.0
2. Specificity		78.4		77.5
3. False acceptances		0.0		0.0
4. False mismatches		21.7		22.5
Soundex	963	6.6	1107	11.5
1. Sensitivity		99.3		99.1
2. Specificity		95.2		91.8
3. False acceptances		0.7		0.8
4. False mismatches		4.8		8.2
LCS 0.60	538	3.7	593	6.2
1. Sensitivity		99.6		99.1
2. Specificity		98.2		97.4
3. False acceptances		0.4		0.9
4. False mismatches		1.8		2.6
LCS 0.40	288	2.0	361	3.8
1. Sensitivity		95.7		95.7
2. Specificity		99.8		99.7
3. False acceptances		4.3		4.3
4. False mismatches		0.2		0.2

of data, but the specificity (78.4 and 77.5%) is the lowest of all the methods we have tested. This is not surprising, because every name pair that did not match exactly was rejected. The false mismatch rate was approximately 22%, which is unacceptably large. On the other hand, the false acceptance rate was 0.0% in both cases, because all mismatches were rejected, and therefore nothing was accepted which should not have been.

The Soundex method performed better than the exact match method; it has a sensitivity of approximately 99% and a specificity of 95.2 and 91.8%. Therefore it only has a false mismatch rate of 4.8 and 8.2%, which is a considerably improvement over the 22% of the exact match method. However the false mismatch rate is still rather high. A significant portion of the false mismatches occur because of cases where there are one or more letters in the name pair which are different. Having several characters different increases the likelihood that one of the different characters occupies one of the first four positions that form the Soundex code. This rate is also not surprising because the Soundex method was devised to correct for typical errors that occur when a name is given verbally, and not in written form. In order to reduce sound-alikes to the same code, this method ignores vowels, punctuation marks, repeating characters, and blanks, and groups consonents into sound-equivalency classes. This method therefore generally does not handle variations, such as typos, letter permutations, and name permutations. It also does not handle a single letter addition or deletion if the letter is a consonent.

Another problem with the Soundex method is that the sound-equivalency classes are formed based on English phonology. These classes are not necessarily valid for the letters associated with foreign languages. Because the origin of the patient population of CPMC comprises many different nationalities, the Soundex code is not the appropriate code to use for pairs of names which do not stem from English.

The false acceptance rate is very low for the Soundex method (0.72 and 0.85%). Three of these cases occurred because the last names were the same but the first names differed; one case occurred where there was a twin of a different sex, and another case occurred where the last names produced the same code but were significantly different in the manual analysis.

The specificity increases further with the LCS 60 method to 97.4% and 98.2%, and the false mismatch rate dropped to 1.8% and 2.6% which is a significant improvement over the Soundex rate of false mismatches. This is not surprising because the LCS method handles more types of name variations than the Soundex code does. There was a total of 4 false acceptances, which consisted of: two cases where the name pairs consisted of the same last name but different first names, one case where the name pair consisted of the same first name which was associated with a male name, and the last names, which were short, were significantly different; the last case consisted of a short last name that is contained in a fragment of the other name in the name pair.

The specificity of the LCS 40 method came the closest to the manual analysis. The specificity rises to 99.8 and 99.7%, and the false mismatch rate dropped to less than 0.3%.

A brief discussion of why the algorithm performs well for each common type of error associated with names will be presented below, with the understanding that the rate of effectiveness varies depending on the threshold value and the two other parameters.

According to our study of patient names at CPMC, the most common type of spelling error is due to the insertion of an additional name, initial, or title. When an initial or title is added, it is usually short compared to the rest of the name; therefore if the rest of the name matches, the likeness measure would generally be high and the mismatch would be considered acceptable. In addition, if the likeness measure is computed based on the length of the shortest name, the additional part of the name would have absolutely no affect on the measure; therefore if the rest of the name pair matches, the likeness measure would be 100%, and the two names would be considered the same.

A second common type of spelling error consists of one letter of the name being different. We will also include those cases where there is one letter deleted and one letter inserted because these cases all are affected by the LCS method in almost the same way. In cases of single letter errors, the likeness measure is $(n - i)/n$, where n is the length of the shortest name in the pair, and i ranges between *1* and *2*, depending on where in the name pair the single error occurs. Since most names are nine characters or more, an approximate lower bound for the likeness measure for single letter errors would be 7/9, or 78%.

The third most common type of spelling errors in names consists of punctuation errors. Since punctuation marks are removed, they should not cause any mismatches.

Another type of common error consists of several letters in the name pairs being different. The LCS technique is effective if the common portion of the name pair is relatively large compared to the length of the shortest name in the pair and to the specified threshold. This case generally covers the situation where one name in the pair contains a nickname, and the other name contains the full first name. Nicknames are often the same as a portion of the first name, and frequently consists of three or more characters of the first name (i.e., *Willy* and *William*), but this is not always the case. For example, *Margaret* and *Meg* have no common subsequences that are more than one letter but they are frequently used synonymously. An algorithm, which investigates variant forms of names and their frequencies, is reported in (*9*) along with a discussion on how tables containing common names and their variants could be an aid in matching name pairs. Our approach does not build such a table because we found the LCS method to be effective enough for our purposes without using a table of synonyms.

A small number of name mismatches occur because parts of the name are permuted. We have demonstrated above that this method handles these types of errors very effectively because if the name pairs are spelled the same, but the order of one of the names in the pair is permuted, the match is 100% using the LCS method.

Another type of error in name pairs occurs because one of the names in the pair is incomplete. Again, this is no problem if the length of the shortest name is used by the LCS method to compute the likeness measure; if a large portion of the shortest name matches the complete name, the measure is likely to be high, and the mismatch tolerated.

However, for the first time there is a significant increase in the false acceptance rates, which are 4.3% using this method. A rise in this rate is not surprising, since this method is so tolerant of spelling variations. One situation that caused a number of false acceptances occurred because certain names are used generically for certain groups of patients. For example, frequently for newborns, the name *female child* or *male child* is recorded in the registry in place of the first name. This situation changes the typical distribution properties of patient names in the registry database, and the likelihood of a match between a pair of names, randomly chosen, is somewhat greater if one of the names occurs more frequently than usual. This problem can be corrected by removing generic names before matching. This requires that a large enough sample of names be analyzed to find those cases where the distribution is much higher than usual.

A second situation that caused some false acceptances occurred because of a formatting convention used in our system: a patient's name is in the form of a last name separated by both a comma and blank, and followed by the first name. The separation between the last and first names therefore occupies two characters and inadvertantly counts more significantly when finding common subsequences. For example, if only the last character of each of the last names of the name pair matched (similarly, if the first character of the first names of the name pair matched), the LCS would equal 3. It would consist of the last character of the last name, the comma, and the blank between names. In order to correct this situation, the comma between names should be removed. At the time of the study, the comma punctuation separating names was not removed from the name string. We intend to correct this situation, and anticipate that there will be a smaller number of false acceptances. The LCS algorithm will be rerun using the same data to determine whether the rate of false acceptances is lowered.

There were also a small number of other cases where some portions of the name pairs were common but where the names referred to different patients. This occurred in two cases in pathology and in one case in the discharge summaries. This situation can not be corrected, but occurs so infrequently that it is acceptable.

5.1. Time Considerations

Pattern matching methods are more time-consuming than Soundex based methods in general. In the LCS method, the time to compute the likeness method is a function of the product of the length of the two names. The time to compute a Soundex code is constant because it only uses the first four consonents of the last name. However, when validating a patient's identification, the amount of time to compute the likeness measure is insignificant compared to the amount of time it takes to retrieve the data from the registry database. Therefore this method does not significantly increase the validation time. More importantly though, since the Soundex method falsely rejects more name pairs than the LCS method, it is more costly overall because clinical data which should be available is not available until the rejected records are returned to the departments as errors, corrected manually, and once again uploaded to the central database.

6. Conclusions

Our study has shown that an algorithm based on a likeness measure is the best method for tolerating spelling errors that occur during patient validation. A manual analysis of spelling errors that occurred during patient validation of pathology and discharge summary reports determined that only a small portion of the errors were identification errors, and that most errors consisted of name variant forms. The LCS algorithm, which we adapted from an algorithm developed by Baskin and Selfridge, accepted almost all of the mismatches that were considered acceptable by a manual analysis. The LCS algorithm rejected less than 0.03% more name pairs than the manual analysis.

Other algorithms, such as the Davidson code, the Koeln Phonetic code, the Megadats-2 code, and the exact match code were considered, but were rejected because previous studies showed that either they performed the same as or worse than the Soundex method. Other algorithms were also considered which were based on the most frequent types of spelling errors that occur when spelling general English words. However, in order to determine their effectiveness, we undertook a study to determine the common types of spelling errors that occurred when validating patient names. We found that different types of errors occurred, and therefore rejected algorithms based on common spelling errors in words because they were inherently limited when applied to names.

References

1. Aho, A. V., Hopcroft, J. E., and Ullman, J. D. "The Analysis of Computer Algorithms." Addison-Wesley, Reading, MA, 1974.

2. Alberga, C. N. String similarity and misspellings. *CACM* **10,** 302–313 (1967).

3. Bickel, M. A. Automatic correction to misspelled names: A fourth-generation language approach. *Commun. ACM* **30**(3), 224–228 (1987).

4. Bilofsky, H. S., and Burks, C. The GenBank genetic sequence databank. *Nucleic Acids Res.* **16**(1861) (1988).

5. Boyer, R. S., and Moore, J. S., A fast string search algorithm. *CACM* **20,** 762–772 (1977).

6. Damerau, F. J. A technique for computer detection and correction of spelling errors. *CACM* **7,** 171–176, 1964.

7. Davidson, L. Retrieval of misspelled names in an airline passenger record system. *CACM* **4,** 169–171 (1962).

8. Braeuer, I., *et al.* The use of phonetic code for patient identification. *In* "Proceedings, Medical Informatics Europe 82," Vol. 16, pp. 812–817. Springer-Verlag, Berlin, 1982.

9. Fair, M. E., Laloned, P., and Newcombe, H. B. Application of exact ODDS for partial agreements of names in record linkage. *Comput. Biomed. Res.* **24,** 58–71 (1991).

10. Goehring, R. Identification of patients in medical databases—Soundex codes versus match code. *Med. Inform.* **10**(1), 27–34, (1985).

11. Greenfield, R. H. An experiment to measure the performances of phonetic key compression techniques. *Meth. Inform. Med.* **16,** 230–233, (1977).

12. Joseph, D. M., and Wong, R. L. Correction of misspellings and typographical errors in a free-text medical English information storage and retrieval system. *Method. Inform. Med.* **18**(4), 228–234 (1978).

13. Morris, J. H., Jr., and Pratt, V. R. A linear pattern matching algorithm. Technical Report 40, Computer Center, University of California, Berkeley, CA, 1970.

14. Gersting, J. M., Jr., Conneally, P. M., and Rogers, K. Two search techniques within a human pedigree database. *In* "Proceedings, Sixth Annual Symposium in Computer Applications in Medical Care," pp. 842–846. IEEE, New York, 1982.

15. Knuth, D. E. "The Art of Computer Programming." Vol. 3. Addison-Wesley, Reading, MA, 1973.

16. Knuth, D. E., Morris, J. H., and Pratt, V. B. Fast pattern matching in strings. *SIAM J. Comput.* **6,** 323–350 (1977).

17. Kruskal, J. B., and Sankoff, D. An anthology of algorithms and concepts for sequence comparisons. *In* "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, Eds.). Addison-Wesley, London, 1983.

18. Mor, M., and Fraenkel, A. S. A hash code method for detecting and correcting spelling errors. *CACM* **25**(1), 935–938 (1982).

19. Peterson, J. L. Computer programs for detecting and correcting spelling errors. *CACM* **23**(12), 676–87 (1980).

20. POSTEL, H. J. Die koelner phonetic. *IBM Nachrichten* **198,** 925–931 (1969).
21. RUSSELL, R. C. U.S. Patent 1,435,663 (1922).
22. SMIT, G. DE V. A comparison of three string matching algorithms. *Software Pract. Exp.* **12,** 57–66 (1982).
23. SMITH, T. F., WATERMAN, M. S., AND FITCH, W. M. Comparative biosequence databank. *J. Mol. Biol.* **18**(38) (1981).
24. TYLER, E. C., MORTON, M. R., AND KRAUSE, P. R. A review of algorithms for molecular sequence comparison. *Comput. Biomed. Res.* **24,** 72–96 (1991).
25. M. S. WATERMAN. Sequence alignments. *In* "Mathematical Methods for DNA Sequences" (M. S. Waterman, Ed.) CRC Press, Boca Raton, FL 1989.

Chapter Two
Data

This chapter describes the selection of States for the study and the characteristics and contents of the administrative databases. The characteristics of the data are presented in detail, consistent with study goals to examine the feasibility of record linkage and consider the potential limitations of administrative data. Procedures for standardizing and cleaning the data prior to record linkage are also documented.

Selection of States for the Study

This study collected administrative data extracts from FSP and WIC programs in three States (Florida, Iowa, and Kentucky). These three States were selected based on the contents of their administrative databases, as reported during the Phase 1 survey conducted for this project. Two criteria were used to select States:

- **Common identifiers.** FSP and WIC client databases each had to have four common individual identifiers as *required* data fields in their client database: name, address, date of birth, and either Social Security Number (SSN) or phone number. A *required* data field is a field that is not supposed to be blank.
- **Record retention.** Participant records must be available for a three-year period, from January 2000 through December 2002. We preferred not to ask States to provide data from offline archives, to minimize burden.

The first criterion was used because individual identifiers such as name, date of birth, SSN, address, and phone must be present to establish a match across files. The presence of four identifiers gave us the flexibility to examine record linkage results under different matching scenarios, defined by the number of match variables. The second criterion was chosen arbitrarily so that we would have "enough" data to examine patterns of participation across the two programs over time.

Among the 26 States surveyed in Phase 1 of the study, only four States met the first criterion: FSP and WIC programs each have name, address, date of birth, and SSN in their client databases as required data fields. These States include the three participating in the study (Florida, Iowa, and Kentucky) plus Tennessee.[18] There were no surveyed States in which both FSP and WIC databases have name, address, date of birth, and phone as required data fields.[19] Table 2 shows all personal identifiers reported to be in the participant databases for the three selected States.[20]

Online record retention varied across the FSP and WIC programs in the three States selected for the study. FSP and WIC programs operate under federal regulations requiring record retention for a minimum of three years (7 CFR 275.4; 7 CFR 246.25), but offline archival can be used to satisfy those requirements. Kentucky FSP and WIC programs reported that client records are never taken

[18] Tennessee was chosen to participate in the study, but the Food Stamp Program was unable to provide data.

[19] Eight additional States met relaxed criteria such that name, address and date of birth are required for FSP and WIC; SSN is required for FSP and available but not required for WIC; and phone number is available but not necessarily required by either FSP or WIC.

[20] In FSP files, address appears on the records of household heads, but can be linked to each household member.

Table 2—Personal identifying information reported to be in FSP and WIC information systems

	First name	Last name	SSN	Date of birth	Address	Mailing address	Phone number	County	Gender	Race/ethnicity	Primary language	Date of first certification	Start & end dates of each cert. period	Monthly indicators of participation	WIC Only		
															Food Stamp case number	TANF case number	Medicaid case number
Florida																	
Food Stamp Program																	
Household head	✓	✓	✓	✓	✓	□	□	✓	✓	✓	✓	□	□	□			
Other family members	✓	✓	✓	✓	—	—	—	—	✓	✓	—	—	—	—			
WIC Program																	
Women	✓	✓	✓	✓	✓	□	□	✓	✓	✓	—	✓	✓	—	□	□	□
Infant/child	✓	✓	✓	✓	✓	□	□	✓	✓	✓	—	✓	✓	—	□	□	□
Iowa																	
Food Stamp Program																	
Household head	✓	✓	✓	✓	□	✓	□	□	□	□	—	✓	□	—			
Other family members	✓	✓	✓	✓	□	✓	□	□	□	□	—	✓	□	—			
WIC Program																	
Women	✓	✓	✓	✓	✓	—	□	✓	✓	✓	✓	—	✓	✓	—	—	□
Infant/child	✓	✓	✓	✓	✓	—	□	✓	✓	✓	✓	—	✓	✓	—	—	□
Kentucky																	
Food Stamp Program																	
Household head	✓	✓	✓	✓	✓	□	□	✓	✓	✓	✓	✓	✓	✓			
Other family members	✓	✓	✓	✓	—	—	—	—	✓	✓	✓	✓	✓	✓			
WIC Program																	
Women	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	✓	✓	—	—	—
Infant/child	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	—	✓	✓	✓	—	—	—

✓ Indicates data field is required to be filled; □ indicates data field is available but not required to be filled; — indicates data field is not available.

Source: *Survey of Food Assistance Information Systems*, Abt Associates, 2002.

offline. Florida FSP and WIC programs take records offline after cases have been inactive for 21 and 30 months, respectively. Iowa FSP and WIC programs take records offline after cases have been inactive for 24 and 66 months, respectively. All programs were asked to provide extracts containing persons active at any time during the three-year period from January 2000 to December 2002. Only Kentucky FSP was unable to provide data for the three-year period and instead provided data for one month (December 2002). It took several months for some State programs to fill the data request. Original requests were made in November 2002 and all data extracts were received by May 2003.

In addition to record retention policies, overwriting policies for individual data fields are relevant when collecting data retrospectively and linking data across systems. For example, when a person's name changes due to marriage, divorce, or adoption, some systems retain the old name in a separate data field, or history file, and some systems overwrite the old data. If a person is active in FSP and WIC at the same point in time, but enrolled at different points in time, then some identifying information may not match. The Phase 1 survey identified only one State (Kentucky) where both FSP and WIC data systems do not overwrite four identifying fields: name, date of birth, SSN, and address. The overwriting/retention rules reported in the Phase 1 survey for the selected States are shown in table 3.

The three selected States vary in caseload size. Table 4 shows FSP and WIC caseload information reported by USDA for the three States in the study. Florida is by far the largest with 5.1 percent of total U.S. food stamp participants and 4.3 percent of total U.S. WIC participants. Iowa has less than one percent of total FSP and WIC participants. Kentucky has 2.4 and 1.5 percent of total FSP and WIC participants, respectively.

Characteristics of Administrative Data Extracts

This section describes the characteristics of FSP and WIC data extracts, in terms of file size and format, records selected for matching, data elements, data quality, and participant dynamics within program over the three-year period. FSP and WIC programs in Florida and Iowa provided data for all persons participating in their program at any time during the period January 2000 through December 2002. Kentucky WIC also provided data for the three-year period, while Kentucky FSP provided data for one month (December 2002).

Table 3 – Overwriting and retention rules for personal identifying information in FSP and WIC programs in selected States

State	Program	Overwriting and retention of Name, Date of birth, SSN, Address, Telephone number[a]
Florida	FSP	Retain all except date of birth
	WIC	Overwrite all
Iowa	FSP	Overwrite all
	WIC	Overwrite all
Kentucky	FSP	Retain all except telephone number
	WIC	Retain all

[a] Indicates whether old information is retained in separate data field when change is made, or whether old information is overwritten and lost.

Source: Cole, Nancy. *Feasibility and Accuracy of Record Linkage To Estimate Multiple Program Participation: Volume I, Record Linkage Issues and Results of the Survey of Food Assistance Information Systems,* E-FAN-03-008-1.

Table 4 – Number of participants in FSP and WIC programs in three selected States

State	Food Stamp Program[a] Number of participants	Food Stamp Program[a] Percent of U.S. total	WIC[b] Number of participants	WIC[b] Percent of U.S. total
Florida	888,000	5.1	340,601	4.3
Iowa	124,000	0.7	62,798	0.8
Kentucky	418,000	2.4	121,098	1.5
U.S. Total	17,297,000	100.0	7,855,537	100.0

[a] Source: *Characteristics of Food Stamp Households: Fiscal Year 2001* (USDA, 2003).
[b] Source: *WIC Participant and Program Characteristics 2000* (Bartlett et al., 2002).

File size and format

The characteristics of the data files are shown in table 5, and the number of unique persons in those files is shown in table 6 (first column). Most data files were provided as flat file ASCII files, except Florida WIC data were in MS-Access format. Within program, file size varied across States according to caseloads. Within Florida and Iowa, FSP files were larger than WIC files due to larger caseloads and different record structure. The FSP files contained records for the entire caseload, even though only records for women of childbearing age, infants, and children would be matched to WIC data.

FSP data files contained one record per participant per month, while WIC data files contained one record per participant per certification.[21] This difference has two implications. First, identification of FSP participants in a given month was straightforward using the "year/month" indicator that was present on the file, while identification of WIC participants in a given month was based on certification date together with length of certification period.[22] The second implication is that FSP files identified participants who received benefits in a given month, whereas WIC files identified enrollees regardless of whether they picked up benefits for a particular month. The distinction between enrollees and participants is not considered important for this study because much of the analysis examined persons participating in FSP or WIC at any time during the three-year period.[23]

Data files from Florida, the largest of the three States, were nearly 10 times larger than data files from Iowa, the smallest State (measured by approximate file size). Florida FSP data consisted of over 30 million person-month records for 2.6 million participants during the three-year period, and the data occupied over 8 gigabytes of disk space. In contrast, data from the Iowa FSP program consisted of nearly 2.5 million person-month records for 337 thousand participants, and occupied less than one gigabyte of disk space.

[21] WIC certification records are also used in the USDA, Food and Nutrition Service biennial *Studies of WIC Participant Characteristics.*

[22] Most WIC applicants are certified for 6-month periods, except infants are certified up until their first birthday.

[23] As discussed later, when FSP files showed a one-month break between two spells of participation, FSP participation was imputed to provide a continuous spell. The elimination of these spurious breaks in participation makes the FSP data more comparable to WIC enrollment data.

Table 5—Administrative data files received from FSP and WIC programs

	File format	Period	Approx. file size	Total number of records[1]
Florida				
Food Stamp Program	ASCII	3 years	8 gigabytyes	32,802,926
WIC Program	MS-Access	3 years	515 MB	1,933,424
Iowa				
Food Stamp Program	ASCII	3 years	883 MB	2,451,181
WIC Program	ASCII	3 years	59 MB	362,494
Kentucky				
Food Stamp Program	ASCII	1 month	75 MB	474,685
WIC Program	ASCII	3 years	121 MB	684,999

[1] Number of records in FSP files is equal to number of person-months during 3-year period. Number of records in WIC files is equal to number of certifications during 3-year period.

