

06-CV-00726-STMT

# Exhibit M

## 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 (NYSIS)[14]. The NYSSIS algorithm has high discriminating power.[15] NYSSIS 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 NYSSIS-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 NYSSIS 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.
NYSSIS encoding of Last Name (LNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
NYSSIS encoding of First Name (FNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
Sum of NYSSIS Names (SNV)	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=5718)
Last Name	3.9	2.1
First Name	12.5	8.2
Phone Num	16.7	9.9
NYSIS Last Name	3.9	1.5
NYSIS First Name	9.5	7.2
NYSIS Sex	12.3	8.3
Gender	0.6	0.6
Month of Birth	3.7	1.4
Day of Birth	3.4	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 NYSIS algorithm had a lower mismatch rate than raw names. However the mismatch rate with NYSIS was not zero, reminding us that phonetic transforms do not equivalence minor name differences like "Bill" and "Gil".

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**

Registration Name	SSN		LN	FN	G	DOB	YB
	H	I					
Hospital A	12345678	87654321	SMITH	FRED	M	12 30 1948	1948
SSDMF Correct Match Link	12345678	87654321	SMITH	FRED	M	12 30 1948	1948
SSDMF Incorrect SSN Link	12345678	98765432	JONES	PAT	F	7 15 1914	1914
<b>Spousal Linkage</b>							
Hospital A	(12345678)	(87654321)	COLLINS	MARY	F	8 20 1947	1947
SSDMF Correct Match Link	(12345678)	(87654321)	COLLINS	MARY	F	8 20 1947	1947
SSDMF Incorrect SSN Link	(12345678)	(98765432)	COLLINS	TOM	M	4 3 1919	1919
<b>Unexplained Collisions</b>							
Hospital A	(12345678)	(87654321)	PLATT	DAVID	M	10 3 1952	1952
SSDMF Correct Match Link	(12345678)	(87654321)	PLATT	DAVID	M	10 3 1952	1952
SSDMF Incorrect SSN Link	(12345678)	(98765432)	PLATT	ERD	F	2 17 1940	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 NYSIS, 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**

Linked Identifiers	Links		Non-Links		Sensitivity (%)	Specificity (%)
	Correct	Incorrect	Correct	Incorrect		
SSN Alone	5450	550	0	0	100	-
<b>Name Criteria:</b>						
SSN, LN, FN	4541	7	543	916	83.2	98.7
SSN, LNY, FNY	4775	7	543	673	87.5	98.7
SSN, SNY	4782	7	543	668	87.7	98.7
<b>Date Criteria:</b>						
SSN, MB, DB, YB	4557	2	548	893	83.6	99.6
<b>Name/Date Criteria with SSN:</b>						
SSN, FN, YB, G	4350	0	550	1100	79.8	100
SSN, FNY, YB, G	4496	0	550	954	82.3	100
SSN, FNY, MB, G	4724	0	550	726	86.7	100
<b>Name/Date Criteria without SSN:</b>						
LN, FN, MB, DB, YB, G	3996*	0	550	1650	70.1	100
<b>Union of (FNY, YB, G), (FNY, MB, G), and (LN, FN, MB, DB, YB)</b>	5053	0	550	593	89.3	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**

Linked Identifiers	Links		Non-Links		Sensitivity (%)	Specificity (%)
	Correct	Incorrect	Correct	Incorrect		
SSN Alone	5719	281	0	0	100.0	-
<b>Name Criteria:</b>						
SSN, LN, FN	5157	2	279	562	90.2	99.3
SSN, LNY, FNY	5247	2	279	474	91.7	99.3
SSN, SNY	5265	2	279	474	91.7	98.9
<b>Date Criteria:</b>						
SSN, MB, DB, YB	5216	2	279	503	91.2	99.3
<b>Name and Date Criteria:</b>						
SSN, FN, YB, G	4997	0	281	727	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
<b>Name and Date Criteria without SSN:</b>						
LN, FN, MB, DB, YB, G	4776*	0	281	1052	81.9	100
<b>Union of (FNY, YB, G), (FNY, MB, G), and (LN, FN, MB, DB, YB)</b>	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 NYSIS 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 linkage 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

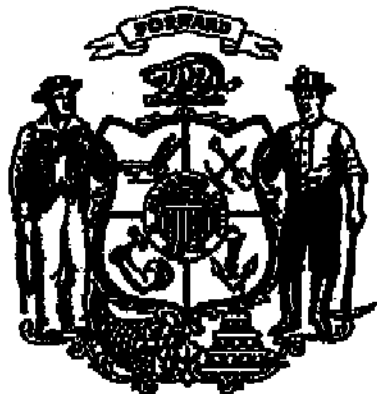
1. Potosky A, Riley G, Lubitz J, et al. Potential for Cancer Related Health Services Research Using a

- Linked Medicare-Tumor Registry Database. *Medical Care* 1993;31(8):732-748.
2. Whalen D, Pepitons 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, Goldboom 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.hhs.gov/hipaa/hipaahtml.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(P12):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 HB. *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 (15 LM-7117-05), the National Cancer Institute (1 U01 CA91343-01), and The Indiana Genomics Initiative (NGEN) of Indiana University, which is supported in part by Lilly Endowment Inc.

# **Exhibit N**



**Project Charter:  
Statewide Voter Registration System**

**Prepared for:  
Wisconsin State Elections Board**

**May 15, 2003**

**Prepared by**



**VirchowKrause  
& company**

Ten Terrace Court  
Madison, WI 53707  
608-249-6622  
[www.virchowkrause.com](http://www.virchowkrause.com)



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..... 5**

A. BACKGROUND ..... 5

B. CURRENT SITUATION..... 5

C. SVRS FUTURE-STATE ..... 8

D. NEXT STEPS ..... 8

E. TOTAL COST OF OWNERSHIP ..... 9

F. COST REDUCTION STRATEGIES ..... 10

**SVRS FINDINGS ..... 12**

A. PROJECT BACKGROUND ..... 12

B. STATEWIDE VOTER DATABASE..... 13

C. VOTER REGISTRATION STATISTICS – COMPARATIVE COMPLEXITY ..... 14

D. FEDERAL AND STATE STATUTES ..... 15

E. POLICIES AND PROCEDURES..... 17

F. INTEGRATION WITH OTHER SEB, COUNTY AND MUNICIPAL INFORMATION SYSTEMS ..... 18

G. INTEGRATION WITH DIRECT IMPACT AGENCIES..... 20

H. PACKAGE VS. CUSTOM SVRS SOLUTION ..... 20

**SVRS PROJECT CHARTER ..... 22**

A. GOAL..... 22

B. OBJECTIVES..... 22

C. SCOPE..... 22

D. ASSUMPTIONS..... 24

E. RISKS AND MITIGATION STRATEGY ..... 26

F. FUTURE SYSTEM PROCESSES ..... 29

G. STATEWIDE VOTER REGISTRATION SYSTEM (SVRS) PLAN ..... 30

H. STATEWIDE PROJECT ORGANIZATION STRUCTURE ..... 48

I. RESOURCE AND STAFFING ESTIMATES ..... 47

J. FIVE YEAR TOTAL COST OF OWNERSHIP ..... 49

**APPENDICES**

## **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.