Table 6—Analysis samples

	Total number of persons[1]	Women, Infants, and Children (W-I-C)[2]	
		All persons active 2000-2002	Active caseload in December 2002
Florida			
Food Stamp Program	2,621,488	1,194,425	388,817
WIC Program	981,464	981,464	403,477
Iowa			
Food Stamp Program	337,083	180,171	60,345
WIC Program	163,649	163,649	70,239
Kentucky			
Food Stamp Program	474,685	na	200,013
WIC Program	329,785	329,778	131,174

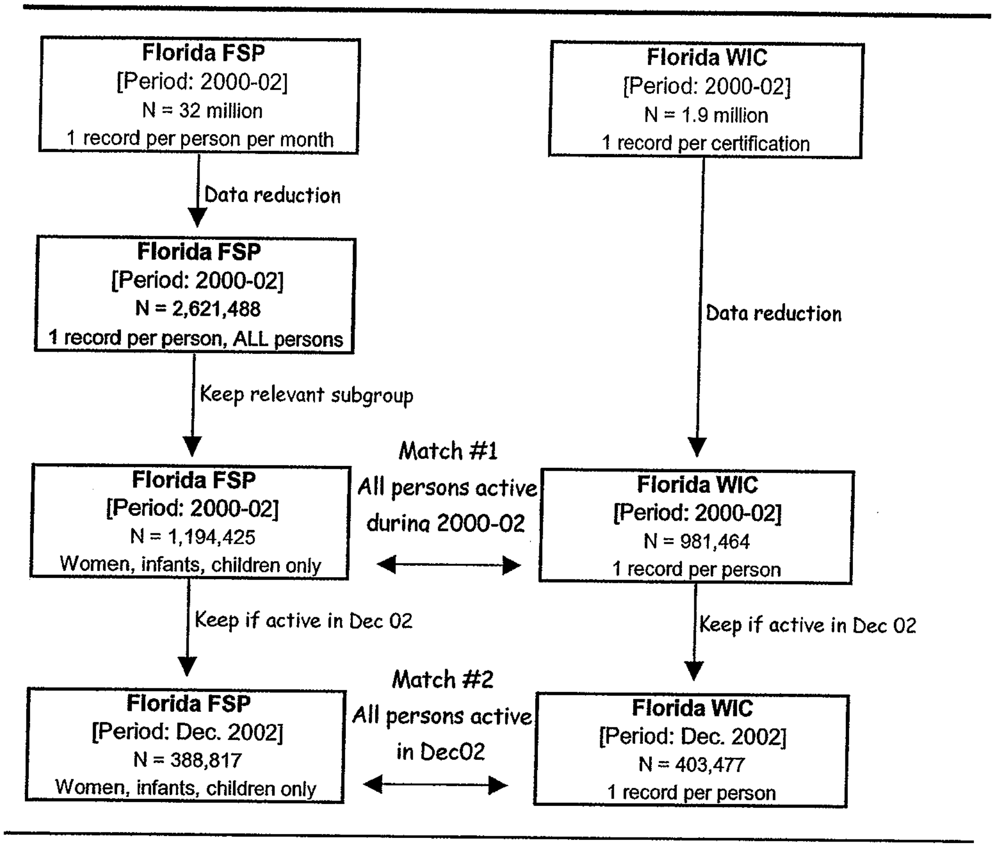
[1] FSP count of persons includes entire caseload and is not limited to women, infants, and children.
[2] W-I-C in FSP caseload are identified by age: women of childbearing age (15-45), infants, and children up to age 5.
na = not available.

Analysis samples

Table 6 shows the number of unique persons in each data file, and the number of unique persons in the analysis samples. Figure 4 provides a flowchart from the original data files to the analysis files, using Florida as the example. Two main steps are shown in the flowchart: data reduction (FSP and WIC) and selection of subgroups (FSP only). The original FSP data files were reduced from one record per person per month to one record per person with an array of monthly participation indicators. Similarly, WIC files were reduced from one record per certification to one record per person with an array of certification dates.[24]

[24] This is a simplified characterization of the data reduction; a more detailed discussion appears in chapter 3.

Figure 4 – Flowchart of data processing and selection of analysis samples



The analysis samples include all WIC participants and the subset of FSP participants identified as women of childbearing age (15-45 years old), infants, or children up to age 5 (hereafter referred to as W- I-C). All WIC participants are used for matching even though only persons with income not exceeding 130 percent of the federal poverty level are potentially eligible for FSP, subject to resource limits (see table 1). This subset of WIC participants cannot be identified with precision, however, because definitions of household and household income vary between FSP and WIC. In addition, the availability of income data in WIC administrative databases varies among States.[25] For these reasons, a subset of records was not selected from WIC databases prior to matching.

Two analysis samples were used: 1) W-I-C who were active at any time during 2000-2002, and 2) W-I-C who were active in December 2002. These samples are denoted "Match #1" and "Match #2" in figure 4. December 2002 was chosen because data from Kentucky FSP were received for that month only.

Table 7 provides a count of women, infants, and children included in the matching routines from the December 2002 caseloads of each State. W-I-C represent 38 to 45 percent of *total* FSP caseloads and

[25] The biennial census of WIC participants reported in *WIC Participant and Program Characteristics 2000* found that income was reported on the records of only 87 percent of WIC participants in April 2000. Administrative records from seven States (including Kentucky) had income missing for over 30 percent of WIC participants.

Table 7—Number and percent of women, infants, and children (W-I-C) in FSP and WIC caseloads, December 2002

	Food Stamps		WIC	
	Number	Percent of total caseload	Number	Percent of total caseload
Florida				
Total W-I-C	388,817	38.19	403,477	100.00
Women				
Age 15-18	33,379	3.28	10,214	2.53
Age 19-34	122,500	12.03	78,815	19.53
Age 35-45	71,677	7.04	9,577	2.37
Total	227,556	22.35	98,606	24.44
Infants	29,953	2.94	112,352	27.85
Children				
Age 1	34,030	3.34	63,476	15.73
Age 2	33,712	3.31	50,384	12.49
Age 3	32,066	3.15	42,822	10.61
Age 4	31,500	3.09	35,837	8.88
Total	131,308	12.90	192,519	47.71
Iowa				
Total W-I-C	60,345	45.03	70,239	100.00
Women				
Age 15-18	4,085	3.05	1,745	2.48
Age 19-34	22,348	16.68	14,171	20.18
Age 35-45	10,782	8.05	1,069	1.52
Total	37,215	27.77	16,985	24.18
Infants	4,655	3.47	17,227	24.53
Children				
Age 1	4,781	3.57	11,650	16.59
Age 2	4,764	3.56	9,242	13.16
Age 3	4,660	3.48	8,232	11.72
Age 4	4,270	3.19	6,903	9.83
Total	18,475	13.79	36,027	51.29
Kentucky				
Total W-I-C	200,013	42.14	131,174	100.00
Women				
Age 15-18	16,205	3.41	3,895	2.97
Age 19-34	75,515	15.91	27,246	20.77
Age 35-45	37,539	7.91	1,597	1.22
Total	129,259	27.23	32,738	24.96
Infants	14,016	2.95	33,965	25.89
Children				
Age 1	14,384	3.03	21,261	16.21
Age 2	14,526	3.06	16,463	12.55
Age 3	14,149	2.98	14,122	10.77
Age 4	13,679	2.88	12,625	9.62
Total	56,738	11.95	64,471	49.15

100 percent of WIC caseloads. FSP enrolls more women of childbearing age than WIC (because FSP enrolls women who are not pregnant or postpartum). WIC enrolls more infants and children than FSP. The ratio of WIC infants to FSP infants varied across States: 3.7 in Florida and Iowa, and 2.4 in Kentucky. The ratio of WIC children to FSP children also varied across States: 1.5 in Florida, 2.0 in Iowa, and 1.1 in Kentucky. These differences are consistent with different Medicaid eligibility provisions, which affect WIC enrollment through WIC adjunct income eligibility.[26]

The number of records entering the matching routine exceeds the number expected to match, for four reasons. First, FSP women of childbearing age include women not eligible for WIC because they are not pregnant or postpartum. Pregnant women cannot be identified in the FSP data and postpartum women may be identified with error if the mother-infant pair does not reside together or if there is a lag in enrolling the infant in the FSP. Second, all FSP W-I-C are income-eligible for WIC, but they may not necessarily meet WIC nutritional risk criteria. Third, some WIC participants across all categories of W-I-C are not eligible for FSP because their income exceeds 130% of poverty (WIC eligibility threshold is 185% of poverty).[27] Fourth, some persons eligible for both programs will not be matched because they have decided to participate in only one program, even though they may be eligible for both.

The subsets of FSP records selected for matching were taken from caseloads that are described in table 8. This table shows the distribution of persons and households participating in FSP by household type, and the percent of persons from each household type that are potentially eligible for WIC. In all three States, approximately 40 percent of all FSP participants are in single-adult households with children. Infants and children under age five represent 15 to 17 percent of FSP participants, and women of childbearing age represent 22 to 27 percent. Children under 5 years of age are present in more than half of single-adult-households-with-children[28], and nearly 90 percent of those households contain women of childbearing age (not shown in table).

Data elements

The data elements provided in each data file are shown in table 9. Five main types of data elements were requested for each program participant: personal identifiers, contact information, program participation dates, household income, and indicators of participation in certain other public programs.

The three States participating in this study were purposefully selected based on the data fields present in their program databases. For the most part, table 9 coincides with table 2 (information reported to the Phase 1 survey). The data extracts from all programs contain first and last name, date of birth, SSN, gender, and race. All FSP programs provided information on participants' relationship to household head. All WIC programs provided certification category and guardian names for infants and children. Contact information includes street address, city, State, ZIP code, and phone number.

[26] Florida and Iowa Medicaid eligibility for infants is 200% of poverty compared to 185% of poverty in Kentucky. Florida and Iowa have Medicaid continuous eligibility provisions, which may explain higher ratios of WIC to FSP children in those States. (See footnote to table 1.)

[27] In addition to income and asset limits, there are non-financial FSP eligibility restrictions – particularly those related to citizenship, residency, and immigration status – that might impact a WIC participant's eligibility for food stamps.

[28] Children under age five are 15.84 percent of the total Florida caseload and 8.69 percent of children under age 5 are in single-adult households: 8.69 / 15.84 = 55%.

Table 8—Distribution of persons and households participating in FSP by household type, December 2002

	Persons		Households		Percent of persons who are		Total percent potentially eligible for WIC
	Number	Percent	Number	Percent	Children under age 5	Women of child-bearing age	
Florida, Total	1,017,979	100.00	480,847	100.00	15.84	22.35	38.20
With children							
Single adult	427,853	42.03	135,425	28.16	8.69	13.16	21.85
Married couple	125,175	12.30	28,787	5.99	2.17	2.75	4.93
Multiple adults	90,695	8.91	20,112	4.18	1.73	2.57	4.30
Children only	80,700	7.93	39,156	8.14	3.24	0.51	3.75
Without children							
Single adult, elderly	109,431	10.75	109,431	22.76	0.00	0.00	0.00
Single adult, not elderly	113,826	11.18	113,826	23.67	0.00	2.71	2.71
Multiple adults, elderly	50,085	4.92	24,366	5.07	0.00	0.11	0.11
Multiple adults, not elderly	20,214	1.99	9,744	2.03	0.00	0.54	0.54
Iowa, Total	134,005	100.00	59,699	100.00	17.27	27.77	45.04
With children							
Single adult	59,548	44.44	20,421	34.21	10.28	15.17	25.46
Married couple	29,738	22.19	6,824	11.43	4.41	5.32	9.73
Multiple adults	10,036	7.49	2,331	3.90	1.55	2.17	3.71
Children only	3,443	2.57	2,051	3.44	1.03	0.39	1.42
Without children							
Single adult, elderly	7,295	5.44	7,295	12.22	0.00	0.00	0.00
Single adult, not elderly	17,775	13.26	17,775	29.77	0.00	3.83	3.83
Multiple adults, elderly	2,335	1.74	1,141	1.91	0.00	0.05	0.05
Multiple adults, not elderly	3,835	2.86	1,861	3.12	0.00	0.84	0.84
Kentucky, Total	474,685	100.00	198,176	100.00	14.91	27.23	42.14
With children							
Single adult	190,618	40.16	65,258	32.93	8.52	13.54	22.05
Married couple	113,939	24.00	27,530	13.89	3.97	6.06	10.03
Multiple adults	56,110	11.82	13,253	6.69	1.97	3.39	5.36
Children only	6,052	1.27	3,327	1.68	0.44	0.28	0.73
Without children							
Single adult, elderly	24,228	5.10	24,228	12.23	0.00	0.00	0.00
Single adult, not elderly	47,089	9.92	47,089	23.76	0.00	2.52	2.52
Multiple adults, elderly	14,163	2.98	6,800	3.43	0.00	0.11	0.11
Multiple adults, not elderly	22,486	4.74	10,691	5.39	0.00	1.34	1.34
U.S. Average, FY2001[1]	17,300,000	100.00	7,450,000	100.00	16.64	27.32	43.96
With children							
Single adult	8,494,000	41.65	2,690,000	31.74	–	–	–
Married couple	2,658,000	13.03	572,000	6.75	–	–	–
Multiple adults	1,426,000	6.99	325,000	3.84	–	–	–
Children only	831,000	4.08	405,000	4.78	–	–	–
Without children							
Single adult, elderly	1,220,000	5.98	1,220,000	14.40	–	–	–
Single adult, not elderly	2,017,000	9.89	2,017,000	23.80	–	–	–
Multiple adults, elderly	712,000	3.49	300,000	3.54	–	–	–
Multiple adults, not elderly	3,034,000	14.88	945,000	11.15	–	–	–

[1] Source: USDA, Food and Nutrition Service. *Characteristics of Food Stamp Households: Fiscal Year 2001*, Alexandria, VA: 2003. From this source, the sum of individual categories does not match the table total because participants and households were counted in multiple categories.

– Data not available.

Table 9—Data elements in the FSP and WIC administrative data extracts

	Food Stamp Programs			WIC Programs		
	Florida	Iowa	Kentucky	Florida	Iowa	Kentucky
Personal identifiers						
Participant ID	✔	✔	(1)	✔	✔	✔
Case number	✔	✔	✔			
First name	✔	✔	✔	✔	✔	✔
Last name	✔	✔	✔	✔	✔	✔
Middle initial			✔		✔	✔
Date of birth	✔	✔	✔	✔	✔	✔
Social Security Number (SSN)	✔	✔	✔	✔	(2)	✔
Sex	✔	✔	✔	✔	✔	✔
Race code	✔	✔	✔	✔	✔	✔
Language			✔		✔	
Relationship to household head	✔	✔	✔			
Certification category				✔	✔	✔
Contact information						
Address						
Street	✔	✔	✔	✔	✔	✔
City	✔	✔	✔	✔	✔	✔
State	✔	✔	✔	✔	✔	✔
Zip code	✔	✔	✔	✔	✔	✔
Phone	✔	(3)	✔	✔	✔	✔
County (office) code	✔	✔	✔	✔	✔	✔
Dates of program participation						
Month/Year indicator	✔	✔	(1)			
Certification date				✔	✔	✔
Certification end date				✔		
WIC family information						
Family ID				✔		
Guardian first name				✔	✔	✔
Guardian last name				✔	✔	✔
Guardian middle initial						✔
Income						
Family (household) size	✔	✔	✔	✔	✔	✔
Income		✔	✔	✔	✔	✔
Participation in other programs						
Food Stamps				✔	✔	✔
Medicaid		✔		✔	✔	✔
TANF				✔	✔	✔
Cash assistance	✔	✔	✔			
FSP/TANF/Medicaid ID	✔			✔		

✔ Indicates data element is present.

(1) These fields were not needed because Kentucky FSP provided only one month of data.

(2) SSN is not a separate data field. Participant ID contains own SSN (women) or mother's SSN (infants/children), if available. Else the participant ID contains the day and year that the record was entered in the system.

(3) Phone numbers do not include area code.

Information about participation in certain other programs is present in both FSP and WIC client databases. The FSP programs in all three States are integrated with TANF, providing a reliable indicator of cash assistance on each person-month record. Iowa FSP records also include an indicator of Medicaid participation. WIC programs are not integrated with other public assistance programs, but their databases contain indicators of adjunct income eligibility (participation in TANF, FSP, or

Medicaid) because applicants may be certified without income documentation if they document participation in these means-tested programs. As noted earlier, indicators of adjunct income eligibility may underestimate actual rates of participation in each program because reporting of one program is sufficient to establish WIC eligibility, even if persons participate in more than one adjunct program. In addition, these indicators capture participation in adjunct programs at the time of WIC certification but do not reflect enrollment in adjunct programs after WIC certification.

FSP and WIC data systems each assign unique participant IDs to individuals. These IDs provide a link between records over time within each system.[29] An additional identifier in the FSP is the participant's case number, identifying the household unit that applied to the program. FSP participants may, however, be associated with multiple case numbers over time if the composition of the household changes. Usually there is a change in case number when there is a change in household head. Some of the analysis presented in chapter 4 excludes "complicated" households containing individuals who changed case number during the study period.[30]

The FSP case number provides a link among FSP household members. In contrast to the FSP, which enrolls households, WIC enrolls individuals. Even so, some WIC programs assign a family ID in addition to a participant ID for use in appointment scheduling and other administrative functions (see Cole, 2003). Among the three WIC programs in this study, only Florida assigns a family ID. As shown in table 9, WIC records for infants and children contain a guardian name that could be used to link family members, but this link was not needed for the analysis presented in this report.[31]

Personal identifiers include first and last name, date of birth, SSN, gender, and race. These data items were the primary items used to link records from FSP and WIC. All identifiers except last name are expected to be stable over time, except for data entry errors or use of abbreviations or nicknames for the first name. Last names may change over time due to marriage, adoption, or divorce.

Contact information consists of components of the address field, telephone number, and county. Contact information is not necessarily stable over time, but it is helpful in linking contemporaneous records from two different data files.

Dates of program participation identify the active caseload at a point in time, and were used to examine the dynamics of program participation and multiple program participation. As discussed above, FSP files contain one record for each active participant each month. Each record has a "year/month" indicator. WIC data contain one record per certification period, and each record contains a certification date. The Florida WIC program also provided the "certification end date" in their data file even though it was not a requested data field; certification end dates were imputed for Iowa and Kentucky based on certification date and program rules. The certification date start and end dates can be used to identify the active caseload at a point in time.

The contents of data extracts are consistent across programs, with the following exceptions:

[29] Some WIC programs have participant IDs that are unique within the local agency, but not unique within the State. In these States, participant IDs change when participants move and transfer to a new agency and the link between longitudinal records is broken (Cole, 2003). Florida, Iowa, and Kentucky assigned unique IDs within the State.

[30] The percent of W-I-C in "complicated households" was 6 percent in Iowa and 8 percent in Florida.

[31] Mother-child pairs could be linked by mother's name and guardian names; siblings could be linked by guardian names.

- Language: Available from only two programs and was not used in matching routines.
- SSN: Iowa WIC does not have SSN in a separate data field. If SSN is provided by an applicant, it is used as part of the participant ID; however, a mother's SSN may be used as part of her child's participant ID. SSNs were extracted from participant IDs for women, but not for infants and children.
- Telephone: Iowa FSP did not include area code.[32]
- Family ID: Kentucky and Iowa WIC do not maintain a family ID.[33]
- Income: Florida FSP did not provide income.
- WIC dates: Florida was the only program to provide certification end dates; these dates were imputed for the other two programs.
- Adjunct ID: Florida WIC was the only WIC program to maintain the FSP/TANF/Medicaid ID number in addition to indicators of participation in those programs. This data field was not used in the record linkage routines, but was used in examining the results of record linkage.

Quality of participant data

Data files were evaluated for prevalence of missing data and standardization of address fields. Missing data are indicated by blank fields or fields filled with zeros or nines. Standardization implies that the same data appears identically within the data file. For example, city names might be standardized at data entry by choosing cities from a list rather than keying in city names, thus eliminating spelling variations.

Examination of the December 2002 caseloads showed that FSP and WIC files in all three States had no missing data for participant names and virtually no missing data (less than .01 percent) for date of birth and gender. Race was almost never missing on WIC records and was missing on less than 2 percent of FSP records. Each of the address components (street address, city, and ZIP code) was missing on less than 2 percent of FSP and WIC records.[34]

The data fields subject to quality problems are shown for FSP and WIC in tables 10 and 11, respectively. SSN and telephone number were subject to missing data; city was not standardized; and

[32] Iowa FSP and WIC data were matched using telephone number without area code. Iowa is divided into five area codes, however, so it was possible that telephone numbers in two different area codes would provide a false match. This was not considered a significant problem because telephone number was only one of several identifiers used for matching.

[33] Florida reported that family IDs are reliable for "some currently participating family members" (Cole, 2003). Family IDs might not be reliable for linking family members whose participation was not contemporaneous.

[34] It is difficult to accurately assess the amount of missing data for street addresses without geocoding the data, which was not done. Casual observation revealed that this data field was occasionally used for comments – for example, to indicate a contact person outside the family.

Table 10—Percent of FSP records with missing or non-standardized data, December 2002

	Number of records	Percent with missing data for		Percent with nonstandardized data for	
		SSN	Telephone	City	ZIP code
Florida					
Total FSP	388,817	1.3	6.2	10.3	0.3
Women	227,556	0.4	6.2	10.5	0.3
Infants	29,953	12.8	6.2	10.0	0.4
Children	131,308	0.3	6.0	10.1	0.3
Iowa					
Total FSP	60,345	1.2	12.9	2.7	0.1
Women	37,215	0.1	13.5	2.7	0.1
Infants	4,655	14.7	12.8	2.8	0.0
Children	18,475	0.2	11.7	2.8	0.1
Kentucky					
Total FSP	200,013	0.0	6.6	28.6	0.1
Women	129,259	0.0	6.4	29.1	0.1
Infants	14,016	0.0	7.3	27.8	0.1
Children	56,738	0.0	6.7	27.6	0.1

Table 11—Percent of WIC certification records with missing or non-standardized data, December 2002

	Number of records	Percent with missing data for		Percent with nonstandardized data for		Quality of income data	
		SSN	Telephone	City	ZIP code	Missing income	Income equal zero
Florida							
Total WIC	403,477	26.6	3.7	25.8	0.5	0.9	3.7
Women	98,606	13.7	3.6	29.2	0.6	0.8	4.3
Infants	112,352	71.4	3.7	22.1	0.6	1.2	4.9
Children	192,519	7.1	3.8	26.1	0.5	0.8	2.7
Iowa							
Total WIC	70,239	76.6	2.8	6.3	0.3	0.0	8.7
Women	16,985	3.0	2.7	6.4	0.3	0.0	12.5
Infants	17,227	100.0	2.7	6.2	0.3	0.0	12.3
Children	36,027	100.0	2.8	6.4	0.3	0.0	5.2
Kentucky							
Total WIC	131,174	11.0	4.1	31.8	0.8	43.1	14.7
Women	32,738	1.2	3.7	31.2	0.7	30.0	14.7
Infants	33,965	33.5	5.5	31.0	1.4	53.1	16.8
Children	64,471	4.2	3.6	32.5	0.6	44.4	13.6

Note: Iowa WIC does not have a separate data field for SSN. See text discussion.

WIC income data showed high percents of missing data or zero values. It is important to note, however, that records with missing data were included in the record linkage procedures. As explained in chapter 3, probabilistic record linkage uses all available information. Missing data in one or more data fields does not necessarily preclude a match.

SSN was never missing on Kentucky FSP records. For Florida and Iowa, SSN was missing on only 1 percent of FSP records overall, but on over 10 percent of infant records. FSP requires an SSN for certification, so it is likely that missing data reflects the delay in SSN issuance for newborns. WIC does not require SSNs for certification, which is reflected in higher rates of missing data compared with FSP. SSNs are missing across all WIC participant categories, although the highest rates are for infants.[35] The table shows that SSN is missing for all Iowa infants and children. As discussed above, Iowa WIC does not have a separate data field for SSN; SSN was extracted from the participant ID for women, but SSNs embedded in the participant ID of infants and children were not extracted because they were likely to be the mother's SSN.

Telephone numbers are potentially valuable for record linkage because they are long numeric fields that are unique to households. Missing telephone numbers are more common in FSP than WIC; 6 and 7 percent of FSP records in Florida and Kentucky, and 13 percent in Iowa are missing telephone number. Only 2 to 4 percent of WIC records are missing telephone number across the three States.

Tables 10 and 11 show the percent of city names and ZIP codes that are not standardized in the sense that they do not match exactly to a master list of place names (cities, towns, county divisions) and ZIP codes in the State.[36] Spelling variations in city names compromise the usefulness of these data for record linkage. For example, there are 960 place names in Florida but over 7,000 unique city names are in the WIC data files (e.g., over 40 spelling variations were recorded for Fort Lauderdale). Kentucky has the highest prevalence of non-standardized city names at 29 and 32 percent of FSP and WIC records, respectively. The percent of non-standardized city names in FSP and WIC was 10 and 26 percent in Florida, and only 3 and 6 percent in Iowa.

The lack of standardized city names was not consistent with responses to the Phase 1 survey: Florida and Kentucky FSP indicated that city and ZIP codes were standardized.[37] Because city names were not standardized, it was not clear that this data field should be used for matching. In addition, ZIP codes could be more useful for matching because they have more geographic precision than cities, which may contain multiple ZIP code areas. Nonetheless, it was thought to be beneficial to include both city and ZIP code in matching routines because potential errors in data are generated differently. ZIP code errors are most likely to result from transposition of numbers, resulting in a ZIP code that references the wrong city. On the other hand, city errors are unlikely to occur in the sense that the wrong city is referenced, but city names are subject to spelling errors and spelling variations.[38]

[35] Missing SSNs on WIC records may reflect enrollment of persons without access to SSNs, such as illegal aliens.

[36] The master list of city names and ZIP codes was current as of September 2003, from the ZIPList5 database available from CD Light, LLC (zipinfo.com).

[37] FSP and WIC programs in all three States indicated that ZIP codes are not validated (source: Phase 1 survey).

[38] As discussed in chapter 3, city and ZIP codes were matched using different criteria with an exact match required for ZIP code but a string comparison (allowing for spelling variations) used for city.

Table 11 shows the quality of income data for WIC programs. The WIC income eligibility cutoff is higher than FSP eligibility limits, so income could potentially be used to select records from WIC files prior to record linkage. As shown in table 11, however, over 40 percent of Kentucky WIC records are missing income data, and a large percent of records in Iowa and Kentucky have zero income (9 and 15 percent, respectively).[39] As noted above, WIC records were not selected based on income prior to record linkage.

Availability of historical data for personal identifiers

In theory, the data files obtained for this study could provide estimates of the rate of change in household information over time (e.g., name changes due to marriage, divorce, adoption) and the rate of mobility (e.g., address and telephone changes) for FSP and WIC populations. In practice, however, the rates of change in individual identifiers depends on whether information systems overwrite or retain data, and the way in which data extracts are created.

Table 3 reported the overwriting and retention rules reported in the Phase 1 survey for name, date of birth (DOB), SSN, address, and telephone number. Iowa FSP, Iowa WIC, and Florida WIC reportedly overwrite all identifiers when information changes (thus losing old information, except in off-system archives). Florida FSP overwrites only DOB; Kentucky FSP overwrites only telephone number; and Kentucky WIC reported no overwriting.

Tables 12 and 13 show the availability and prevalence of historical changes in identifying information observed in the data. There are some inconsistencies between reported overwriting policies and actual data, which may be due to the methods used to create data extracts. Florida FSP data show no change in personal identifiers (even though there is reportedly no overwriting) and high rates of change in contact information. Iowa FSP data show near zero rates of change in personal identifiers (consistent with overwriting of all information) and high rates of change in contact information (not consistent with overwriting).[40]

WIC data from Florida and Iowa are consistent with the overwriting policies discussed above –these programs reportedly overwrite all data and the data files show no change in personal identifiers over time. Florida WIC shows small rates of change in contact information indicating a possible change in policy over time or an effort to standardize data.

Observed rates of change in personal identifiers are likely to be equal or close to true rates of change for Iowa FSP and Kentucky WIC. The rates of change are measured over a one-year period.[41]

[39] Missing income data has been reported in the *WIC Participant and Program Characteristics* reports as associated with adjunct income eligibility (Bartlett et al., 2000 and 2002). However, adjunct income eligibility does not imply income below a single national cutoff because Medicaid eligibility thresholds vary by State and may exceed the WIC threshold of 185% of poverty. As of October 1, 2000, Medicaid eligibility was 200 percent of the poverty level for pregnant women and infants in Iowa, and for infants in Florida.

[40] Florida FSP provided person-level records and case-level records (contact information) in separate files and most likely lost the historic person-level information in the way that the data were extracted. Iowa FSP data extracts were created from month-end archives, thereby preserving the historical data.

[41] This analysis was based on the most recent 12 months of participation for FSP participants with at least 6 months of participation, and the two most recent certification records for WIC participants. Restricting the sample to a one-year period eliminates the potential downward bias if long-term participants are more stable than short-term participants.

Table 12—Availability and prevalence of historical changes in personal identifying information for FSP participants[1]

	Florida				Iowa			
	Total	Women	Infants	Children	Total	Women	Infants	Children
Number participants with > 6 months of participation	595,245	351,592	63,332	180,321	93,229	56,558	9,353	27,318
Percent with change in personal identifiers[2]								
First name	0.0	0.0	0.0	0.0	0.3	0.2	0.6	0.3
Last name	0.0	0.0	0.0	0.0	2.0	2.9	1.0	0.5
Date of birth	0.0	0.0	0.0	0.0	0.2	0.1	0.6	0.2
SSN	0.0	0.0	0.0	0.0	0.5	0.2	3.1	0.3
Gender	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Race	0.0	0.0	0.0	0.0	1.0	0.8	2.6	0.8
Percent with change in contact information								
Telephone	42.7	41.2	44.4	45.2	20.4	19.3	23.8	21.6
Street name[3]	49.1	47.8	51.6	50.8	36.0	34.3	41.1	37.8
City	13.4	12.6	14.5	14.3	13.3	12.5	15.6	14.1
County	4.6	4.4	5.0	4.9	11.6	11.3	12.7	11.9
ZIP code	23.4	22.2	25.5	25.1	21.6	20.4	25.2	23.1

[1] Prevalence of change in identifying information is evaluated over the last 12 months of participation for participants with at least 6 months of participation.
[2] Change from missing to nonmissing is not counted, and vice-versa.
[3] Change is evaluated after parsing the street name from the address field with the Census standardization software.

Table 13—Availability and prevalence of historical changes in personal identifying information for WIC participants[1]

	Florida				Iowa				Kentucky			
	Total	Women	Infants	Children	Total	Women	Infants	Children	Total	Women	Infants	Children
Number participants with multiple certfications	523,049	184,636	77,135	261,278	95,968	31,200	11,430	53,338	176,375	60,871	20,971	94,533
Percent with change in personal identifiers[2]												
First name	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.9	3.3	0.5
Last name	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	3.1	6.1	4.1	1.0
Date of birth	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.8	0.2
SSN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gender	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.9	0.2
Race	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.2	0.9	0.2
Percent with change in contact information												
Telephone	1.7	1.4	2.5	1.7	0.0	0.0	0.0	0.0	28.8	30.6	36.0	26.1
Street name[3]	1.6	1.4	2.4	1.5	0.0	0.0	0.0	0.0	26.3	28.8	36.9	22.4
City	7.5	5.7	9.7	8.2	0.0	0.0	0.0	0.0	10.2	11.3	14.0	8.6
County	0.2	0.2	0.3	0.2	0.0	0.0	0.0	0.0	4.0	4.8	5.6	3.1
ZIP code	1.0	0.8	1.5	1.0	0.0	0.0	0.0	0.0	13.2	14.6	18.7	11.1

[1] Prevalence of change in identifying information is evaluated over the last two certification records for participants with multiple certifications.
[2] Change from missing to nonmissing is not counted, and vice-versa.
[3] Change is evaluated after parsing the street name from the address field with the Census standardization software.

Evidence from these programs suggests that personal identifiers are unlikely to change over time. Rates of changes in first name, DOB, SSN, gender, and race are less than one percent (except for infants) and most likely reflect corrections to erroneous entries and not true changes. Changes in last name are rare (one percent or less) for children in both Iowa FSP and Kentucky WIC. Evidence from Kentucky WIC, however, suggests that approximately five percent of women and infants change last name within a one-year period, possibly reflecting changes in marital status after childbirth.

Observed rates of change in contact information are likely to be equal or close to true rates of change for Florida FSP, Iowa FSP, and Kentucky WIC. Evidence from these programs indicate that 20 to 43 percent of program participants change telephone number within a year, 26 to 49 percent change street address, and 10 to 13 percent move to a new city.[42] Unfortunately, none of the three States provide direct within-State comparison of the mobility of FSP participants versus WIC participants.

Because none of the three States provided historical changes in identifiers for both FSP and WIC, record linkage results could be biased. Loss of data due to overwriting policies increases the potential for false negatives – that is, a failure to find a match when a match exists. The low rates of change for most personal identifiers suggest that this is not a large problem. However, changes in last name for WIC participants can pose a problem in establishing a match to FSP because marriage is a primary trigger for exit from FSP (Blank, 1993). Women who participate in both FSP and WIC but exit FSP after marriage may be observed with their maiden name in FSP and married name in WIC.

The high rates of change in contact information must be taken into account when specifying criteria for establishing a match between FSP and WIC records. For example, criteria can be specified such that corresponding address information helps to establish a match, while non-corresponding address information does not preclude a match.

Participation Dynamics Within FSP and WIC

The data for this study were collected retrospectively, resulting in a three-year snapshot of FSP and WIC caseloads, except for Kentucky FSP. For individuals observed in these files, participation histories may be truncated because participation may have started prior to the sample period (left-truncation) or continued after the sample period (right-truncation). Only one cohort of children is observed for a 36-month period from birth – infants born in January 2000.

Tables 14 and 15 show the distributions of FSP and WIC participants by total months of participation during the 3-year period. The total months need not be continuous, which means that the distributions contain participants with either single spells or multiple spells within the period. Infants and children are categorized according to age when first observed in the data file. The tables show unconditional and conditional percents. Unconditional percents are calculated over all participants observed during the three-year period. Conditional percents are calculated using a conditioned sample consisting of participants first observed more than 6 months (or 12 or 24 months) prior to the end of the sample period. For example, the conditional percent of participants with at least 12 months of participation is calculated over all participants who entered the program more than 12 months before the end of the sample period. Conditional percents provide better estimates of the distribution of months of participation in the face of right-truncation.

[42] Among Florida FSP participants with a change in telephone number, 2 percent changed area code without changing the remaining 7-digits of the telephone number. Iowa FSP data did not include area codes.

Table 14—FSP participant dynamics: Number of months of participation during 2000-2002

	Unconditional percent		Conditional percent			
	Florida	Iowa	Florida		Iowa	
	Percent		Conditioned Sample Size	Percent[1]	Conditioned Sample Size	Percent[1]
Total W-I-C						
Number participants	1,194,425	180,171	–	–	–	–
Cumulative duration of						
> 6 months	60.0	59.3	1,082,163	66.2	160,894	66.4
> 12 months	38.5	38.9	950,731	48.4	142,600	49.2
> 24 months	16.4	15.7	711,949	27.5	102,335	27.7
Ever received cash assistance	33.8	55.9	–	–	–	–
Women of childbearing age						
Number participants	710,771	109,037	–	–	–	–
Cumulative duration of						
> 6 months	58.3	58.2	645,609	64.2	97,894	64.9
> 12 months	37.0	38.3	571,808	46.0	87,396	47.8
> 24 months	15.8	16.2	433,888	25.8	63,711	27.7
Ever received cash assistance	29.8	49.3	–	–	–	–
Infants						
Number participants	199,759	28,685	–	–	–	–
Cumulative duration of						
> 6 months	58.3	58.4	173,588	67.1	24,194	69.3
> 12 months	34.7	35.3	140,594	49.3	19,846	51.0
> 24 months	11.2	10.2	87,221	25.8	11,495	25.4
Ever received cash assistance	41.4	69.0	–	–	–	–
Children						
Number participants	283,895	42,449	–	–	–	–
Cumulative duration of						
> 6 months	65.3	62.5	262,966	70.5	38,806	68.4
> 12 months	45.0	42.8	238,329	53.6	35,358	51.4
> 24 months	21.6	18.3	190,840	32.1	27,129	28.6
Ever received cash assistance	38.5	64.1	–	–	–	–

[1] The denominators of the conditional percents are the conditioned sample sizes, which are the numbers of FSP participants first observed more than 6 months (or 12 or 24 months) prior to the end of the sample period. For example, the number of Florida FSP participants who had cumulative durations greater than 6 months was 716,297 (i.e., 59.97 percent of all 1,194,425 participants), which represents 66.2 percent of the 1,082,163 in the conditioned sample.

– Not applicable

Table 14 shows the percent of FSP participants with greater than 6, 12, and 24 months of participation on both a conditional and unconditional basis.[43] The unconditional percentages indicate that about 60 percent of FSP participants in Florida and Iowa are observed with more than 6 months of participation within a three-year period, 39 percent have more than 12 months participation, and 16 percent have more than 24 months participation. Conditional percentages indicate that 66 percent, about 50 percent, and 28 percent have more than 6, 12, and 24 months of participation, respectively.

[43] The percent of FSP participants with duration in a particular range can be obtained from the difference in cumulative percents. For example, the percent of Florida FSP participants with 12 to 24 months of participation is equal to the percent with ">12 months" less the percent with ">24 months", which is 38.5 – 16.4 = 22.1 percent.

Table 15—WIC participant dynamics: Number of months of participation during 2000-2002

	State					
	Florida		Iowa		Kentucky	
	Unconditional percent	Conditional percent	Unconditional percent	Conditional percent	Unconditional percent	Conditional percent
Total WIC						
Sample size[1]	981,464	856,511	163,649	145,252	329,778	295,585
Cumulative duration of						
> 6 months	76.0	87.1	83.6	94.2	83.6	93.3
> 12 months	39.8	53.5	47.3	61.2	46.7	59.6
> 24 months	10.5	22.5	13.5	26.1	10.1	20.2
Percent with multiple certifications	53.3	64.2	58.6	69.6	53.5	63.0
Percent with continuous multiple certifications[2]	29.5	35.1	43.0	50.7	41.0	48.0
Women						
Sample size[1]	336,940	289,509	52,441	45,831	102,262	90,290
Cumulative duration of						
> 6 months	64.5	75.1	82.8	94.8	82.9	93.9
> 12 months	20.7	29.3	33.5	45.3	36.1	48.3
> 24 months	0.7	1.8	2.0	4.5	2.0	4.4
Percent with multiple certifications	54.8	62.6	59.5	67.9	59.5	68.1
Percent with continuous multiple certifications[2]	38.5	43.6	47.8	54.2	49.0	55.7
Infants						
Sample size[1]	352,390	290,956	52,391	43,412	108,463	90,835
Cumulative duration of						
> 6 months	80.4	97.4	82.5	99.6	82.4	98.4
> 12 months	46.6	70.6	52.9	79.7	48.1	71.0
> 24 months	11.8	36.4	14.5	44.2	7.6	22.2
Percent with multiple certifications	43.0	62.2	47.4	69.3	35.5	50.2
Percent with continuous multiple certifications[2]	22.9	32.6	33.4	48.2	28.0	39.2
Children						
Sample size[1]	292,134	276,046	58,817	56,009	119,053	114,460
Cumulative duration of						
> 6 months	83.9	88.8	85.3	89.6	85.4	88.8
> 12 months	53.6	60.4	54.6	60.7	54.6	59.8
> 24 months	20.2	27.9	22.8	30.4	19.5	28.0
Percent with multiple certifications	64.0	67.6	67.9	71.1	64.6	67.3
Percent with continuous multiple certifications[2]	27.0	28.4	47.3	49.4	46.0	47.8

[1] The sample size for unconditional percents is the total number of persons participating in WIC at any time during the three-year period (e.g., 981,464 in Florida). The sample size for conditional percents is different for each measure, but can be derived from the table. The sample size shown for conditional percents is the conditioned sample size is for duration > 6 months. Conditioned sample size is equal to (unconditional percent) / (conditional percent) x (unconditional sample size). For example, 0.76/0.871 x 981464 = 856,386, which differs from 856,511 shown in table due to rounding of percents.

[2] Continous participation is defined by a "next" certification date within 30 days of the previous termination date, for all certification periods.

Within a three-year period, children have more months of FSP participation than women and infants. For example, using the conditional figures, 32 percent of children in Florida FSP participated longer than 24 months, compared with 26 percent of women and infants (the difference between children and others is smaller in Iowa). Intra-group differences between unconditional and conditional percents indicate that the impact of right-truncation is greatest for infants.

Duration of WIC participation is shown in table 15. Between 87 and 94 percent of WIC participants in the three States were enrolled in WIC for more than 6 months (on a conditional basis); 53 to 61 percent were enrolled at least 12 months; and 20 to 26 percent were enrolled more than 24 months. Table 15 shows that over 95 percent of WIC infants in all three States were enrolled more than 6 months, compared with 75-95 percent of women and 89-90 percent of children. Durations of more than 6 months are consistent with regulations allowing infants to be certified up until their first birthday, while women and children may be re-certified after an initial 6-month period. The conditional percent of WIC participants with multiple certifications (shown in table) is slightly higher for children (67-71 percent across States) compared with women (63-68 percent) and infants (50-69 percent).

WIC women have the shortest participation durations in the data, consistent with WIC eligibility that is limited to periods around childbirth. Table 15 shows that it is very unlikely for women to be enrolled in WIC more than 24 months within a 36-month period. WIC durations for women vary by State; the percent enrolled more than 12 months is 20 percentage points higher in Iowa and Kentucky, compared with Florida. WIC infants are initially enrolled in WIC up until their first birthday, and then, if still eligible, may be recertified as children.[44] The conditional percents show that 71 to 80 percent of infants were enrolled more than 12 months (i.e., re-enrolled as children) and fewer than half remain in WIC more than 24 months (the percents range from 22 to 44 percent across States). Compared with infants, those initially observed as children have somewhat lower conditional percentages of enrollment for at least 12 months (60-61 percent vs. 71-80 percent) and lower conditional percentages of enrollment for at least 24 months (28-30 percent vs. 22-44 percent).

Comparison of tables 14 and 15 shows that, within a three-year time period, women participate in FSP longer than in WIC – about 27 percent of FSP women and less than 5 percent of WIC women participate longer than 24 months. In contrast, infants participate in WIC longer than in FSP –about 26 percent of FSP infants and 22-44 percent of WIC infants participate longer than 24 months. Children are more likely to participate in WIC for more than 6 months, compared with FSP. But duration of at least 12 or 24 months is comparable for children in FSP and WIC.

[44] Most WIC infants are enrolled during the first three months after birth (91 percent in Iowa, 89 percent in Kentucky, and 83 percent in Florida).

BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2001/04

Quality of Very Large Databases

William E. Winkler
Statistical Research Division
Methodology and Standards Directorate
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: 07/25/2001

Quality of Very Large Databases

William E. Winkler, U.S. Bureau of the Census, william.e.winkler@census.gov

Abstract

Analyses and data mining of large computer files are affected by the quality of the information in the files. For large population registers and for files that are created by merging two or more files, duplicate entries must be identified. Duplicate identification can depend on record linkage software that can deal with name, address, and date-of-birth data containing many typographical errors. Quantitative and qualitative data must be edited to assure that mutually contradictory or missing items are changed automatically and quickly. This paper describes computational methods and software that are suitable for groups of files where individual files contain between 1 million and 4 billion records.

Keywords: record linkage, editing, imputation, data mining

1. INTRODUCTION

There is significant interest in improving the quality of registers, groups of files that might be used in creating data warehouses, merging lists, and identifying duplicates within lists. With the substantial increases in computational power and storage, more groups are able to attempt projects in which single files or groups of files are cleaned to identify and correct erroneous information such as duplicates and contradictory information.

This paper consists of a number of subsections. The second section gives background and covers examples of how duplicates can arise even in well-designed situations. The third section gives background on two methods for improving the quality of data files. The first method is for identifying duplicates. It is based on the Fellegi-Sunter model of record linkage (Fellegi and Sunter 1969). The second set of methods is for assuring the logical consistency of information within a record or group of records. They are based on the Fellegi-Holt model of statistical data editing (Fellegi and Holt 1976). The fourth section covers further examples in which truly enormous files having possibly billions (10^9) of records may be processed. The final section consists of concluding remarks.

2. BACKGROUND AND INTRODUCTORY EXAMPLES

A *duplicate* is a record that cannot be correctly linked with another record to which it corresponds. In a population register, if a record is not given a correct unique identifying number (UID), then it may not be properly connected with other records that are associated with an individual. There are ways to minimize error. The most important is to have a check digit or check digits that are added at the end of the UID. A single check digit can help eliminate 90 percent of erroneous keying and transcription errors and a double check digit can eliminate 99 percent.

If there are no check digits, other quality control methods may not be entirely effective. It is estimated that 2-3 percent of the Social Security Numbers (SSNs) that are used in the California Quarterly Employment Files are in error in any given quarter. Over a period of twenty years, the records with each individual can expect to contain at least two errors where the SSN has been miskeyed or transcribed improperly. The SSN does not have a check digit. For the State of California in the U.S., the twenty-year quarterly employment file contains 1.1 billion records that need to be unduplicated.

The methods of unduplicating the file may involve use of name, date-of-birth information if available, employer, address, and SSN. Each of the identifying fields such as name may contain typographical error. Some of the identifying fields such as employer and address are time dependent. They may not be unique over a period of years.

In some situations, a group may wish to combine multiple files into a large merged database such as a data warehouse. If the files come from a variety of sources, then the files are unlikely to have a UID that allows them to be easily linked. Typically, name and address information may be all that is available for linkages. If a file has been poorly maintained, then the name and address information may be difficult or nearly impossible to use for linkage. Name and a full date of birth are better identifying information than name and address. Even with well maintained business lists, it may be difficult to keep track of the different name variations and different addresses associated with a business over a period of years.

The following table illustrates the difficulty with unduplicating using name information. In line 1, Janice Mary Smith is the current legal married name. The second line, Jan Smith, contains the nickname Jan and might be the form that appears on most lists. In parts of the U.S., it is still possible that many women are listed as in line 3. The form of the name is essentially the husband's name. The fourth line contains two minor typographical errors of the name in line 2. The fifth line is the maiden name that she used prior to being married.

Table 1. Free-form Name Fields in U.S. Lists

1. Janice Mary Smith
2. Jan Smith
3. Mrs. John Robert Smith
4. Jon Smuth
5. Janice Mary Brown

The above names cannot be used for exact character-by-character matching. Name-parsing software (described in the next section) can break a name into components that allow comparison of corresponding components. To facilitate matching, both the married and maiden names need to be maintained in the large administrative list if it is used over a period of years. The Social Security Administration carries the major legal variants of names in its files. Each name is in a separate record that contains the correct SSN. A flag in a separate field indicates what name variant is the currently used version. A name variation such as 3 is essentially unusable. It may be usable if there is auxiliary information that variation 3 corresponds to other variations such as 1. Without additional corroborating information such as address or date-of-birth, it is generally impossible to match on the first name Janice and the last name Smith because they are so common. There are three million individuals with the last name Smith in the U.S. There are 60,000 John Smiths.

The following table indicates variants of addresses. The first three variants all are intended to be actual location where the individual lives at a given point in time. The fourth variant might be a Post Office Box where the individual receives some of her mail. The fifth variant might be the address of an accountant that files the tax forms for the individual. Again, address parsing and standardization software can help with the first three variants of the address. The only way to deal with the last two variants of the address in to carry them as auxiliary information in the address file associated with the individual. Because address information is highly time dependent (in some of the areas of the U.S., twenty percent of the individuals move each year), tracking address information is very difficult.

Table 2.

1. 123 East Main Street
2. 123 E. Main St.
3. 123 E. Main Street, Unit 1
4. P.O. Box 5465
5. 6879 Maple Avenue, Suite 1001

Date-of-birth (dob) information is available in many different forms as illustrated in Table 3. Line 2 is the European convention of day first, whereas line 1 has the U.S. variant with month first. Line 3 is the variant that records the year as two digits, and line 5 is the variant that records the dob in the MMDDYYYY variant in which the year is given four digits. Line 4 has minor typographical errors in both month-of-birth and year-of-birth.

Table 3.

1. January 5, 1960
2. 5 January 1960
3. 01/05/60
4. 01/06/69
5. 01/05/1960

With many U.S. lists, the full date-of-birth is missing with over half of the records. The year-of-birth may all that is available. With a rare U.S. name such as Callahan Zabrinsky, a minor typographical error in the dob field such as given in line 4 of Table 3 may still allow correct matching. With the 60,000 John Smiths, any typographical error in dob is likely to match a John Smith with the incorrect John Smith.

One of the main uses of a large administrative list such as a national health register is in matching it with various hospital, doctor, and regional health records. Each of the lists would need to be statistically edited and imputed to remove or eliminate inconsistent or missing information. For instance, the codes of female for sex and prostrate cancer for disease are inconsistent. Other information in a record might be used to change the sex code to male. More information related to registers in available in Gill (2001).

For various economic analyses, several files might be combined using the name, address, and other information. The merged files might contain quantitative and other data from the source files. Any analyses would need to be corrected for matching error. Some of the quantitative information might require editing and imputing both prior and after matching.

3. METHODS

This section describes methods for record linkage and for statistical data editing and imputation. All of the methods have been implemented and used at National Statistical Institutes. With a few exceptions, most of the software can be used on a variety of computer systems.

3.1. Record Linkage Methods

Fellegi and Sunter (1969) introduced a formal mathematical foundation for record linkage. Their model makes rigorous concepts introduced by Newcombe et al. (1960). Two files A and B are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

If $R > T_\mu$, then designate pair as a match.
If $T_\lambda \leq R \leq T_\mu$, then designate pair as a possible match and
hold for clerical review. (2)
If $R < T_\lambda$, then designate pair as a nonmatch.

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on the rates μ and λ of false matches and false nonmatches, respectively. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small.

Pairs with weights above the upper cut-off are referred to as *designated matches*. Pairs below the lower cut-off are referred to as *designated nonmatches*. The remaining pairs are referred to as *designated potential matches*. The probabilities $P(\gamma\in\Gamma \mid M)$ *and* $P(\gamma\in\Gamma \mid U)$ *are referred to as the m-probability and the u-probability, respectively.* In practice, the probabilities may be difficult to estimate.

The *matching parameters or probabilities* given in the numerator and denominator of (1) can be estimated based on priori experience or via an optimization method such as the EM algorithm (see e.g., Winkler 1995). With very large register files, optimal parameters can be estimated prior to matching and will work well when smaller files are matched against the register (Gill 1999, 2001). If good matching parameters are not available prior to matching, then the parameters can be re-estimated based on a review of the initial matching results.

String comparators are needed because of the large amount of typographical error in files. In some geographic subregions of a major Decennial Census application, as much as 25 percent of first names and 15 percent of last names of records that are true matches contain typographical errors. Typographical error is best dealt with via string comparators that return values between 1 (perfect character-by-character agreement) and 0 (pure disagreement). Table 4 compares the string comparator values returned by the Jaro and Winkler string comparators (see e.g. Jaro 1989, Winkler 1995) with a bigram string comparator that is widely used in computer science. The likelihood ratio in (1) is adjusted for the string comparator values that are strictly between 0 and 1.

Table 4. Comparison of String Comparators Using Last Names, First Names, and Street Names

Two strings		String comparator values		
		Jaro	Winkler	Bigram
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.700
DUNNINGHAM	CUNNIGHAM	0.896	0.896	0.889
NICHLESON	NICHULSON	0.926	0.956	0.625
JONES	JOHNSON	0.790	0.832	0.204
MASSEY	MASSIE	0.889	0.933	0.600
ABROMS	ABRAMS	0.889	0.922	0.600
HARDIN	MARTINEZ	0.000	0.000	0.365
ITMAN	SMITH	0.000	0.000	0.250
JERALDINE	GERALDINE	0.926	0.926	0.875
MARHTA	MARTHA	0.944	0.961	0.400
MICHELLE	MICHAEL	0.869	0.921	0.617
JULIES	JULIUS	0.889	0.933	0.600
TANYA	TONYA	0.867	0.880	0.500
DWAYNE	DUANE	0.822	0.840	0.200
SEAN	SUSAN	0.783	0.805	0.289
JON	JOHN	0.917	0.933	0.408
JON	JAN	0.000	0.000	0.000

Current record linkage software (Winkler 2000) is relatively fast in that it processes approximately 10,000 pairs of records per second. Some commercial software (see e.g. the listing at http://caravel.inria.fr/~galharda/cleaning.html) can be upwards as one third as fast. Most software requires that each input file be sorted by blocking criteria. Blocking criteria are a set of characteristics such as first and last name that every pair must agree exactly (i.e., character-by-character). Sorts that can be prohibitively expensive for a file of one billion records in terms of CPU time (6 days on a fast machine) and disk storage (3.0 terabytes for a 1.0 terabyte file). Software (Yancey and Winkler 2001) that gets around some of the limitations is described in section 4.

To properly match files using name and address information, the components of the names and the components of the addresses must be parsed into components that must be compared. Table 5 illustrates name parsing and standardization. The output is from general business name software (Winkler 1993) that also works well with certain types of person names.

Table 5. Examples of Name Parsing and Standardization

	Standardized
1.	DR John J Smith MD
2.	Smith DRY FRM
3.	Smith & Son ENTP

Parsed

	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith		Son	ENTP	

Addresses are considerably more difficult to standardize and parse because they represent far more differing patterns. There are many good commercial address standardization software packages available because of the wide-spread use of mailing lists. Table 6 illustrates examples of address-parsing and standardization subroutines developed by Beck (1994) that is in use at the U.S. Census Bureau.

Table 6. Examples of Address Parsing

	Standardized
1.	16 W Main ST APT 16
2.	RR 2 BX 215
3.	Fuller BLDG SUITE 405
4.	14588 HWY 16 W

Parsed (1)

	Pre2	Hsnm	Stnm	RR	Box
1.	W	16	Main		
2.				2	215
3.					
4.		14588	HWY 16		

Table 6 (continued)

	Parsed (2)				
	Post1	Post2	Unit1	Unit2	Bldg
1.	ST		16		
2.					
3.				405	Fuller
4.		W			

Porter and Winkler (1998) wrote generalized, parameter-driven software that calls the name and address standardization routines.

3.2. Statistical Data Editing and Imputation

A good overview of the principles of Statistical Data Editing is given in Granquist and Kovar (1997). A combination of macro editing can be used to target the largest and most important records for processing manually. The view is further described in De Waal et al. (2000). In some situations, there may be too much data to review clerically. For instance in the 1997 U.S. Census of Manufactures, 100,000 records (4% of 2.5 million records) may contain errors or missing data. Because most of the 100,000 records are associated with small businesses, an automated method can deal with those records. The records of the largest businesses are additionally given a semi-automated clerical review.

The Fellegi and Holt (1976) provided a mathematical model for statistical data editing in which all edits reside in easily maintained tables. In conventional editing, thousands of lines of if-then-else code need to be maintained and debugged. In a Fellegi-Holt system, the code of the main mathematical routines can be easily maintained. It is possible to check the logical consistency of the system prior to the receipt of data. In one pass through the data of an edit-failing record, it is possible to fill in and change values of variables so that the record satisfies all edits. If a complete set of implicit edits can be logically derived prior to editing, then the integer-programming routines that determine the minimal number of fields to change in a record are relatively fast. Implicit edits are those edits that can be logically derived from a set of explicitly defined edits. Generally, it is difficult to derive all implicit edits prior to editing (Garfinkel et al. 1986, Winkler 1997). When most of the implicit edits are available, an efficient way of determining the approximate minimal number of fields to change is described in Winkler and Chen (2001).

In the Fellegi-Holt model, a set of edits is a set of points determined by edit restraints. An edit is failed if a record intersects the set of points. Generally, discrete restraints have been defined for discrete data and linear inequality restraints for continuous data. For continuous x's,

$\Sigma_i \; a_{ij} \; x_j \; \leq C_j$ for j=1,2,...,n.

For discrete data,

{Age ≤ 15, marital status = Married}.

If a record r falls in the set of restraints defined by the edit, then the record fails the edit. It is intuitive that one field (variable) in a record r must be changed for each failing edit. There is a major difficulty. If fields associated with failing edits are changed, then other edits that did not fail originally will fail. Fellegi and Sunter (1976) showed that implicit edits provide information about edits that do not originally fail but may fail as a record is changed.

The SPEER97 system (Draper and Winkler 1997) for ratio editing and balancing (assuring that items add to totals) is relatively fast (1000 records per second). The DISCRETE edit system (Winkler 1997, Chen 1998) is also fast (1000 records per second). The SPEER97 system requires that most of the implicit edits be computed in advance. The DISCRETE system requires that all of the implicit edits be computed in advance. Both SPEER97 and DISCRETE have modules that assure that imputed records satisfy edits. SPEER97 is known to adequately process relatively large files in which a modest proportion of records have substantial error. As shown by Draper and Winkler (1997), 10,000 (0.4% of 2.5 million) records needed to have 6 or more variables imputed. Of the 10,000, 99.0% were imputed automatically in a manner that assured that the resultant record satisfied edits. Overall, 99.9% of the edit-failing records were imputed in a manner so that the resultant record satisfied edits.

Because computing implicit edits in advance is not always possible, other systems do most of the computation to determine the minimal number of fields to change "on-the-fly". The GEIS system of Statistics Canada (see e.g., Kovar and Winkler 1996) uses a variant of Chernikova's algorithm to perform general linear inequality editing. It processes approximately 10 records per second. A more sophisticated LEO system from Statistics Netherlands (De Waal 2000) simultaneously does linear inequality and general editing. LEO is contained in an edit/imputation system that includes an AutImp module for imputation and an ECS module for finding edit-passing records that are close to imputed records. AutImp does not impute records that satisfy edits. The overall edit/impute system may process as many as 5 records per second. The LEO system is at an early stage of development. Neither GEIS nor LEO/AutImp can assure that records satisfy edits. Both are intended for relatively small situations having 20 or fewer variables in which less than 6 variables need to be changed.

The dramatic reduction in resources by using a Fellegi-Holt type of system is illustrated by Garcia and Thompson (2000). They compared the AGGIES system (Todaro 1999) on a large capital expenditures survey. The edits for the survey are complicated because there are ratio edits and there is some nesting of balance equations. Ten analysts worked up to six months to clerically edit and impute the data. Their changes involved manually making changes and then determining whether the resultant changed record satisfied all edits. By iterating, the analysts were eventually able to produce a record that satisfied all edits. They changed three times as much data as the AGGIES system. The AGGIES system automatically edited and imputed the data in less than 24 hours.

Bankier's Nearest Neighbour Imputation Method (NIM) is an effective alternative edit/imputation methodology. NIM performs well when there are many high quality hot-deck donors (Bankier 1991, Bankier et al. 1997, Bankier 2000). Like pure Fellegi-Holt systems, edits reside in tables that are much more easily maintained than thousands of lines of if-then-else rules. NIM has been used effectively on Canadian and Brazilian censuses. Because NIM is the most thoroughly tested system, the system is likely to be more robust than other systems. It is sufficiently fast to process files with millions of records. The methodology is known to be consistent with the Fellegi-Holt model (Winkler and Chen 2001).

4. RECORD LINKAGE FOR EXCEPTIONALLY LARGE FILES

Many individuals believe that identifying duplicates is one of the most difficult of the data quality issues. For a large matching situation such as matching the main Social Security Administration file of 600 million records against the 2000 Decennial Census file of 300 million records, this may entail the detailed comparison of 600 trillion pairs of records. Matching must be done using name, address, and date-of-birth information because the Census file does not contain the Social Security Number. Matching is done on secure administrative-record machines having two additional sets of firewalls inside the main firewalls protecting Census Bureau computers. To match efficiently, the files are matched in a series of blocking passes. During a *blocking pass*, only those pairs agreeing on certain characteristics are considered. For instance, on one blocking pass, only those pairs agreeing on first and last name may be considered. Other characteristics such as dob and address are used to determine whether a pair is a match. On another pass, only those pairs agreeing on date-of-birth may be considered. Prior to each matching pass according to a given blocking criteria, the files must be sorted according to the blocking criteria. Whereas the string comparators are useful once a pair of records has been brought together, they cannot be used for bringing pairs together. Twelve blocking passes have been used in some applications. A sort of a file requires three times the storage of the file being sorted. To sort a 600 million record file of 0.7 terabytes necessitates 2.1 terabytes of storage. The sort can require 3 days on a fast machine. Ten pairs of sorts and associated matching passes can take more than 40 days CPU time and substantial disk storage for intermediate files. The slowest part of the process can sometimes be the amount of skilled programmer intervention that is needed for tracking steps of the processing, backing off intermediate files, and writing auxiliary programs needed for analysis and evaluation.

BigMatch software (Yancey and Winkler 2001) allows the matching of a relatively small file having between 1 million and 100 million records against a large file of 4 billion records. The software allows up to ten simultaneous blocking criteria. For the above situation, the Census file could be divided in three subsets of 100 million records and matched against the Social Security Administration File. For ten blocking criteria, the match would take less than three days (one day for each subset of the Census file). The overall disk space requirement might be as little as 3 terabytes. Very little special programmer intervention would be needed.

BigMatch software begins by storing the smaller file in memory. It proceeds to dynamically build the structures needed for the sort keys, sorts the file by successive sort keys, and stores summary information about the beginning of blocks and the location of individual records within the blocks. Once the data structures are created, matching can proceed. After a record from the large file is input, it is paired with the records in the second file. For each blocking criteria, two files are output. The first file contains the matching weight of the pair, summary information associated with the matching process, and the information from the two pairs that was used in computing the matching weight. The second file contains the record from the larger file that was matched against the smaller file. For each blocking criteria, a special reformatting program creates a printout of pairs by decreasing blocking weight. Another preprocessing programming determines, within each blocking criteria, the sizes of the largest blocks in the smaller file. If blocks are too large, then the blocking criteria can be modified.

A special version of the BigMatch software allows identification of duplicates within a file.

5. CONCLUDING REMARKS

With large registers and data warehouses that may contain a billion (10^9) or more records, there is increased need for methods that can identify duplicates within and across files and to statistically edit and impute for missing and contradictory data. This paper describes some of the fastest methods that have been implemented in software.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

Alvey, W. and Jamerson, B. (eds.) (1997), *Record Linkage Techniques -- 1997* (Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, in Arlington VA), Washington, DC: Federal Committee on Statistical Methodology (available at http://www.fcsm.gov).

Bankier, M. (1991), “Alternative Method of Doing Quantitative Variable Imputation,” Statistics Canada Memorandum.

Bankier, M., Houle, A.-M., Luc, M. and Newcombe, P. (1997), “1996 Canadian Census Demographic Variables Imputation,” *American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods*, 389-394.

Bankier, M. (2000), “2001 Canadian Census Minimum Change Donor Imputation Methodology,” U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Cardiff, UK, October 2000 (also available at http://www.unece.org/stats/documents/2000.10.sde.htm).

Beck, B. (1994), “Address Parsing Software,” unpublished computer system and documentation, U.S. Bureau of the Census.

Chen, B.-C. (1998), “Set Covering Algorithms in Edit Generation,”*American Statistical Association, Proceedings of the Section on Statistical Computing*, 91-96.

De Waal. (2000), “New Developments in Automatic Edit and Imputation at Statistics Netherlands,” U.N. Economic Commission for Europe Work Session on Statistical Data Editing, Cardiff, UK, October 2000 (also available at http://www.unece.org/stats/documents/2000.10.sde.htm).

Draper, L., and Winkler, W.E. (1997), “Balancing and Ratio Editing with the New SPEER System,” Statistical Research Division Report 97/05 (a shorter version appeared in American Statistical Association, Proceedings of the 1997 Section on Survey Research Methods, pp. 582-587).

Fellegi, I. P. and D. Holt, (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17-35.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.

Garcia, M. and J. E. Thompson (2000), "Applying the Generalized Edit/Imputation System AGGIES to the Annual Capital Expenditures Survey," International Conference on Establishment Surveys, II, Buffalo, NY, June 2000, to appear in the Conference Proceedings.

Gill, L. (1999), "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.

Gill, L. (2001), *Methods for Automatic Record Matching and Linking and their use in National Statistics*, National Statistics Methodology Series, London: National Statistics.

Granquist, L. and J. G. Kovar, (1997), "Editing of Survey Data: How much is Enough?" in *Survey Measurement in Data Quality*, New York: Wiley, pp. 415-435.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.

Kovar, J.G., and Winkler, W.E., (1996), "Editing Economic Data", *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 81-87 (a version is available at http://www.census.gov/srd/www/byyear.html as report rr00/04).

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

Porter, E. H. and W. E. Winkler (1998), "General Business Name and Address Parsing Software," unpublished computer system and documentation, U.S. Bureau of the Census.

Todaro, T. A. (1999), "Overview and Evaluation of the AGGIES Automated Edit and Imputation System," Room paper presented at the Conference of European Statisticians, 2-4 June, 1999, Rome, Italy.

De Waal, T., F. Van de Pol, and R. Rennsen (2000), "Graphical Macro Editing: Possibilities and Pitfalls" Proceedings of the International Conference on Establishment Surveys, II. 579-588.

Winkler, W. E. (1993), "Business Name Parsing Software," unpublished computer system and documentation, U.S. Bureau of the Census.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report rr94/05 available at http://www.census.gov/srd/www/byyear.html as report rr94/05).

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W. E. (1997), "Set Covering and Editing Discrete Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 564-569 (longer version report 98/01 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1999), "The State of Statistical Data Editing," in *Statistical Data Editing*, Rome: ISTAT, 169-187 (also available at http://www.census.gov/srd/www/byyear.html as report rr99/01).

Winkler, W. E. (2000), Record linkage system with documentation, U.S. Bureau of the Census.

Winkler, W. E. and B.-C. Chen (2001), "Extending the Fellegi-Holt Model of Statistical Data Editing," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

Yancey, W. E. and W.E. Winkler (2001), "Bigmatch software," unpublished computer system and documentation, U.S. Bureau of the Census.

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. CXIX August 2004 Issue 3

THE CAUSES AND CONSEQUENCES OF DISTINCTIVELY BLACK NAMES*

ROLAND G. FRYER, JR. AND STEVEN D. LEVITT

In the 1960s Blacks and Whites chose relatively similar first names for their children. Over a short period of time in the early 1970s, that pattern changed dramatically with most Blacks (particularly those living in racially isolated neighborhoods) adopting increasingly distinctive names, but a subset of Blacks actually moving toward more assimilating names. The patterns in the data appear most consistent with a model in which the rise of the Black Power movement influenced how Blacks perceived their identities. Among Blacks born in the last two decades, names provide a strong signal of socioeconomic status, which was not previously the case. We find, however, no negative relationship between having a distinctively Black name and later life outcomes after controlling for a child's circumstances at birth.

I. INTRODUCTION

On May 17, 1954, the landmark Supreme Court decision in *Brown v. Board of Education of Topeka, Kansas* ruled, unanimously, that segregation in public schools was unconstitutional.

* We are grateful to George Akerlof, Susan Athey, Gary Becker, David Card, John Donohue, Esther Duflo, Edward Glaeser, Shoshana Grossbard-Shechtman, Daniel Hamermesh, Lawrence Katz, Glenn Loury, Douglas McAdam, Andrei Shleifer, and Steven Tadelis for helpful comments and suggestions. We thank seminar participants too numerous to mention for their comments and suggestions. We also thank two editors and three anonymous referees for detailed feedback that improved the paper. Michael Roh provided excellent research assistance. Financial support was provided by the National Science Foundation. A portion of this paper was written while Fryer was an NSF Post-Doctoral Fellow at the University of Chicago and Levitt was a fellow at the Center for Advanced Study in the Behavioral Sciences in Stanford. Correspondence can be addressed to Fryer at Harvard University, Littauer Center, Cambridge MA 02138 (e-mail: rfryer@fas.harvard.edu); or Levitt at Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637 (e-mail: slevitt@midway.uchicago.edu).

The Quarterly Journal of Economics, August 2004

This ruling paved the way for the fall of Jim Crow and large-scale desegregation. In the 1960s a series of further government actions were taken with the goal of achieving racial equality and integration, most notably the Civil Rights Act of 1964, Executive Order 11246 in 1965, and the Fair Housing Act of 1968. The civil rights movement arguably represents one of the most profound social transformations in American history [Woodward 1974; Young 1996].

Nonetheless, an enormous racial divide persists. There are large disparities between Blacks and Whites in the United States on many indicators of social and economic welfare including income [Bound and Freeman 1992; Chandra 2003; Heckman, Lyons, and Todd 2000; Smith and Welch 1989], educational achievement [Jencks and Phillips 1998], out-of-wedlock childbearing [Ventura and Bachrach 2000], health (see Kington and Nickens [2001]), and criminal involvement [Reno et al. 1997]. The degree of residential segregation by race, though lower today than in 1970, remains high [Cutler, Glaeser, and Vigdor 1999; Massey 2001].

Racial differences also persist, and in some cases have become even more pronounced, on a wide range of cultural dimensions including musical tastes [Waldfogel 2003], linguistic patterns [Wolfram and Thomas 2002], and consumption choices. For instance, the cigarette brand Newport has a 75 percent market share among Black teens, but just 12 percent among White teen smokers; 65 percent of White teens smoke Marlboro compared with only 8 percent of Blacks [Johnston et al. 1999]. Seinfeld, one of the most popular sitcoms in television history among whites, never ranked in the top 50 among Blacks. Indeed, of the top ten shows with the highest viewership among Whites during the 1999–2000 television season, only one show was also among the top ten for blacks: NFL Monday Night Football (Nielsen Media Research: http://www.nielsenmedia.com/ethnicmeasure/).

Understanding whether cultural differences are a *cause* of continued economic disparity between races is a question of great social importance. Cultural differences may be a cause of Black economic struggle if Black culture interferes with the acquisition of human capital or otherwise lowers the labor market productivity of Blacks (as argued in the culture of poverty paradigm in sociology; see Hannerz [1969], Lewis [1966], Riessman [1962], and implicitly, Anderson [1990]). For instance, high-achieving Black children may be ostracized by their peers for "acting white," potentially leading to lower investment in human capital

[Fordham and Ogbu 1986; Austen-Smith and Fryer 2003]. Speaking "Ebonics" may interfere with the ability to interact with White coworkers and customers, or disrupt human capital acquisition more directly [Orr 1989]. On the other hand, the presence of a Black culture may simply be the *consequence* of past and current segregation and economic inequality, but play no role in perpetuating economic disparity. If differences in tastes do not influence human capital acquisition or labor market productivity, then there is little reason to believe that such tastes will have a causal negative economic impact on Blacks. For example, "soul food" [Counihan and Van Esterik 1997] and traditional African-American spirituals [Jackson 1944] can be traced to the social conditions endured during slavery, but are unlikely to be causes of current poverty. Eliminating cultural differences in this scenario would have no overall impact on Black welfare relative to Whites.

A primary obstacle to the study of culture has been the lack of quantitative measures. In this paper we focus on one particular aspect of Black culture—the distinctive choice of first names—as a way of measuring cultural investments.[1] Our research builds upon a growing literature by economists devoted to understanding a diverse set of social and cultural phenomena [Akerlof and Kranton 2000; Berman 2000; Fryer 2003; Glaeser, Laibson, and Sacerdote 2002; Iannaccone 1992; Lazear 1999]. In contrast to these earlier papers, however, our contribution is primarily empirical.

Using data that cover every child born in California over a period of four decades, our analysis of first names uncovers a rich set of facts. We first document the stark differences between Black and White name choices in recent years.[2] For example, more than 40 percent of the Black girls born in California in recent years received a name that not one of the roughly 100,000 White girls born in California in that year was given.[3] Even

[1] Other than the audit studies of resumes discussed below, the only other economic analysis of name choices that we are aware of is Goldin and Shim [2003] which examines the issue of women retaining their maiden names at marriage. The seminal work on names outside of economics has been done in a series of papers by Stanley Lieberson and coauthors, culminating in Lieberson [2000].

[2] There are multiple dimensions along which a name can be considered "black" or "white." For example, Lieberson and Mikelson [1995] study distinctive patterns of phonemes that are characteristic of Black names. In this paper we study only one dimension of the issue: the relatively frequency with which Blacks and Whites choose a given name for their children.

[3] Lieberson and Mikelson [1995], using a sample of names from birth records in Illinois, find that approximately 30 percent of black baby girls born

among popular names, racial patterns are pronounced. Names such as DeShawn, Tyrone, Reginald, Shanice, Precious, Kiara, and Deja are quite popular among Blacks, but virtually unheard of for Whites.[4] The opposite is true for names like Connor, Cody, Jake, Molly, Emily, Abigail, and Caitlin. Each of those names appears in at least 2,000 cases (between 1989–2000), with less than 2 percent of the recipients Black.[5] Overall, Black choices of first names today differ substantially more from Whites than do the names chosen by native-born Hispanics and Asians.

More surprising, perhaps, is the time series pattern of Black first names. In the 1960s the differences in name choices between Blacks and Whites were relatively small, and factors that predict distinctively Black names in later years (single mothers, racially isolated neighborhoods, etc.) have much lower explanatory power in the 1960s. At that time, Blacks who lived in highly racially segregated neighborhoods adopted names that were almost indistinguishable from Blacks in more integrated neighborhoods and similar to Whites. Within a seven-year period in the early 1970s, however, a profound shift in naming conventions took place, especially among Blacks in racially isolated neighborhoods. The median Black female in a segregated area went from receiving a name that was twice as likely to be given to Blacks as Whites to a name that was more than twenty times as likely to be given to Blacks. Black male names moved in the same direction, but the shift was less pronounced. On the other hand, among a subset of Blacks encompassing about one-fourth of Blacks overall and one-half of those in predominantly White neighborhoods, name choices actually became more similar to those of Whites during this same period.

We argue that these empirical patterns are most consistent with a model in which the rise of the Black Power movement influenced Black identity. Other models we consider, such as ignorance on the part of Black parents who unwittingly stigma-

between 1920 and 1960 have unique names. Starting in the early 1960s, there was a remarkable increase in the prevalence of unique names, resulting in a peak in 1980 in which 60 percent of Black girls were given unique names. A similar, though less pronounced, phenomenon existed among Black boys.

[4] There are 463 children named DeShawn, 458 of whom are Black. The name Tyrone is given to 502 Black boys and only 17 Whites. 310 out of 318 Shanice's are Black, as are 431 out of 454 girls named Precious, and 591 out of 626 girls named Deja.

[5] The most extreme case is for the name Molly, in which only 9 of 2,248 children given the name are Black.

tize their children with such names, simple price theory models, and signaling models, all contradict the data in important ways.

The paper concludes by analyzing the relationship between distinctively Black names and life outcomes. Previous studies have found that distinctively Black names are viewed negatively by others (e.g., Busse and Seraydarian [1977]). Most persuasive are audit studies in which matched resumes, one with a distinctively Black/ethnic minority name and another with a traditionally White name, are provided to potential employers [Jowell and Prescott-Clarke 1970; Hubbick and Carter 1980; Brown and Gay 1985; Bart et al. 1997; Bertrand and Mullainathan 2003]. Such studies repeatedly have found that resumes with traditional names are substantially more likely to lead to job interviews than are identical resumes with distinctively minority-sounding names. The results suggest that giving one's child a minority name may impose important economic costs on the child. In our data, however, we find no compelling evidence of a negative relationship between Black names and a wide range of life outcomes after controlling for background characteristics. Although seemingly in conflict with prior audit studies using Black names on resumes, there are three interpretations of the data that reconcile the two sets of results: (1) Black names are used as signals of race by discriminatory employers at the resume stage, but are unimportant once an interview reveals the candidate's race, or (2) Black names provide a useful signal to employers about labor market productivity after controlling for information on the resume, or (3) names themselves have a modest causal impact on job callbacks and unemployment duration that we are unable to detect.

The remainder of the paper is structured as follows. Section II describes the data used in the analysis. Section III summarizes the basic patterns observed in the data. Section IV attempts to reconcile the stylized facts with a range of potential theories. Section V analyzes the relationship between names and life outcomes and attempts to reconcile our results with previous audit studies. Section VI concludes. A data appendix describes the details of our sample construction.

II. The Data

The data used in this paper are drawn from the Birth Statistical Master File maintained by the Office of Vital Records in the California Department of Health Services. These files provide (between 0–20). The fraction of Whites steadily shrinks as one moves from left to right in the figure. More than half of all Blacks have names that are at least four times as likely to be given to Blacks (between 81–100). For both races there is very little weight in the middle of the distribution (41–60), implying that there are relatively few individuals carrying names that are similarly likely for Blacks and Whites.

One might suspect that the sharp differences across races in Figure I(A) may in part be an artifact of how we construct our measure of Black names using the observed empirical distribution. In other words, we might miscategorize a name as being distinctively Black or White simply because for many names we observe only a few individuals with that name. Limiting the sample to names that appear at least twenty times in the data, however, does little to change the picture. Figure I(B), which is identical to Figure I(A) except that it compares the naming patterns of Whites with that of American-born Asians, further demonstrates that the result for Blacks is not an artifact of our measure. With the exception of a small fraction (approximately 10 percent) of the Asian population adopting names that are rare among Whites, name choices of American-born Asians strongly parallel White name choices. A comparison of native-born Hispanics and Whites in Figure I(C) shows differences in naming patterns among these two groups, although there is still substantially more overlap than for Blacks and Whites.[10]

An important racial difference in naming patterns is the greater usage of unique or nearly unique names in the Black community (see also Lieberson and Michelson [1995]). Figures II(A) and II(B) report, by race and gender, the number of children born in California in that same year (regardless of race) with that child's name. Remarkably, nearly 30 percent of Black girls receive a name that is unique among the hundreds of thousands of children born annually in California. Among Whites, that fraction is only 50 percent. Similarly, the fraction of unique names among Black boys is six times higher than for White boys, although only about half the rate of Black girls. The median Black child shares

[10] We have also compared the names chosen by Whites with different levels of education. There are systematic differences in name choices (larger, in fact, than between Asians and Whites overall), but these differences are much smaller than either the Black-White or Hispanic-White gap.

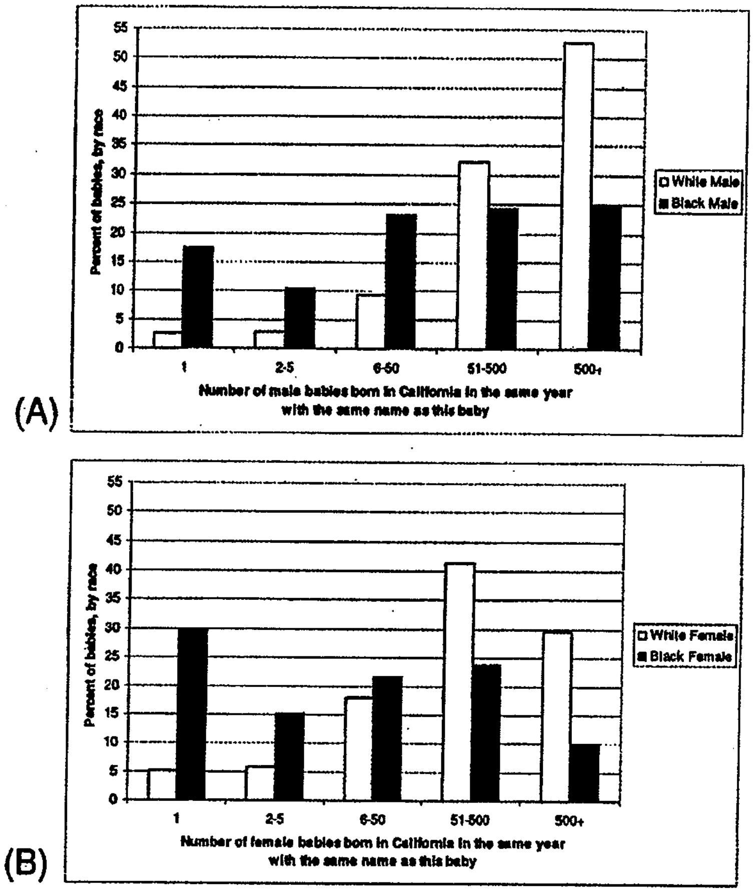


FIGURE II
Distribution of Male (A) and Female (B) Babies by How Many Share a Name, 1989–2000
(among children of all races born in California in a year)

his or her name with 23 other children; the number is almost fifteen times greater for Whites (351).[11]

Perhaps the most interesting findings in the data are the

[11] The differences between Blacks and Whites in Figure II are not primarily due to the fact that there are many more Whites than Blacks born in California each year. If we randomly select a subset of Whites equal in size to the number of Blacks born each year, a similar pattern of results persists.

the dismal science

A Roshanda by Any Other Name

How do babies with super-black names fare?

By Steven D. Levitt and Stephen J. Dubner

Posted Monday, April 11, 2005, at 6:32 AM ET

Which is more dangerous: a gun or a swimming pool? How much does campaign spending really matter? What truly made crime fall in the 1990s? These are the sort of questions raised—and answered—in the new book Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. *In today's excerpt, the first of two, authors Steven D. Levitt and Stephen J. Dubner explore the impact of a child's first name, particularly a distinctively black name. Tomorrow's excerpt shows how names work their way down the socioeconomic ladder.*

It has been well established that we live in an age of obsessive, even competitive, parenting. The typical parent is led to believe that her every move will greatly influence her child's future accomplishments. This belief expresses itself in the first official act a parent commits: giving the baby a name. Many parents seem to think that a child will not prosper unless it is hitched to the right one; names are seen to carry great aesthetic and even predictive powers.

This might explain why, in 1958, a New York City father named Robert Lane decided to call his baby son Winner. The Lanes, who lived in a housing project in Harlem, already had several children, each with a fairly typical name. But this boy—well, Robert Lane apparently had a special feeling about him. Winner Lane: How could he fail with a name like that?

Three years later, the Lanes had another baby boy, their seventh and last child. For reasons that no one can quite pin down today, Robert decided to name this boy Loser. Robert wasn't unhappy about the new baby; he just seemed to get a kick out of the name's bookend effect. First a Winner, now a Loser. But if Winner Lane could hardly be expected to fail, could Loser Lane possibly succeed?

Loser Lane did in fact succeed. He went to prep school on a scholarship, graduated from Lafayette College in Pennsylvania, and joined the New York Police Department, where he made detective and, eventually, sergeant. Although he never hid his name, many people were uncomfortable using it. To his police colleagues today, he is known as Lou.

And what of his brother? The most noteworthy achievement of Winner Lane, now in his late 40s, is the sheer length of his criminal record: more than 30 arrests for burglary, domestic violence, trespassing, resisting arrest, and other mayhem.

These days, Loser and Winner barely speak. The father who named them is no longer alive. Though he got his boys mixed up, did he have the right idea—is naming destiny? What kind of signal does a child's name send to the world?

These are the sort of questions that led to "The Causes and Consequences of Distinctively Black Names," a research paper written by a white economist (Steven Levitt, a co-author of this article) and a black economist (Roland G. Fryer Jr., a young Harvard scholar who studies race). The paper acknowledged the social and economic gulf between blacks and whites but paid particular attention to the gulf between black and white culture. Blacks and whites watch different TV shows, for instance; they smoke different cigarettes. And black parents give their children names that are starkly different than white children's.

The names research was based on an extremely large and rich data set: birth-certificate information for every child born in California since 1961. The data covered more than 16 million births. It included standard items like name, gender, race, birthweight, and the parents' marital status, as well as more telling factors: the parents' ZIP code (which indicates socioeconomic status and a neighborhood's racial composition), their means of paying the hospital bill for the birth (again, an economic indicator), and their level of education.

The California data establish just how dissimilarly black and white parents have named their children over the past 25 years or so—a remnant, it seems, of the Black Power movement. The typical baby girl born in a black neighborhood in 1970 was given a name that was twice as common among blacks than whites. By 1980, she received a name that was *20* times more common among blacks. (Boys' names moved in the same direction but less aggressively—likely because parents of all races are less adventurous with boys' names than girls'.) Today, more than 40 percent of the black girls born in California in a given year receive a name that not *one* of the roughly 100,000 baby white girls received that year. Even more remarkably, nearly 30 percent of the black girls are given a name that is unique among every baby, white and black, born that year in California. (There were also 228 babies named Unique during the 1990s alone, and one each of Uneek, Uneque, and Uneqqee; virtually all of them were black.)

What kind of parent is most likely to give a child such a distinctively black name? The data offer a clear answer: an unmarried, low-income, undereducated, teenage mother from a black neighborhood who has a distinctively black name herself. Giving a child a super-black name would seem to be a black parent's signal of solidarity with her community—the flip side of the "acting white" phenomenon. White parents, meanwhile, often send as strong a signal in the opposite direction. More than 40 percent of the white babies are given names that are at least four times more common among whites.

So, what are the "whitest" names and the "blackest" names? Click here for the top 20 each for girls and here for the top 20 each for boys. (For the curious, we've also put together a list of the top 20 crossover names—the ones that blacks and whites are most likely to share.) And how much does your name really matter? Over the years, a series of studies have tried to measure how people perceive different names. Typically, a researcher would send two identical (and fake) résumés, one with a traditionally white name and the other with an immigrant or minority-sounding name, to potential employers. The "white" résumés have always gleaned more job interviews. Such studies are tantalizing but severely limited, since they offer no real-world follow-up or analysis beyond the résumé stunt.

The California names data, however, afford a more robust opportunity. By subjecting this data to the economist's favorite magic trick—a statistical wonder known as regression analysis—it's possible to tease out the effect of any one factor (in this case, a person's first name) on her future education, income, and health.

The data show that, on average, a person with a distinctively black name—whether it is a woman named Imani or a man named DeShawn—does have a worse life outcome than a woman named Molly or a man named Jake. *But it isn't the fault of his or her name.* If two black boys, Jake Williams and DeShawn Williams, are born in the same neighborhood and into the same familial and economic circumstances, they would likely have similar life outcomes. But the kind of parents who name their son Jake don't tend to live in the same neighborhoods or share economic circumstances with the kind of parents who name their son DeShawn. And that's why, on average, a boy named Jake will tend to earn more money and get more education than a boy named DeShawn. DeShawn's name is an indicator—but not a cause—of his life path.

sidebar

Return to article

The 20 Whitest Girl Names

1. Molly
2. Amy
3. Claire
4. Emily
5. Katie
6. Madeline
7. Katelyn
8. Emma
9. Abigail
10. Carly
11. Jenna
12. Heather
13. Katherine
14. Caitlin
15. Kaitlin
16. Holly
17. Allison
18. Kaitlyn
19. Hannah
20. Kathryn

The 20 Blackest Girl Names

1. Imani
2. Ebony
3. Shanice
4. Aaliyah
5. Precious
6. Nia
7. Deja
8. Diamond
9. Asia
10. Aliyah
11. Jada
12. Tierra
13. Tiara
14. Kiara
15. Jazmine
16. Jasmin
17. Jazmin
18. Jasmine
19. Alexus
20. Raven

sidebar

Return to article

The 20 Whitest Boy Names

1. Jake
2. Connor
3. Tanner
4. Wyatt
5. Cody
6. Dustin
7. Luke
8. Jack
9. Scott
10. Logan
11. Cole
12. Lucas
13. Bradley
14. Jacob
15. Garrett
16. Dylan
17. Maxwell
18. Hunter
19. Brett
20. Colin

The 20 Blackest Boy Names

1. DeShawn
2. DeAndre
3. Marquis
4. Darnell
5. Terrell
6. Malik
7. Trevon
8. Tyrone
9. Willie
10. Dominique
11. Demetrius
12. Reginald
13. Jamal
14. Maurice
15. Jalen
16. Darius
17. Xavier
18. Terrance

19. Andre
20. Darryl

---

sidebar

Return to article

Most Popular Girl Crossover Names

1. Andrea
2. Whitney
3. Alicia
4. Kendra
5. Alexandria
6. Natasha
7. Tiffany
8. Brittany
9. Amber
10. Talia
11. Erika
12. Brianna
13. Ariel
14. Gabrielle
15. Veronica
16. Alana
17. Kyra
18. Ashley
19. Breanna
20. Erica

Most Popular Boy Crossover Names

1. Vincent
2. George
3. Troy
4. Christian
5. Martin
6. Corey
7. Brandon
8. Eric
9. Craig
10. Frank
11. Cameron
12. Shawn
13. Micah
14. Gregory
15. Nathaniel

16. Marc
17. Aaron
18. Dominic
19. Theodore
20. Isaac

---

sidebar

Return to article

Regression analysis is a powerful—if limited—tool that uses statistical techniques to identify otherwise elusive correlations. *Correlation* is nothing more than a statistical term that indicates whether two variables move together. It tends to be cold outside when it snows; those two factors are positively correlated. Sunshine and rain, meanwhile, are negatively correlated. Easy enough—as long as there are only a couple of variables. But with a couple *hundred* variables, things get harder. Regression analysis is the tool that enables an economist to sort out these huge piles of data. It does so by artificially holding constant every variable except the two he wishes to focus on, and then showing how those two co-vary.

In the case of a complicated data set that concerns, for instance, the test scores of 20,000 schoolchildren, it might help to think of regression analysis as performing the following task: converting each of those schoolchildren into a sort of circuit board with an identical number of switches. Each switch represents a single category of the child's data: his first-grade math score, his third-grade math score, his first-grade reading score, his third-grade reading score, his mother's education level, his father's income, the number of books in his home, the relative affluence of his neighborhood, and so on. Now a researcher is able to tease some insights from this very complicated set of data. He can line up all the children who share many characteristics—all the circuit boards that have their switches flipped the same direction—and then pinpoint the single characteristic they *don't* share. This is how he isolates the true impact of that single switch on the sprawling circuit board. This is how the effect of that switch— and, eventually, of every switch—becomes manifest. (From pages 161-162 of *Freakonomics.*)

Steven D. Levitt teaches economics at the University of Chicago and is a recipient of the John Bates Clark Medal, awarded every two years to the best American economist under 40. Stephen J. Dubner is a New York City journalist and author of two previous books: Turbulent Souls *and* Confessions of a Hero-Worshiper.

Article URL: http://www.slate.com/id/2116449/

The State of Record Linkage and Current Research Problems

William E. Winkler, U. S. Bureau of the Census[1]

ABSTRACT

This paper provides an overview of methods and systems developed for record linkage. Modern record linkage begins with the pioneering work of Newcombe and is especially based on the formal mathematical model of Fellegi and Sunter. In their seminal work, Fellegi and Sunter introduced many powerful ideas for estimating record linkage parameters and other ideas that still influence record linkage today. Record linkage research is characterized by its synergism of statistics, computer science, and operations research. Many difficult algorithms have been developed and put in software systems. Record linkage practice is still very limited. Some limits are due to existing software. Other limits are due to the difficulty in automatically estimating matching parameters and error rates, with current research highlighted by the work of Larsen and Rubin.

Keywords: computer matching, modeling, iterative fitting, string comparison, optimization

RÉSUMÉ

Cet article donne une vue d'ensemble sur les méthodes et les systèmes qui ont été mis en place pour le couplage d'enregistrements. Newcombe, qui développe une aproche nouvelle, et Fellegi et Sunter avec leur model mathématique, nous ont laisse les bases nécessaires pour un traitement moderne de la discipline du couplage d'enregistrement. Dans leur travail fondamental, Fellegi et Sunter ont introduit des méthodes puissantes pour l'estimation des paramètres sous-jacents, ainsi que des idées qui continuent d'influencer la pratique du couplage d'enregistrement. La recherche sur le couplage d'enregistrement se charactérise par une synergie de la statistique, de l'informatique, et de la recherche opérationnelle. Malgré l'intégragion sous formes de logiciels de plusieurs algorithmes difficiles, la pratique du couplage d'enregistrement n'en reste pas moins limitée. Cette limitation est due en partie aux defauts des logiciels eux-mêmes, mais aussi aux difficultés à estimer de facon systématique les paramètres sous-jacents ainsi que les taux d'erreurs encourues. Le problème de l'estimation automatique des taux d'erreurs encourues font l'object d'une recherche récente par Larsen et Rubin.

Mots Clés: couplage d'enregistrements, modeling, comparaison de chaîne de caractères, optimisation

1. INTRODUCTION

Record linkage is the methodology of bringing together corresponding records from two or more files or finding duplicates within files. The term record linkage originated in the public health area when files of individual patients were brought together using name, date-of-birth and other information. In recent years, advances have yielded computer systems that incorporate sophisticated ideas from computer science, statistics, and operations research. Some of the work originated in epidemiological and survey applications. Very recent work is in the related areas of information retrieval and data mining.

The ideas of modern record linkage originated with geneticist Howard Newcombe (Newcombe et al. 1959, 1962) who introduced odds ratios of frequencies and the decision rules for delineating

[1] William E. Winkler, Statistical Research Division, Room 3000-4, Bureau of the Census, Washington, DC, 20233-9100 USA, bwinkler@census.gov

matches and nonmatches. Newcombe's ideas have been implemented in software that is used in many epidemiological applications and often rely on odds-ratios of frequencies that have been computed a priori using large national health files. Fellegi and Sunter (1969) provided the formal mathematical foundations of record linkage. Their theory demonstrated the optimality of the decision rules used by Newcombe and introduced a variety of ways of estimating crucial matching probabilities (parameters) directly from the files being matched.

The outline of this paper is as follows. The second section provides more details on intuition about and the theoretical model for record linkage. Ideas of Newcombe have had the most important application in the development of national health files of individuals. The more general ideas of Fellegi and Sunter have been instrumental in estimating crucial matching parameters and estimating error rates for wide classes of lists. Methods for overcoming messy-data problems are described systematically in relation to the formal model of Fellegi and Sunter. In the third section, some of the basic research problems are covered. Although some of the problems have been (partially) solved for high quality pairs of lists, the solution methods do not easily extend to most matching situations. The fourth section describes three research areas that have arisen in recent years and depend heavily on record linkage ideas. The first is microdata confidentiality and associated re-identification methods. The second is analytic linking as introduced by Scheuren and Winkler (1993, 1997). *Analytic linking* refers to the merging and proper analysis of data (quantitative and discrete) taken from two or more files. The analysis is intended to adjust for the biases due to linkage error. The third presents some of the methods of information retrieval and machine learning as used by computer scientists in web search engines and data mining applications. Concluding remarks are given in the final section.

2. BACKGROUND ON RECORD LINKAGE

Howard Newcombe had crucial insights that led to computerized approaches for record linkage. The first was that the relative frequency of the occurrence of a value of a string such as a surname among matches and nonmatches could be used in computing a binit weight (score) associated with the matching of two records. The second was the scores over different fields such as surname, first name, age, etc. could be added to obtain an overall matching score. More specifically, he considered odds ratios

$$\log_2(p_L) - \log_2(p_F) \qquad (1)$$

where p_L is the relative frequency among links and p_F is the relative frequency among nonlinks. Since the true matching status is often not known, he suggested approximating the above odds ratio with the following ratio

$$\log_2(p_R) - \log_2(p_R)^2 \qquad (2)$$

where p_R is the frequency of a particular string (first, initial, birthplace, etc). If one matches a large universe file with itself, then the second ratio is a good approximation of the first ratio. Newcombe's ideas have been extended in a variety of ways (e.g., Newcombe et al., 1988, 1992, Gill 1999)

Fellegi and Sunter (1969) introduced a formal mathematical foundation for record linkage. To begin, notation is needed. Two files **A** and **B** are matched. The idea is to classify pairs in a product space **A** × **B** from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter, making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \varepsilon \Gamma \mid M) / P(\gamma \varepsilon \Gamma \mid U) \tag{3}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \varepsilon \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

If R > *UPPER*, then designate pair as a match.

If *LOWER* $\leq R \leq$ *UPPER*, then designate pair as a possible match and hold for clerical review. (4)

If $R <$ *LOWER*, then designate pair as a nonmatch.

The cutoff thresholds *UPPER* and *LOWER* are determined by a priori error bounds on false matches and false nonmatches. Rule (4) agrees with intuition. If $\gamma \varepsilon \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \varepsilon \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \varepsilon \Gamma$ consists primarily of disagreements, then ratio (3) would be small.

Pairs with weights above the upper cut-off are referred to as *designated matches* (or links). Pairs below the lower cut-off are referred to as *designated nonmatches* (or nonlinks). The remaining pairs are referred to as *designated potential matches* (or potential links).

If one considers a situation where there are three matching fields and only simple agree/disagree weights are considered, then a conditional independence assumption can be made to simplify computation.

P(agree first, agree last, agree age | M)

= P(agree first | M) P(agree last | M) P(agree age | M) (5a)

Similarly,

P(agree first, agree last, agree age = | U)

= P(agree first | U) P(agree last | U) P(agree age | U) (5b)

This conditional independence assumption must hold on all combinations of fields (variables) that are used in matching. The probabilities P(agree first | M), P(agree last | M), P(agree age | M), P(agree first | U), P(agree last | U), and P(agree age | U) are called *marginal probabilities*. P(| M) & P(| U) are called the m- and u-probabilities, respectively. The natural logarithm of the ratio R of the probabilities is called the *matching weight or total agreement weight*. The logarithms of the ratios of probabilities associated with individual fields (marginal probabilities) are called the *individual agreement weights*. The m- and u-probabilities are also referred to as *matching parameters*.

Fellegi and Sunter showed that it is possible to compute the unknown m- and u- probabilities directly in the 3-variable, conditional independence case. More generally, in the conditional independence situation, the parameters can be computed via a straightforward application of the EM algorithm (Winkler 1988). If the conditional independence assumption does not hold, then the parameters can be computed by generalized EM methods (Winkler 1988, 1989a, 1993b, Armstrong and Mayda 1993, see also Meng and Rubin 1993), by scoring (Thibaudeau 1993), and by Gibbs sampling (Larsen 1996, Larsen and Rubin 1999). The methods of Larsen and Rubin (1999) are the most general. These methods can yield more accurate matching parameters and better decision rules. These parameter-estimation methods do not always yield sufficiently accurate probability estimates for estimating record linkage error rates. An error-rate estimation method that is somewhat supplemental to these is due to Belin and Rubin (1995). Although the method of Belin and Rubin requires calibration data, it is known to work well in a narrow range of situations (Winkler and Thibaudeau, 1991; Scheuren and Winkler, 1993). The situations are those in which there is substantial separation of the curves of log frequency versus matching weight for matches and nonmatches.

Generally, good separation of curves occurs with high-quality lists of individuals containing only moderate amounts of typographical error and reasonable amounts of homogeneity in the characteristics respectively used in classifying pairs as matches and nonmatches. With some administrative lists and most agricultural and business lists, such homogeneity does not occur. For instance, if names or address do not standardize, then it is unlikely that true matches having nonstandardized names or addresses can be identified. If homogeneity holds, then most matches have similar characteristics within the group of matches. Most nonmatches have similar characteristics within the group of nonmatches. In some situations, difficulties with business lists can be dealt with via software loops that deal with list-specific nonhomogeneity. Each of the major departures from homogeneity due to severe typographical error must be dealt with via a separate software loop. Other departures from nonhomogeneity occur when either the class of matches or the class of nonmatches naturally divide into subclasses. For instance, when matching persons within household, the class of nonmatches naturally divides into those that agree on address (household characteristics) and those that do not. Some of the general methods for dealing with nonhomogeneity of identifying characteristics are described in Winkler (1993b). EM methods and ideas for dealing with one major type of nonhomogeneity similar to Winkler (1988, 1989, 1993b) have recently been applied to the general problem of text classification in machine learning and data mining by Nigam et al. (1999). The methods of Winkler are more general because they allow for dependencies of fields and convex constraints on probabilities (either class or marginal) that predispose estimates to subregions of the parameter based on prior knowledge from similar matching situations.

2.1 String Comparators

In many matching situations, it is not possible to compare two strings exactly (character-by-character) because of typographical error. Dealing with typographical error via approximate string comparison has been a major research project in computer science (see e.g., Hall and Dowling, 1980). In record linkage, one needs to have a function that represents approximate agreement, with agreement being represented by 1 and degrees of partial agreement being represented by numbers between 0 and 1. One also needs to adjust the likelihood ratios (3) according to the partial agreement values. Having such methods is crucial to matching. For instance, in a major census application for measuring undercount, more than 25% of matches would not have been found via exact character-by-character matching. Three geographic regions are considered in Table 1. The function Φ_n represents exact agreement when it takes value one and represents partial agreement when it takes values less than one. In the St Louis region, for instance, 25% of first names and 15% of last names did not agree character-by-character among pairs that are matches.

Table 1 Proportional Agreement by String Comparator Values Among Matches Key Fields by Geography

	StL	Col	Wash
First			
$\Phi_n=1.0$	0.75	0.82	0.75
$\Phi_n\geq 0.6$	0.93	0.94	0.93
Last			
$\Phi_n=1.0$	0.85	0.88	0.86
$\Phi_n\geq 0.6$	0.95	0.96	0.96

Jaro (1976, see also 1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic Jaro algorithm has three components: (1) compute the string lengths, (2) find the number of common characters in the two strings, and (3) find the number of transpositions. The definition of common is that the agreeing character must be within half the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value (rescaled for consistency with the practice in computer science) is:

$$\Phi_j(s1,s2) = 1/3(\ \#\text{common}/\text{str\_len1} + \#\text{common}/\text{str\_len2} + 0.5\ \#\text{transpositions}/\#\text{common}),$$

where s1 and s2 are the strings with lengths str_len1 and str_len2, respectively.

Using truth data sets, Winkler (1990b) introduced crude methods for modeling how the different values of the string comparator affect the likelihood in the Fellegi-Sunter decision rule. Winkler also showed how a variant of the Jaro string comparator Φ_n dramatically improves matching efficacy in comparison to situations when string comparators are not used. The variant employs some ideas of Pollock and Zamora (1984) in a large study for the Chemical Abstracts Service. They provided empirical evidence about how the probability of keypunch errors increased as the character position in a string moved to the right. Budzinsky (1993) in a review of twenty string comparators concluded that the methods of Jaro and Winkler worked second best and best, respectively. The Winkler string comparator Φ_n is used in the Generalized Record Linkage System software of Statistics Canada.

2.2 Heuristic Improvement by Forcing 1-1 Matching

Jaro (1989) introduced a linear sum assignment procedure (lsap) to force 1-1 matching because he observed that greedy algorithms often made erroneous assignments. A greedy algorithm is one in which a record is always associated with the corresponding available record having the highest agreement weight. Subsequent records are only compared with available remaining records that have not been assigned. In the following, the two households are assumed to be the same, individuals have substantial identifying information, and the ordering is as shown. A lsap algorithm causes the wife-wife, son-son, and daughter-daughter assignments correctly because it optimizes the set of assignments globally over the household. Other algorithms such as greedy algorithms can make erroneous assignments such as husband-wife, wife-daughter, and daughter-son.

HouseH1	HouseH2
husband	
wife	wife
daughter	daughter
son	son

$$\begin{array}{ccc} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \\ c_{41} & c_{42} & c_{43} \end{array}$$

4 rows, 3 columns
Take at most one in each
row and column

c_{ij} is the (total agreement) weight from matching the ith person from the first file with the jth person in the second file. Winkler (1994) introduced a modified assignment algorithm that uses 1/500 as much storage as the original algorithm and is of equivalent speed. The modified assignment algorithm does not induce a very small proportion of matching error (0.1-0.2%) that is caused by the original assignment algorithm.

2.3 Why the methods do not always work well.

The record linkage methods described above can perform well when there is little typographical variation and other forms of nonhomogeneity in the identifying characteristics of lists. The methods may not work well due to failures of the assumptions used in the models, lack of sufficient variables for matching, sampling or lack of overlap in lists, and extreme variations in the messiness of data. The idiosyncrasies of messy data are most easily described. Each of the following types of errors provides examples of situations where pairs of records will not have homogeneous identifying characteristics.

1. Records that do not address standardize.
2. Records that do not name standardize.
3. Records that have more information or missing matching variables.
4. Records that do not have easily comparable fields.

Name	Ralph Smith	R J Smith
Address	123 Main St	PO Box 9128
Age	54	50

If the PO Box address in the right-most column were replaced by a street address that corresponds almost exactly to the street address given in the second column, then it might be possible to accurately match. If R J Smith is actually Roberta Joan Smith, then the match would be in error. Inconsistencies of name and address information are typically even greater with agriculture and business lists. During name and address standardization, commonly occurring words such as

Mister, Road, Post Office Box, etc. are replaced by standardized spellings and the components of the names and addresses are placed in fixed locations. If standardization fails for a record, then automatic matching in software may be impossible. This is due to specific information needed for comparison and computing weights that is missing. If two lists of individuals are small samples, then we may not be able to match on certain commonly occurring names such as John Smith without substantial corroborating information. The difficulty of estimating the overlap of samples has most effectively been dealt with by Deming and Gleser (1959) in situations where there is no matching error. When there is matching error, the estimation can be more difficult.

3. BASIC RESEARCH PROBLEMS

The basic research problems have been open since the work of Newcombe et al. (1959) and Fellegi and Sunter (1969). Partial progress in solving the problems has occurred. The major difficulties in all situations have been determining how identifying information can be used and what the relative value of a field is in matching in comparison with other fields.

3.1 When can frequency-based matching improve over simple agree/disagree matching?

The ideas of frequency-based (value-specific) matching were introduced by Newcombe et al. (1959). Fellegi and Sunter (1969) gave two methods for computing frequency-based weights in the context of their formal model that have been extended by Winkler (1988, 1989). The basic idea is that agreements on rarely occurring values of a field (variable) are better at distinguishing matches than agreements on commonly occurring values of a field. The agreement on a rare value is also better than the general yes/no agreement (i.e., non-value-specific) on a field. For instance,

P(agree last name = 'Zabrinsky', agree first name 'Zbigniew' | M) >

P(agree last name, agree first name | M) > (6)

P(agree last name = 'Smith', agree first name 'James' | M) .

Reasonably correct frequencies are computed and used in matching. The intuition is that frequency-based weights given by the first and third probabilities in (6) are better able to delineate matches and nonmatches than the simple agree/disagree probabilities given in the second probability in (6). Names by themselves are seldom effectively used in matching. Additional fields such as components of the address, age or full date-of-birth, maiden name, sex, and race are also needed to reduce error rates to acceptable levels. In some early experiments, frequency-based matching often did better than simple agree/disagree matching. With the development of more sophisticated models for estimating agree/disagree matching parameters via the EM algorithm, simple agree/disagree weights sometimes performed better. The reason is due to the fact that, in many files, a moderate number of false matches agree on relatively rarely occurring names. In those situations, pairs that might be in the clerical review region given in (4) might move upward to the designated match region. If there is a substantial number of fields available for matching, then the redundancy provided by the extra fields can reduce matching error. If redundancy is sufficient to reduce matching error, then it seems likely that frequency-based matching is not needed. Raising the total agreement weights for pairs associated with less frequently values of a variable will not improve matching.

There are, nevertheless, a number of important situations when it is likely that frequency-based matching may be demonstrated to work at least as well as simple agree/disagree matching. The major situations all involve large national health files that have been significantly cleaned for typographical error and for which accurate probabilities can be computed a priori using true population counts. The research question is "Are there situations for which it can be shown that frequency-based matching improves over simple agree/disagree matching?" It seems that with many business lists, agriculture lists, and general administrative lists that frequency-based matching may not yield improvements because of the large amounts of typographical variation. These lists often have moderate to large proportions of records that fail standardization, have excessively high typographical error rates, and have only moderate overlap. If any one of these three situations occurs, then frequency-based matching may be seriously compromised.

3.2 What is the best method for estimating parameters under conditional independence when non-1-1 (or 1-1) matching is done?

Parameter estimates obtained under the conditional independence EM can be superior to other parameter estimates (Winkler, 1990b) and can be obtained more easily. The conventional methods estimate the marginal probabilities P(agree field | M) and P(agree field | U) directly using samples for which truth has been obtained via possibly time-consuming manual review. The estimates are obtained more easily because the known truth of matches on subsets is not needed (Winkler, 1988). The reason that the EM parameters work better is that they effectively represent the conditional probabilities such as the following

$$P(\text{agree field 1, agree field 2, agree field 3} \mid M) = \qquad (7)$$

$$P(\text{agree field 1} \mid M)\, P(\text{agree field 2} \mid \text{field 1}, M)\, P(\text{agree field 3} \mid \text{field 1, field 2}, M).$$

The EM algorithm decides what ordering of the fields in (7) is optimal in estimating the likelihoods. These probabilities implicitly perform a minor automatic adjustment for the lack of conditional independence. The EM algorithm still makes a *homogeneity* assumption because it assumes that the same ordering can be applied to all pairs conditional on whether they are a match or nonmatch. Because the EM-parameters are designed to maximize the likelihood, they produce better decision rules than the probabilities estimating under the conventional methods. The conventional parameters do not maximize the likelihood because of the strong conditional independence assumption that is made. Winkler (1990b) provided an exact comparison of decision rules using parameters obtained by the two estimation techniques. Caution in the automatic use of the EM-probabilities is needed because the EM may not exactly divide the set of pairs into two classes that correspond exactly to matches and nonmatches. The difficulty of having EM-determined classes that correspond to true matching classes has been addressed by Winkler (1993b) and by Nigam et al. (1999). The caution may not apply to conventionally estimated parameters because the clerical review can better assure that estimated parameters are consistent with model assumptions.

The EM probabilities are estimated using all pairs and often used in matching software that forces 1-1 matching. Although the mechanisms for forcing 1-1 matching are not explicitly accounted for, the probabilities are known to work well in those situations. The research question is "When can the EM-probabilities estimated under conditional independence be effectively used in 1-1 matching decision rules?" If marginal probabilities are conventionally estimated via samples, when can they be effectively used in 1-1 matching?

3.3 When does accounting for dependencies help in matching?

If conditional independence does not hold, then

$$P(\text{agree first name, agree last name} \mid M) \neq P(\text{agree first} \mid M)\, P(\text{agree last} \mid M)\,.$$

Decision rules that apply probabilities estimated under the conditional independence assumption may be suboptimal. Smith and Newcombe (1975 gave a modified decision rule that adjusts for the lack of dependence that have been effectively extended and applied by others (Gill, 1999). The modified decision rules are heavily dependent on the assumption that the adjustments based on a sample for which truth is known can be used in a variety of matching situations. The assumption is likely to be reasonable in situations of large national health files for which truth is known on a large subset. Winkler (1989a), Thibaudeau (1993), Armstrong and Mayda (1993), and Larsen and Rubin (1999) have all given formal models for estimating the record linkage parameters (probabilities) under general dependence models. Winkler (1989a) also showed that the values of matching parameters vary significantly from one list to another. The variation occurs even when the lists have the same matching variables and the same amount of overlap but represent different geographic regions. All of the authors have shown that the development of appropriate dependence models takes considerable skill and suitable software. They have also shown that probabilities estimated under dependence are more accurate. None of the authors has been able to show whether the new parameter-estimation method can be assured to yield appropriately good decision rules in actual record linkage software on a day-to-day basis. A basic research question is "For what types of files and matching situations can general dependence-based probabilities and decision rules improve matching?" There is still considerable empirical evidence that matching under the conditional independence assumption is effective in practice. Winkler (1993b, 1994) demonstrated that matching under the conditional independence assumption worked nearly as well as matching under more general dependency models in certain situations. The situations included population files having multiple individuals per household in which 1-1 matching was forced. Winkler (1994) did suggest accounting for dependencies might yield better automatic estimates of error rates.

3.4 What are (suitable) ways of estimating error rates?

The method of Belin and Rubin (1995) is currently the only method for automatically estimating record linkage error rates. Belin and Rubin were able to achieve highly accurate estimates (Winkler and Thibaudeau 1991, Scheuren and Winkler 1993) in a narrow range of situations. The situations generally involved population files where there was good separation between the matching weights associated with nonmatches and matches. If there is not good separation, then methods that use more information from the matching process may ultimately yield suitable estimates in a larger range of situations as suggested by Winkler (1994) and Larsen and Rubin (1999). The estimation methods and the means of evaluating the fits of the latent class models are quite difficult because the usual Chi-square methods do not work (Rubin and Stern, 1993). The basic research question is "How does one automatically estimate error rates?"

4. ADVANCED RESEARCH PROBLEMS

Three areas use methods and underlying models that are closely related to the basic ideas of record linkage. Confidentiality of microdata is most closely related because record linkage methods can be used for evaluating the re-identification risk in public-use files. Since the quantitative data in a public-use file are typically masked, new metrics for comparing quantitative data can yield higher re-identification rates. Analytic Linking is the methodology (Scheuren and Winkler 1997) for using not directly comparable data items to improve matching and to account for the effect of matching error in analyses. For instance, if one administrative file has receipts and another has income, an additional variable, predicted income, can be added to the first file to improve matching. The matching can also be improved by targeting outliers and systematic errors in the merged files in a manner that identifies likely false matches. Data mining and some models for information retrieval in computer science use Bayesian networks for classifying documents using free-form textual information. The representations in Bayesian networks from machine learning can be viewed as a special case of representations in the Fellegi-Sunter model. Recent advances in applying the EM algorithm in machine learning settings give insight into how to better use training data (if available), how to better structure the models, and how to use free-form text in a rigorous model.

4.1 Confidentiality

There is substantially increased need to supply researchers with large, general-purpose public-use files that can be used for a variety of analyses. Balancing the analytic needs are the requirements that agencies not release individually identifiable data. If a public-use file is created, then agencies must determine if the file meets analytic needs and is confidential. Record linkage methods (Winkler 1998) that employ new metrics for comparing somewhat related quantitative data provide a useful enhancement and yield higher re-identification rates than less sophisticated methods. If an agency can effectively determine that a small percentage of records might be re-identified, it can take additional precautions.

Methods for masking data are intended to make re-identification more difficult. Existing masking methods cover a variety of areas. Global recoding and local suppression (DeWaal and Willenborg 1996, 1998; Sweeney, 1999) have been successfully used to create public-uses files and other security procedures. The advantage of the methods is that available general software is often straightforward to apply. The associated research problems relate to how seriously analytic properties are compromised. Additive noise is known to preserve some of the analytic properties of files (Kim 1986, 1989; Fuller 1993). Research problems are whether general software can be developed and whether files are free of disclosures. Combinations of additive noise and limited swapping have been used by Kim and Winkler (1995) and Winkler (1998). Data perturbation methods (Tendick and Matloff 1994) are closely related to additive noise. The methods are good at preserving confidentiality and yielding totals on a number of subdomains that are consistent with unreleased confidential data. The basic research problems are whether the methods can be extended to preserved second order and higher statistics as the additive noise methods do. Camouflage (Gopal, Goes, and Garfinkel 1998) is a sophisticated method that returns intervals rather than point estimates for large classes of functions on arbitrary subdomains. A basic research question is whether these methods can produce the types of information that users of the public-use files need. Microaggregation (see e.g., Mateo-Sanz and Domingo-Ferrer 1998) is a method of replacing values of individual variables in ranges with means. The algorithms can be quite sophisticated. The research questions are: "Do these methods compromise analytic validity seriously?" and "What are re-identification rates with certain classes of files?" The most sophisticated methods involve models for re-identification risk and analytic properties of files. Fienberg, Makov and Sanil (1997) and Fienberg, Makov, and Steele (1998) have introduced some promising ideas that need extension to encompass different classes of data and to achieve computational tractability. With all of these methods the basic research question is "If analytic validity is preserved, then what is the re-identification rate?" If good source files for matching and suitable re-identification software are available to an intruder, then what is the re-identification rate?

4.2 Analytic Linking

Researchers often have the need to analyze large amounts of data that result from the merger of two or more administrative files in which unique identifiers are unavailable. Scheuren and Winkler (1993) showed how regression analyses might be adjusted for biases due to linkage errors. In the simplest situation of two variables, the dependent variable might be taken from one file and the independent variable from another file. If there is matching error, then the dependent and independent variables associated with false matches generally will not correspond as closely as those associated with true matches. The adjustments were highly dependent on accurate probabilities obtained by the methods of Belin and Rubin (1995). If error-rates are estimated accurately, then the bias-adjustments for matching error were reasonably accurate.

One administrative file may have a number of data fields (variables) that are correlated or otherwise related to a number of data fields in another administrative file. Scheuren and Winkler (1997) introduced analytic linking methods that place predictors in one file that can be used to improve matching with another file. After each matching pass, data are again modeled to refine the predictors. Through a series of iterations in which predictors and matching are improved, Scheuren and Winkler showed how matching could be performed in situations that were previously considered impossible. If matching error is low, then adjustment methods may not be needed (Scheuren and Winkler 1993). If matching error is moderate, then the adjustment method of Scheuren and Winkler (1993) may help. The basic research problems are "What are more generally applicable adjustments methods for matching error?" How can all of the information in two files be used? Scheuren and Winkler (1997) used simple predicted values that may not account for many types of matching error and may not be suitable as a global set of predictions. The work of Scheuren and Winkler has a strong visual component. Summary representations in graphs (images) are successively improved as erroneous data due to false matches are eliminated. Much of the erroneous data shows up as outliers that detract from the graph that would be obtained from the true model having no noise. If the underlying analytic model and the effects of some of the matching error are effectively modeled (i.e., learned), then the images associated with the process also improve. Improving the methods may involve advanced image resolution ideas (Besag et al. 1974, 1986, 1995; Geman and Geman 1984). Other improvements may be due to better modeling of the components of the iterative analytic linking process. Van Dyk (1999) has recently introduced methods for speeding up EM-type computations associated with hierarchical models that contain ideas that might improve the methods of Winkler and Scheuren. Although the specific types of speed-ups may not be needed, the insight that Van Dyk offered into modeling a large number of components of a process seems to be needed.

Winkler (1999) also indicated how large *bridging* files can be used to improve matching with two smaller files. A *bridging* file is a large universe file that approximately contains the two smaller files. Bridging files might be a large file such as the main Social Security Administration file of the U.S. population or a large credit database with associated information. Although the large bridging file does not generally have sufficient information for matching all records in the smaller files, it has sufficient information for reducing the set of potential matches to small subsets. Additional matching runs on the smaller data files can then yield higher proportions of matches. The research question is "What are effective ways of using bridging files?" Bridging files also should have significant power for improving re-identification experiments.

4.3 Data Mining

Machine learning algorithms that employ Bayesian networks are tools being applied to classify text into different groups. Bayesian networks are one of the standard tools in data mining. They are also used for information retrieval methods such as used in some of the web search engines. The latest algorithms (Nigam et al., 1999) utilize EM-based methods that are closely related to methods used by Winkler (1988, 1989, 1993b) and Larsen and Rubin (1999). The EM-based algorithms for finding maximum likelihood estimates in the latent classes models of record linkage are a direct generalization of ideas for automatically estimating parameters given in Fellegi and Sunter (1969). The basic research problems are quite difficult. The first is how to automatically obtain parameters and latent classes that allow automatic accurate determination of error rates. The second is how to effectively use combinations of training data for which true classification is known and general data for which true classification is unknown. Presently, some of the examples in machine learning suggest that appropriate training data – often obtained via very expensive clerical review – can be useful in some situations. Because of the additional structure available in record linkage, some authors (Winkler 1993b, 1994; Larsen and Rubin 1999) have been able to obtain good matching results without subsets of training data. The advantage of training data is that it implicitly imposes additional structure on the learning with general text. With record linkage, the additional structure is due to knowing that fields such as first name, last name, house number, and date-of-birth need to be compared. With general text, the algorithms of machine learning must create a structure for comparing that is facilitated by the training data. The machine learning algorithms are useful in record linkage situations when free-form names or addresses cannot be parsed into components. Winkler (1993b) and Nigam et al. (1999) have shown that each of the latent classes may be best estimated as a further mixture of latent classes. A third research problem emphasized by Nigam et al. (1999) is when the classes obtained under the theoretical latent class models correspond to true classes into which individuals might want to classify the data. Winkler (1989) showed that the parameters of the latent classes sometimes yield very poor matching performance if the latent classes do not correspond to the true classes of matches and nonmatches. Winkler (1993b) showed that dramatic improvements in matching can occur if the class of nonmatches is estimated as a mixture of two subclasses. To better make use of a priori information, Winkler (1988, 1989, 1993b) showed how convex constraints such as P(disagree first | M) < a, 0 < a < 1, or P(M) < b, 0 < b < 1, could be used to force estimates obtained via versions of the EM algorithm into regions of the subspace of parameters.

5. CONCLUDING REMARKS

This paper describes current research problems in record linkage and some related research in microdata confidentiality, information retrieval and data mining.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. A shorter version appeared in the 1999 Statistical Society of Canada Proceedings of Survey Methods. The translation of the abstract to French was facilitated by Dr. Yves Thibaudeau.

REFERENCES

Aarts, E. H. L., and J. K. Lenstra (eds.) (1997), *Local Search in Combinatorial Optimization*, New York, NY: Wiley-Interscience.

Alvey, W. and Jamerson, B. (eds.) (1997), *Record Linkage Techniques -- 1997* (Proceedings of An International Record Linkage Workshop and Exposition, March 20-21, 1997, in Arlington VA), Washington, DC: Federal Committee on Statistical Methodology.

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, **90**, 694-707.

Belin, T. R. (1993) "Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment", *Survey Methodology*, **19**, 13-29.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society, B*, **34**, 192-236.

Besag, J. (1986), "On the Statistical Analysis of Dirty Pictures (with discussion)," *Journal of the Royal Statistical Society, B*, **46**, 25-37.

Besag, J., P. J. Green, D. Higdon, and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science*, **10**, 3-41.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.

Budzinsky, C. D. (1991), "Automated Spelling Correction," Statistics Canada.

Copas, J. R., and F. J. Hilton (1990), "Record Linkage: Statistical Models for Matching Computer Records,"*Journal of the Royal Statistical Society*, **A**, **153**, 287-320.

DeGuire, Y. (1988), "Postal Address Analysis," *Survey Methodology*, **14**, 317-325.

De Waal, A.G. and L.C.R. J. Willenborg (1996), "A View on Statistical Disclosure Control of Microdata," *Survey Methodology*, **22**, 95-103.

De Waal, A.G. and L.C.R. J. Willenborg (1998), "Optimal Local Suppression in Microdata," *Journal of Official Statistics*, **14**, 421-435.

Deming, W. E., and G. J. Gleser (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, 403-415.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **B**, **39**, 1-38.

Fayad U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) (1996), *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: The MIT Press.

Fellegi, I. P. (1999), "Record Linkage and Public Policy: A Dynamic Evolution" in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 3-12.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.

Fienberg, S. E., U. E. Makov, and A. P. Sanil (1997), "A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data," *Journal of Official Statistics*, **13**, 75-89.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998), "Disclosure Limitation and Related Methods for Categorical Data," *Journal of Official Statistics*, **14**, 485-502.

Frakes, W. B., and Baeza-Yates, R. (ed.) (1992), *Information Retrieval: Data Structures & Algorithms*, Upper Saddle River, NJ: Prentice-Hall PTR.

Friedman, J., T. Hastie, R. Tibshirani (1998), "Additive Logistic Regression: a Statistical View of Boosting," Stanford University, Statistics Department Technical Reprort.

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, **9**, 383-406.

Geman, S. and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," IEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721-741.

Gill, L. (1999), "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.

Gopal, R., P. Goes, and R. Garfinkel, "Confidentiality Via Camouflage: The CVC Approach to Database Query Management," in *Statistical Data Protection '98*, Eurostat, Brussels, Belgium, 1-8.

Hall, P. A. V., and Dowling, G. R. (1980), "Approximate String Comparison," *Computing Surveys*, **12**, 381-402.

Heckerman, D. (1996), "A Tutorial on Learning with Bayesian Networks," Microsoft Research, Technical Report MSR-TR-95-06.

Jaro, M. A. (1976), "UNIMATCH," Software system (no longer available).

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey*

Research Methods, 303-308.

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods,* 456-461.

Larsen, M. D. (1996), "Bayesian Approaches to Finite Mixture Models," Ph.D. Thesis, Harvard University.

Larsen, M. D., and D. B. Rubin (1999), "Iterative Automated Record Linkage Using Mixture Models," Statistics Department Technical Report, Harvard University.

Mateo-Sanz, J. M. and J. Domingo-Ferrer (1998), "A method for Data-Oriented Multivariate Microaggregation,"in *Statistical Data Protection '98*, Eurostat, Brussels, Belgium, section 1.

Meng, X., and D. B. Rubin (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm," *Journal of the American Statistical Association*, **86**, 899-909.

Meng, X., and D. B. Rubin (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.

Mitchell, T. M. (1997), *Machine Learning*, New York, NY: McGraw-Hill.

Neter, J., E. S. Maynes, and R. Ramanathan, (1965), "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, **60**, 1005-1027.

Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press (out of print).

Newcombe, H. B., M. E. Fair, and P. Lalonde (1992), "The Use of Names for Linking Personal Records (with discussion), Journal of the American Statistical Association, **87**, 1193-1208.

Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.

Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (1999), "Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, to appear.

Pollock, J. and Zamora, A. (1984), "Automatic Spelling Correction in Scientific and Scholarly Text," *Communications of the ACM*, **27**, 358-368.

Porter, E. H., and W. E. Winkler (1999), "Approximate String Comparison and its Effect in an Advanced Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 190-199.

Rogot, E., Sorlie, P., and Johnson, N. J. (1986), "Probabilistic Methods in Matching Census Samples to the National Death Index, " *Journal Chronological Disease*, **39**, 719-734.

Rubin, D. B. and H. S. Stern (1993), "Testing in latent class models using a posterior predictive check distribution," in *Analysis of latent variables in developmental research*, (eds., Clogg, C. and von Eye, A.).

Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, 39-58.

Scheuren, F., and W. E. Winkler (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, **23**, 157-165.

Sekar, C. C., and W. E. Deming (1959), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association*, **44**, 101-115.

Smith, M. E. and H. B. Newcombe (1975), "Methods for Computer Linkages of Hospital Admission-Separation Records into Cumulative Health Histories," *Meth. Inform. Medicine*, 18-25.

Sweeney, L. (1999), "Computational Disclosure Control for Medical Microdata: The Datafly System" in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 442-453.

Tendick, P. and N. Matloff (1994), "A Modified Random Perturbation Method for Database Security, " *ACM Transactions on Database Systems*, **19**, 47-63.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," in *Proceedings of the Section on Statistical Computing, American Statistical*

Association, 283-288.

Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, 31-38.

Titterington, D. M., A. F. M. Smith, U. E. Makov (1988), *Statistical Analysis of Finite Mixture Distributions*, NewYork: J. Wiley.

Van Dyk, D. A. (1999), "Nesting EM Algorithms for Computational Efficiency, " *Statistica Sinica*, to appear.

Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.

Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, 101-117.

Winkler, W. E. (1989c), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783.

Winkler, W. E. (1990a), "Documentation of record-linkage software," unpublished report, Washington DC: Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (1990b), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Assn.*, 354-359.

Winkler, W. E. (1993a) "Business Name Parsing and Standardization Software," unpublished report, Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

Winkler, W. E. (1993b), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report 94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W. E. and Scheuren, F. (1995), "Linking Data to Create Information," *Proceedings of Symposium 95, From Data to Information - Methods and Systems*, Statistics Canada, 29-37.

Winkler, W. E. and Scheuren, F. (1996), "Recursive Analysis of Linked Data Files," *Proceedings of the 1996 Census Bureau Annual Research Conference*, 920-935.

Winkler, W. E. (1997), "Producing Public-Use Microdata That are Analytically Valid and Confidential," Proceedings *of the Section on Survey Research Methods, American Statistical Association*, 41-50.

Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, **1**, 87-104.

Winkler, W. E. (1999), "Issues with Linking Files and Performing Analyses on the Merged Files," *Proceedings of the Section on Social Statistics, American Statistical Association*, to appear.

Winkler, W. E. and Scheuren, F. (1991), "How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis," U.S. Bureau of the Census, Statistical Research Division Technical Report.

Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical report RR91/09.

Analysis of Identifier Performance using a Deterministic Linkage Algorithm

Shaun J. Grannis MD, J. Marc Overhage MD PhD, Clement J. McDonald MD

Regenstrief Institute for Health Care, Indiana University, Indianapolis IN

ABSTRACT

As part of developing a record linkage algorithm using de-identified patient data, we analyzed the performance of several demographic variables for making linkages between patient registry records from two hospital registries and the Social Security Death Master File. We analyzed samples from each registry totaling 6,000 record-pairs to establish a linkage gold-standard. Using Social Security Number as the exclusive linkage variable resulted in substantial linkage error rates of 4.7% and 9.2%. The best single variable combination for finding links was Social Security Number, phonetically compressed first name, birth month, and gender. This found 87% and 88% of the links without any false links. We achieved sensitivities of 90% to 92% while maintaining 100% specificity using combinations of social security number, gender, name, and birth date fields. This represents an accurate method for linking patient records to death data and is the basis for a more generalized de-identified linkage algorithm.

INTRODUCTION

Because the information needed to answer important health research, management, and policy questions is usually scattered across many independent databases, methods for accurate linkage of patient records from independent sources are critical. Researchers have successfully used a variety of linkage methodologies[1,2].

Automated linkage methodologies are conceptually divided into two broad categories: deterministic and probabilistic.[3] Deterministic algorithms employ a set of rules based on exact agreement/disagreement results between corresponding fields in potential record pairs. Such algorithms are designed to match on a reliable identifier with high discriminating power and then perform verification using additional parameters. For example, linkage may be attempted using Social Security Number (SSN), which is then verified by first and last names.[1] If linkage is unsuccessful, one uses another composite key such as first and last name verified by other identifiers.

Probabilistic algorithms use statistical methods [2,4,5]. Frequency of identifier agreement and disagreement is derived from potential linked and non-linked record-pairs in the data sets. From this information, likelihood scores are calculated for each potential record-pair[5]. The likelihood scores for all potential record-pairs ideally form a bimodal distribution where low scores represent non-links, high scores represent probable links, and intermediate scores represent indeterminate links.

In addition to exact matching, methods exist for establishing agreement between fields such as approximate string comparison[6], phonetic encoding, and nearness metrics[7].

Although probabilistic methods may discriminate better than deterministic methods, in some cases their results require human intervention, and agreement likelihood information may not be readily available for all data.[8] Additionally, deterministic approaches often require less development time and still achieve acceptable results[1,3,4].

While much information can be gained from linked databases, steps must be taken to assure confidentiality of patient records.[9] We are developing a linkage method using data de-identified by a one-way hash function [10,17]. Nearness metrics cannot be used for data de-identified in this way because nearness information is lost in hash functions. Therefore, we must find other mechanisms to reduce variation that might otherwise be accounted for by nearness measures. It is important to avoid mechanisms that require human supervision, because that would break confidentiality in many circumstances, and the cost of supervised matching can be high. Consequently, we have implemented a deterministic, or exact match linkage method.

METHODS

This work was performed as part of the Shared Pathology Information Network (SPIN) project for which we received IRB approval. Using records from two hospital systems' patient registries, our goal was to maximize the chance for an individual to link to the Social Security Death Master File (SSDMF) even after applying a one-way hash function to all identifiers. This problem has general relevance to all medical databases and registries because a match to the SSDMF provides the best indicator of vital status (i.e. whether the patient is living or deceased). Mortality is an important outcome variable for many research questions[11] and we believe the SSDMF is the best source for that data.

The SSDMF is a publicly available database containing demographic data for over 65 million deceased individuals. A one-time snapshot can be purchased for $1,750 and monthly updates are available for $6,900 per year. The database has fields for SSN, name, date of birth, date of death, state or country of residence, ZIP code of last residence, and ZIP code of lump-sum payment. The Social Security Administration (SSA) receives approximately 90% of its death notifications from funeral homes, friends, and relatives of the deceased; postal authorities and financial institutions contribute another 5%. The remaining 5% are derived from computer matches with Federal and State agency data. The file is updated with additions, deletions, and modifications on a weekly basis.[12] The SSA maintains that absence from the database is not proof the patient is alive because some deaths are not recorded. The CDC lists 2,391,043 decedents for 1999 compared to 2,154,018 (90.1%) included in the SSDMF for that year.

For this study we used patient registries from two hospitals in central Indiana. Hospital A is a public inner-city hospital system with a large Medicare/Medicaid population. Hospital B is a private urban hospital system that invested in extensive patient registry clean-up in 1999.

Selecting Indiana Death Records: Patient registries were obtained in December, 2001. We developed an Indiana subset of the SSDMF to speed up the matching process described below. An SSDMF record was included in this subset if any of the fields indicated the patient worked in, lived in, or obtained their SSN from Indiana using following data in the SSDMF: first 3 digits of SSN in the range 303-317; ZIP code for last residence or lump-sum payment ZIP code in the range 46000-47999; or an Indiana state of residence.

Preprocessing: Names and other variables can include variations and errors such that exact string matches may fail when a human reader might recognize them or the equivalent (e.g. "Jim" and "James"). To achieve de-identified matching, we plan to apply a one-way hash function to all fields before attempting linkage, and all information that could help in close matches will be lost. We thought that pre-processing names using a phonetic compression algorithm would help overcome such variations and errors. There are several phonetic compression algorithms; examples include Soundex[13], Metaphone, and the New York State Identification and Intelligence System algorithm (NYSIIS).[14] The NYSIIS algorithm has high discriminating power.[15] NYSIIS codes for first and last names were generated for each data set.

To eliminate last name, first name order reversal errors, we converted names from base 27 (A-Z) to base 10, summed them together, and re-converted to base 27. In this way "JOHN SMITH" and "SMITH JOHN" both produce the sum "SWYAV". We applied this same process to the NYSIIS-transformed first and last names.

Gender was available in the patient registries, but the SSDMF contains no fields for gender. When gender was missing from the hospital registration we imputed it using the non-intersecting names from the top 1000 male and female first names derived from 1990 U.S. Census data. We did the same for all SSDMF records.

Birth date and SSN are also subject to errors, but there is nothing analogous to Soundex-like rules for these variables. To accommodate errors in birth date, we decomposed it into month, day, and year variables; we used various combinations to attempt linkage. When SSN was erroneous we used other linkage criteria such as full name, birth date, and gender.

We preprocessed the data from each of the candidate match fields shown in Table 1. Because identifiers such as race, mother's maiden name, and institutional identifiers that were present in the hospital records were not present in the SSDMF, they were not included in matching rules. We used only the preprocessed variables in our analysis. In the context of anonymous linkage, we could perform this preprocessing at each source system before applying a one-way hash without compromising confidentiality. However, we examined the performance of both the raw and NYSIIS names. The preprocessing was intended to increase the chance of a correct match.

Manual Analysis: We developed a gold standard for measuring the error rates of the linkage variables and for comparing the matching accuracy of various combinations of these variables as follows. Using SSN as the single identifier, we linked the patient registries to the Indiana subset of the SSDMF resulting in potentially linked record pairs. If a hospital record linked to more than one record in the SSDMF, the first record pair was used. As the first stage, we obtained a random sample of n=1000 record-pairs from each institutions' potential links. The two samples were then manually reviewed and record pairs were labeled as correct or incorrect links.

Retrospective analysis of both 1000 patient samples revealed that all incorrect links based on SSN alone mismatched either on first names or birth years. In hospital A, the 84/1000 manually-labeled incorrect links were found among record pairs mismatched either on first name or birth year. Similarly, in hospital B, the 39/1000 incorrect links were found among record pairs meeting the same mismatch criteria.

To create a larger set of test cases, we took a random sample of 5000 record pairs linked by SSN alone from each hospital and manually reviewed all cases that mismatched on first name or birth date. Of the 5000 record pairs in each sample, 1,367 record-pairs from hospital A (27.3%) and 825 record pairs from hospital B (16.5%) were manually reviewed and labeled as correct or incorrect links. The n=1000 and n=5000 samples from each hospital were then combined to form gold standards of n=6000 record-pairs. We determined sensitivities and specificities for multiple combinations of candidate

Table 1: Preprocessed identifiers

Identifier	Values	Preprocessing Rules
Social Security Number (SSN)	0-9	Remove non-numeric characters; nullify if not 9 digits; nullify if not valid
Last Name (LN)	A-Z	Remove non-alphabetic characters, suffix and prefix nullify invalid names.
First Name (FN)	A-Z	Remove non-alphabetic characters, suffix and prefix nullify invalid names.
Name Sum (NS)	A-Z, zero	Produced after pre-processing of Last and First Names.
Gender (G)	M, F	If null, or ≠ (M F), attempt imputation from first name list based on census list.
NYSIIS encoding of Last Name (LNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
NYSIIS encoding of First Name (FNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
Sum of NYSIIS Names (SNY)	A-Z, zero	Sum of LNY and FNY
Month of birth (MB)	0-9	Convert from alphabetic month, 0 if < 0 or > 13
Day of Birth (DB)	0-9	0 if (< 0 or > 31)
Year of Birth (YB)	0-9	0 if (< 1800 or > 2001)

linkage variables within these gold-standard record pairs.

Non-SSN Linkage: For SSN record pairs labeled as incorrect links, we attempted a second linkage to the Indiana SSDMF using first name, last name, gender, and birth date. These were manually reviewed and labeled as correct or incorrect links. The correct links not generated by SSN were then compared to the initial incorrect SSN-generated links.

RESULTS

A substantial number of patient registration records, approximately 35%, lacked SSNs at each institution. Only the hospital records with valid SSNs were used in this study. When we linked these hospital records to the Indiana subset of the SSDMF, 57,446 (8.4%) of hospital A's records linked to a record in the Indiana SSDMF, and 147,878 (10%) records from hospital B linked.

We used the patient registry records that linked by SSN alone to the SSDMF to obtain the gold standard·data set of 6000 record pairs. Among the 6000 gold standard record pairs, using SSN as the exclusive match variable, hospital A had 550 incorrect links, indicating a 9.2% SSN error rate, and hospital B had 281 incorrect links, indicating a 4.7% SSN error rate.

Table 2 shows the individual identifier mismatch rates among correct links based on SSN alone. Assuming that the SSDMF carries the correct information, these data provide an estimate of the error rates in the recorded information for each of the listed patient identifier fields.. However, we cannot consider mismatches on first and last names to be strict errors because interchange between first names, nicknames and varying uses of first and middle initials confound this comparison. Further, the gender figures are not precise because all of the gender values in the SSDMF file are imputed.

Table 2: Identifier Error Rates Among Correct SSN-based Links

	Error Rates (%)	
	Hospital A (n=5450)	Hospital B (n=5719)
Last Name	5.9	2.1
First Name	12.5	8.2
Name Sum	16.7	9.9
NYSIIS Last Name	3.9	1.5
NYSIIS First Name	9.5	7.2
NYSIIS Sum	12.3	8.3
Gender	0.6	0.6
Month of Birth	3.7	1.8
Day of Birth	8.4	5.3
Year of Birth	8.2	4.2

There are some interesting observations we can make from this Table. Error rates were higher at hospital A as compared to hospital B, which had invested in a major clean up of their registration systems 3 years ago. It is notable that the month of birth is more accurate than year or day of birth. Also as expected, the NYSIIS algorithm had a lower mismatch rate than raw names. However the mismatch rate with NYSIIS was not zero, reminding us that phonetic transforms do not equivalence minor name differences like "Bill" and "Gill".

Among the record pairs not linkable by SSN, the use of name and birth date criteria identified an additional 196 correct links between hospital A and the Indiana SSDMF, while the same process identified another 109 correct links in hospital B. Using these links we analyzed the original SSN-linked record pairs for errors.

SSN errors consisted of three types shown in Figure 1. The most common error appeared to be due to spousal mix-ups (56% hospital A, 39% hospital B) in that a female of one record was linked to a male record sharing the same last name. Typographical errors (41% hospital A, 30% hospital B) and SSN collisions of unknown etiology (3% hospital A, 31% hospital B) accounted for the remainder of the errors. Figure 1 shows examples using fictitious data.

Figure 1: SSN Error Examples

	SSN	LN	FN	G	MB	DB	YB
	Typographical Errors						
Institution Data	123446789	SMITH	FRED	M	12	20	1908
SSDMF Correct Name Link	123456789	SMITH	FRED	M	12	20	1908
SSDMF Incorrect SSN Link	123446789	JONES	PAT		7	13	1914
	Spousal Linkage						
Institution Data	987654321	COLLINS	MARY	F	8	29	1917
SSDMF Correct Name Link	987708245	COLLINS	MARY	F	8	29	1917
SSDMF Incorrect SSN Link	987654321	COLLINS	TIM	M	4	3	1915
	Unexplained Collisions						
Institution Data	586192376	PRATT	DAVID	M	10	3	1932
SSDMF Correct Name Link	666245764	PRATT	DAVID	M	10	3	1932
SSDMF Incorrect SSN Link	586192376	WILSON	ERIN		5	17	1940

The rows in Tables 3 and 4 describe sets of identifiers that could be used for linking patients and their corresponding false positive and false negative link rates. The best single combination of identifiers for finding matches was SSN, first name transformed by the NYSIIS, month of birth, and gender. This combination found 87% to 88% of the possible links without finding any false links. Taking the union of more than one set of keys – that is link by one set of keys, then link by another set of keys, and include all of the links from any of these steps in the final result – yielded an 89% to 90% link rate without picking up any false links. Adding links on first name, last name, and full birth date increased these yields to 90-92%.

DISCUSSION

Hospital registries contain substantial numbers of errors in SSNs that prohibit the use of SSN as a single linkage key. Additional fields have to be added to avoid incorrect links. Similar error rates in the SSN have been reported previously.[16] Nearly half of the SSN errors are due to spousal mix-ups, almost certainly due to a mix up between the guarantor's SSN and that of the patient, or beneficiary. Additional linkage identifiers such as gender and first name help to avoid incorrect links between beneficiaries and guarantors. We recommend that health care systems develop registration procedures to avoid the incorrect assignment of guarantor's SSN to a beneficiary.

Linkage criteria that include SSN combined with variables from both name and birth date maximize the match rate while keeping the false positive rate near zero. Identifier variations are not independent; people with the same last names may end up using the same SSN because of beneficiary or other errors. The first name and

Table 3: Results of 6,000 random samples taken from 57,446 record-pairs linked by SSN between Hospital A and SSDMF Indiana

	Links		Non-links			
Linked Identifiers	Correct	Incorrect	Correct	Incorrect	Sensitivity (%)	Specificity (%)
SSN Alone	5450	550	0	0	100	--
Name Criteria:						
SSN, LN, FN	4541	7	543	916	83.2	98.7
SSN, LNY, FNY	4775	7	543	675	87.6	98.7
SSN, SNY	4782	7	543	668	87.7	98.7
Date Criteria:						
SSN, MB, DB, YB	4557	2	548	893	83.6	99.6
Name/Date Criteria with SSN:						
SSN, FN, YB, G	4350	0	550	1100	79.8	100
SSN, FNY, YB, G	4496	0	550	954	82.5	100
SSN, FNY, MB, G	4724	0	550	726	86.7	100
Name /Date Criteria without SSN:						
LN, FN, MB, DB, YB, G	3996*	0	550	1650	70.1	100
Union of (FNY,YB,G), (FNY, MB, G), and (LN,FN,MB,DB,YB)	5053	0	550	593	89.5	100

* Potential links for non-SSN matches = 6196

Table 4: Results of 6,000 random samples taken from 147,848 record-pairs linked by SSN between Hospital B and SSDMF Indiana

	Links		Non-Links			
Linked Identifiers	Correct	Incorrect	Correct	Incorrect	Sensitivity (%)	Specificity (%)
SSN Alone	5719	281	0	0	100.0	--
Name Criteria:						
SSN, LN, FN	5157	2	279	562	90.2	99.3
SSN, LNY, FNY	5247	2	279	474	91.7	99.3
SSN, SNY	5245	2	279	474	91.7	98.9
Date Criteria:						
SSN, MB, DB, YB	5216	2	279	503	91.2	99.3
Name and Date Criteria:						
SSN, FN, YB, G	4997	0	281	722	87.4	100
SSN, FNY, YB, G	5048	0	281	671	88.3	100
SSN, FNY, MB, G	5181	0	281	538	90.6	100
Name and Date Criteria without SSN:						
LN, FN, MB, DB, YB, G	4776*	0	281	1052	81.9	100
Union of (FNY,YB,G), (FNY, MB, G), and (LN,FN,MB,DB,YB)	5331	0	281	497	91.5	100

* Potential links for non-SSN matches = 6109

gender provide important protections against such errors. Gender is included to avoid the theoretical possibility of an incorrect NYSIIS linkage between family members with similar first names who share SSN and birth date parameters.

The preprocessed linkage variables that perform reasonably well in this study are suitable for a de-identified linking mechanism. After being preprocessed at the local information system, identifiers can be encrypted via a secure one-way hash, using a one-time seed shared by all sites. The hashed keys can be sent to a trusted third party for linking and that party can assign random codes to each patient.[17]

We restricted the matching to the Indiana subset of the SSDMF to limit file size and computer time. To find all possible deaths in a local population of patients, one would link to the entire SSDMF. We would expect to find more links between patients in the registration files but also to encounter higher error rates, because the larger number of individuals in the target file would provide greater chances for links between different individuals who happen to have the same identifiers.

These results are based on modest sample sizes, and further analysis of larger populations is warranted. Our methods apply to decedent matches and patients from the Midwest. This may not generalize to other populations with high percentages of Hispanic or Asian names. By its nature, the death index contains an older population; linkage performance in a younger, more diverse population may differ. Further, assuming that the SSDMF file contains much cleaner data than the average hospital registration file, we would expect a lower link rate and more errors when data from both files are derived from patient registries.

This is an accurate method of linking patient records to death data, and will be the basis for a more generalized de-identified linkage algorithm. Future work includes linking registry data to the entire SSDMF to study the error properties and match rates using a larger data set. Work will also be directed toward improving non-SSN name matches. We will also consider use of some statistical properties such as name and birth date frequencies to improve matching precision.

REFERENCES

1. Potosky A, Riley G, Lubitz J, et al. Potential for Cancer Related Health Services Research Using a

Linked Medicare-Tumor Registry Database. Medical Care 1993;31(8):732-748.

2. Whalen D, Pepitone A, Graver L, Busch JD. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.
3. Liu S, Wen SW. Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission. Chronic Diseases in Canada 1999; 20(2):77-81.
4. Gill, L., Methods for Automatic Record Matching and Linking and their use in National Statistics. Her Majesty's Stationary Office, Norwich, 2001.
5. Fellegi, I.P., & Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183-1210.
6. Porter E, Winkler W. Approximate String Comparison and its Effect on an Advanced Record Linkage System. Record Linkage Techniques--1997: Proceedings of an International Workshop and Exposition. National Academy Press, Washington DC 1999.
7. Sideli R, Friedman C. Validating Patient Names in an Integrated Clinical Information System. Symposium on Computer Applications in Medical Care, Washington, DC. November 1991:588-592.
8. Van Den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunan PMH. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. Int J Epidemiol 1990; 19:553-8.
9. Department of Health and Human Services, Office of the Secretary. The Health Insurance Portability and Accountability Act of 1996, Standards for Privacy of Individually Identifiable Health Information; Final Rule. Federal Register 65 FR 82462; December 28, 2000. Available at: http://www.hcfa.gov/hipaa/hipaahm.htm
10. Burrows, JH. Secure Hash Standard. Federal Information Processing Standards, Publication FIPS PUB 180-1 <http://www.itl.nist.gov/fipspubs/fip180-1.htm> website accessed 3/1/2002.
11. Pates R, Scully W, et al. Adding Value to Clinical Data by Linkage to a Public Death Registry. MedInfo 2001;10(Pt 2):1384-8.
12. Social Security Administration, Office of the Inspector General, Unresolved Death Alerts Over 120 Days Old (A-09-00-10001). Audit Report; 2001 August.
13. Knuth DE. The Art of Computer Programming, Volume 3/Sorting and Searching, Second Edition. Addison-Wesley Publishing Company, 1998.
14. Lynch BT, Arends WL. Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division, 1977.
15. Newcombe HE. Handbook of Record Linkage, Methods for Health and Statistical Studies, Administration, and Business. Oxford University Press, 1988.
16. Newman T, Brown A. Use of Commercial Record Linkage Software and Vital Statistics to Identify Patient Deaths. J Am Med Inform Assoc. 1997 May-June; 4 (3): 233-237.
17. Schadow G, McDonald CJ Maintaining Patient Privacy in a Large Scale Multi-Institutional Clinical Case Research Network. AMIA Proceedings (2002 Submission).

ACKNOWLEDGEMENTS

This work was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was supported in part by grants from the National Library of Medicine (T15 LM-7117-05), the National Cancer Institute (1 U01 CA91343-01), and The Indiana Genomics Initiative (INGEN) of Indiana University, which is supported in part by Lilly Endowment Inc.

Project Charter:
Statewide Voter Registration System

Prepared for:

Wisconsin State Elections Board

May 15, 2003

Prepared by

Virchow Krause
& company

Ten Terrace Court
Madison, WI 53707
608-249-6622
www.virchowkrause.com

Virchow Krause
& company

May 15, 2003

Mr. Kevin J. Kennedy, Executive Director
Wisconsin State Elections Board
132 East Wilson Street, Suite 200
Madison, WI 53701-2973

Dear Kevin,

Re: Statewide Voter Registration System Project Charter

Enclosed is the final report of Virchow Krause & Co., LLP for the study of the statewide voter registration system (SVRS) for the Wisconsin State Elections Board. The SVRS is a significant initiative for the entire state of Wisconsin. Successful deployment of a new system will require involvement and investment at the state, county, and local levels of government. It will require a complex and lengthy implementation.

For such an initiative, it is imperative that consensus be reached on the SVRS Project Charter before the work begins. To that end we submit our report, divided into three sections plus appendices:

1. Executive Summary
2. SVRS Findings
3. SVRS Project Charter
4. Appendices

The Executive Summary provides a high level overview of the SVRS initiative, describing the current situation in Wisconsin and the implementation steps the State of Wisconsin must take to achieve compliance with the SVRS provisions of the federal Help America Vote Act (HAVA). Estimated costs of the SVRS and cost reduction strategies are also outlined.

The SVRS Findings section provides a review of the SVRS study project, and presents important information discovered during the analysis. The findings are the result of extensive discussions with county and local officials, state agencies affected by the SVRS, the Department of Electronic Government (DEG), other states, and three system vendors with previous experience in statewide voter lists. The findings are logically grouped into major themes, each of which will have a significant impact on the SVRS implementation and which therefore also shape the Project Charter recommendations.

The Project Charter is the definition of the SVRS implementation initiative. The objectives, scope, assumptions and known risks of the SVRS initiative are documented, along with proposed draft statutory changes. Flowcharts show the business processes and technical architecture of the new system across the local, county, and state levels. Phased implementation plans provide the approach, high level work steps, resources, and organization required to finalize the operational model at the county and municipal levels, select the system vendor, and implement the system. Detailed 5 year total cost of ownership schedules show a year-by-year cost estimate broken down by cost element, based on vendor RFI responses and other agency cost estimates. Combined, these components provide a high level design of the SVRS system which should be used as the basis for final policy, technical, operational, and funding decisions.

Thank you for the opportunity to work with you and your team. While it has been a significant amount of work in a very short period of time, it has been a pleasure and a privilege to work on this important project with the State Elections Board, the Department of Electronic Government, the Department of Transportation's Division of Motor Vehicles, the Department of Corrections, the Department of Health and Family Services, and the county and municipal officials. We appreciate the effort and cooperation of all involved.

Sincerely,
Virchow Krause & Co., LLP

Keith Downey, Partner

Table of Contents

Executive Summary

A. Background

In October 2002, the federal government passed the Help America Vote Act of 2002 (HAVA). This legislation created new election administration requirements for all states and called for an upgrade of voting systems to better accommodate persons with disabilities. Specifically, HAVA calls for the creation of a single, uniform, official, centralized, interactive computerized statewide voter registration list defined, maintained, and administered at the state level that contains the name and registration information of every legally registered voter in the state. The current timeline for HAVA calls for election officials to meet the majority of the HAVA requirements by January 1, 2004, and the remainder by January 1, 2006. Extensions of the initial deadline (to January 1, 2006) are permissible and Wisconsin plans to submit an extension request and expects it will be accepted. .

In December 2002, the Legislature provided funds for the State Elections Board (SEB) to study and prepare specific recommendations for implementing a statewide voter registration database system (SVRS), including a proposal for the system's cost and proposed legislation required to initially implement such a system. The Elections Board retained Virchow Krause & Co. LLP to assist with the study, analyze the central and local system requirements, and develop and issue a request for information (RFI) to potential vendors of statewide voter systems and other interested vendors. This report and the enclosed Project Charter represent the results of that study and the RFI.

B. Current Situation

The State of Wisconsin does not currently have a formalized statewide voter registration system or process. Consider the following:

- Under the present statutes, only municipalities that have a population over 5,000 are required to register electors.
- There are some individual municipalities that have voter registration systems to comply with statutory requirements.
- Some counties maintain voter registration data for municipalities and some municipalities electronically maintain elector lists but do not register voters.
- Most municipalities have no record (manual or electronic) of its electors. Consider the following data, collected from the November 2002 election:

Number of municipalities **without** voter registration	1,530
Number of voters in the November 2002 election	563,272
Number of municipalities **with** voter registration	320
Number of registered electors	2,625,353
Number of voters in the November 2002 election	1,363,789
Estimated size of statewide voting age population	4,100,000

Furthermore, there are over fifty different software solutions (e.g., Workhorse Software, Town Hall Software, etc.) and fifty custom applications being employed by those 320 municipalities, including custom applications in the state's largest municipalities—Milwaukee and Madison. Associated with those varied solutions are a myriad of policies and procedures.

In summary, the current activities and processes supporting voter registration are:

- Currently managed at the local levels,
- Largely decentralized and non-standard, and
- Effective in maintaining the integrity of the local electoral process.

To comply with HAVA regulations, the state will require new SVRS applications, processes and procedures for the centralized voter list.

C. SVRS Future-State

Based on information from other states and from vendors who have implemented statewide voter systems, it is clear that a statewide voter registration system is significantly different from a municipal system both in kind and in degree.

A statewide voter system is not simply a municipal system with more records. A statewide system has different integration processes (between municipalities and with other state agencies); it has different security issues; different validation processes; different purge processes, and different scalability requirements. The system has to accommodate various levels of technological sophistication and volumes of transactions ranging from the City of Milwaukee, Milwaukee County (with up to 100,000 election day registrations) to the Town of Butler, Clark County with its 70 voting age electors.

The complexities of implementing a statewide system involving municipalities, counties, and multiple state agencies require a very large scale project effort. Statutory, policy, funding, process, organizational, and technical issues must be carefully addressed in order for this project to succeed. This is a unique challenge for the state in a critical area where the right of citizens to vote is affected.

Furthermore, there is a body of expertise in statewide voter registration found in the SVRS package vendors. No vendor with statewide voter registration experience proposed the development of a custom application. There is a significant opportunity to leverage existing HAVA-specific software functionality, expertise, and lessons learned. The State Elections Board SVRS initiative is much more than a software development and implementation project, and the overall initiative may benefit greatly from leveraging the knowledge and experiences of SVRS package vendors to ensure success.

The future statewide voter registration system will need to take into consideration the following factors (see the Findings section for more detail on these elements):

- Statewide voter database—one centralized, unified list at the core of the electoral process.
- Municipal information—the SVRS is more than just voters; it requires the maintenance of address, municipal, and voting jurisdiction information.
- State statutes—revised in "cosmetic" ways, to modify language pertaining to municipal registration and revised substantially to address new issues created by the existence of one statewide elector database.
- Policies and procedures for the 72 counties and 1,850 municipalities—to insure the integrity of the database, the number of users must be controlled and the policies and procedures that gather the data must be standardized.
- Integration with direct impact agencies—how often to integrate, what data to extract, how to match and what to do with the other agencies' data.
- Cost—the cost to the state and municipalities will vary significantly depending on the number of users (i.e., degree of consolidation), the magnitude of conversion, and whether the state or the vendor will host the application and provide on-going maintenance and support.
- Statewide implementation and roll-out—the nature, timing and duration of activities to bring the system live.

- Initial and on-going training—a burst of activity initially, but on-going training as well, because of the natural turnover in the office of municipal and county clerks.
- Large scale technical architecture—the size and complexity of the system result in very robust software, hardware, and connectivity requirements.
- On-going operation and maintenance of both central and distributed system components—another cost of doing business that must be factored into state and local budgets.

The proposed future-state process map for Wisconsin's statewide voter registration is depicted at a high level in Figure 1, below.

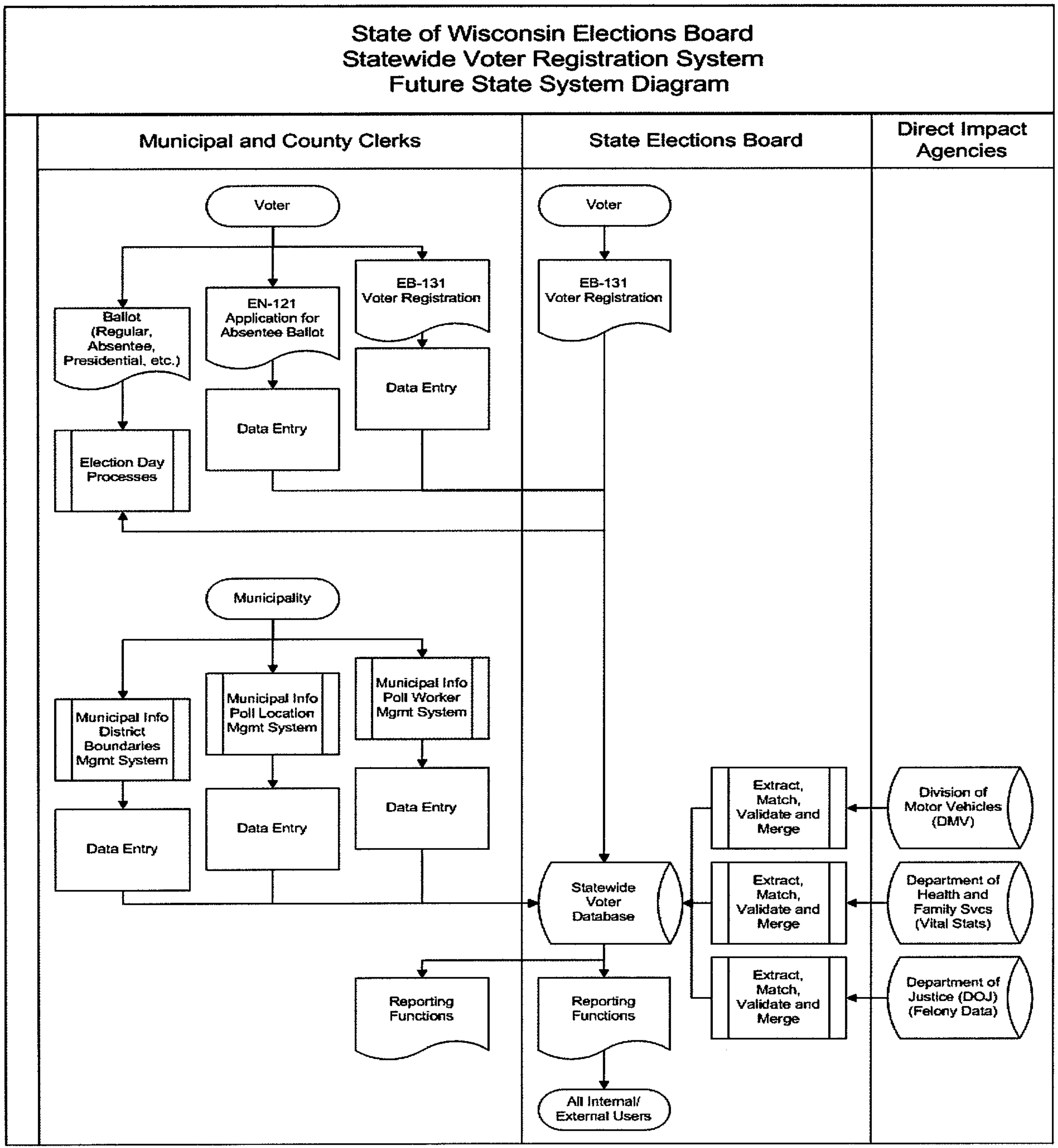

State of Wisconsin Elections Board
Statewide Voter Registration System
Future State System Diagram
Municipal and County Clerks
State Elections Board
Direct Impact Agencies
Voter
Voter
Ballot (Regular, Absentee, Presidential, etc.)
EN-121 Application for Absentee Ballot
EB-131 Voter Registration
EB-131 Voter Registration
Data Entry
Data Entry
Election Day Processes
Municipality
Municipal Info District Boundaries Mgmt System
Municipal Info Poll Location Mgmt System
Municipal Info Poll Worker Mgmt System
Data Entry
Data Entry
Data Entry
Statewide Voter Database
Extract, Match, Validate and Merge
Extract, Match, Validate and Merge
Extract, Match, Validate and Merge
Division of Motor Vehicles (DMV)
Department of Health and Family Svcs (Vital Stats)
Department of Justice (DOJ) (Felony Data)
Reporting Functions
Reporting Functions
All Internal/ External Users

Figure 1. High-Level Statewide Voter Registration Process Map

However, there is a future scenario where a direct connection between the SVRS and the voting system is possible. Connecting ballot creation and the recording and tallying of votes would allow for "anywhere-voting." That is, the system could be connected so that a voter could use voting equipment at a Wisconsin polling place that is electronically connected to the voter database and the ballot creation system. Then, the voter could sign in on the voting equipment which would then pull up the appropriate ballot. Thus, a voter could vote at any location that was connected to the SVRS. Furthermore, this scenario is not far behind the concept of internet voting; that is, the voter signing into the voting system via the internet and then receiving and casting their ballot. Most likely, the technology for this future state will exist long before statutes enable it.

As the state Elections Board pursues the selection and implementation of its SVRS, it should work to ensure that the solution does not preclude it from the flexibility of considering anywhere-voting and Internet voting.

G. Integration with Direct Impact Agencies

HAVA calls for agreements between the SEB and the DMV. It calls on the SEB to also use data on deaths and felony/civil rights status from other direct impact agencies (e.g., DHFS and DOJ). The agreements must specify the following:

- The specific elements of data requested (e.g., name, address, driver's license number),
- The format of the data,
- The frequency of data requests, and
- The cost for data.

In addition, the following issues must be addressed:

- Programs must be written to match the extracted data to the statewide voter database.
- Policies and procedures would be developed for dealing with
  - Records that match 100%,
  - Records that partially match, and
  - Records that do not match at all.

Name matching and validation issues are very complex (e.g., matching Margie L. Smith with Margaret Smith), and are made even more complex when aliases and name changes are considered. The timing and error correction routines of the interfaces to other state agencies is extremely important. Even a 1% error rate on an interface validating names, driver license numbers, etc. could generate tens of thousands of bad matches in an error log, well beyond any ability for the users to manually verify the errors. Again, a high degree of accuracy is imperative prior to the modification of voter records.

One vendor proposed (and has implemented) an option where records that match 100% be "pushed" automatically into the statewide voter registration database. Two others suggested, based on their experiences, that records that match 100% be distributed to appropriate municipalities for their approval prior to updating the statewide voter database. This second scenario appears to be more aligned with Wisconsin's philosophy related to electors and voters. All vendors suggested that incomplete or unmatched records be ignored, because the time to resolve, cost to resolve, and potential for error and disenfranchisement was too high.

H. Package vs. Custom SVRS Solution

The RFI responses led the study team to focus on vendors who have knowledge and experience with statewide voter registration systems. The objective was to leverage that knowledge and expertise in order to be in a position to create a viable procurement process, including preparing the state for the financial impact of this project. If the findings of this study included a conclusion that the state's requirements were very unique, then it would be more likely that a custom solution could be a viable option.

In order to prepare an RFP to which custom developers could provide a credible (i.e., fiscally responsible) reply, the state would need to expend a significant amount of up-front investment (see Project Charter, section G). Because, in addition to specifying business requirements (as this study did through the RFI), a "custom development RFP" would need to identify and develop detailed design and functional specifications required by the application. The RFP would need to provide screen layouts, report designs, and many other system elements.

The study found that the requirements of Wisconsin's SVRS are not very unique. That is, vendors with existing SVRS solutions bring knowledge, expertise, and additional functionality. Thus, selection of a custom application does not appear to be warranted.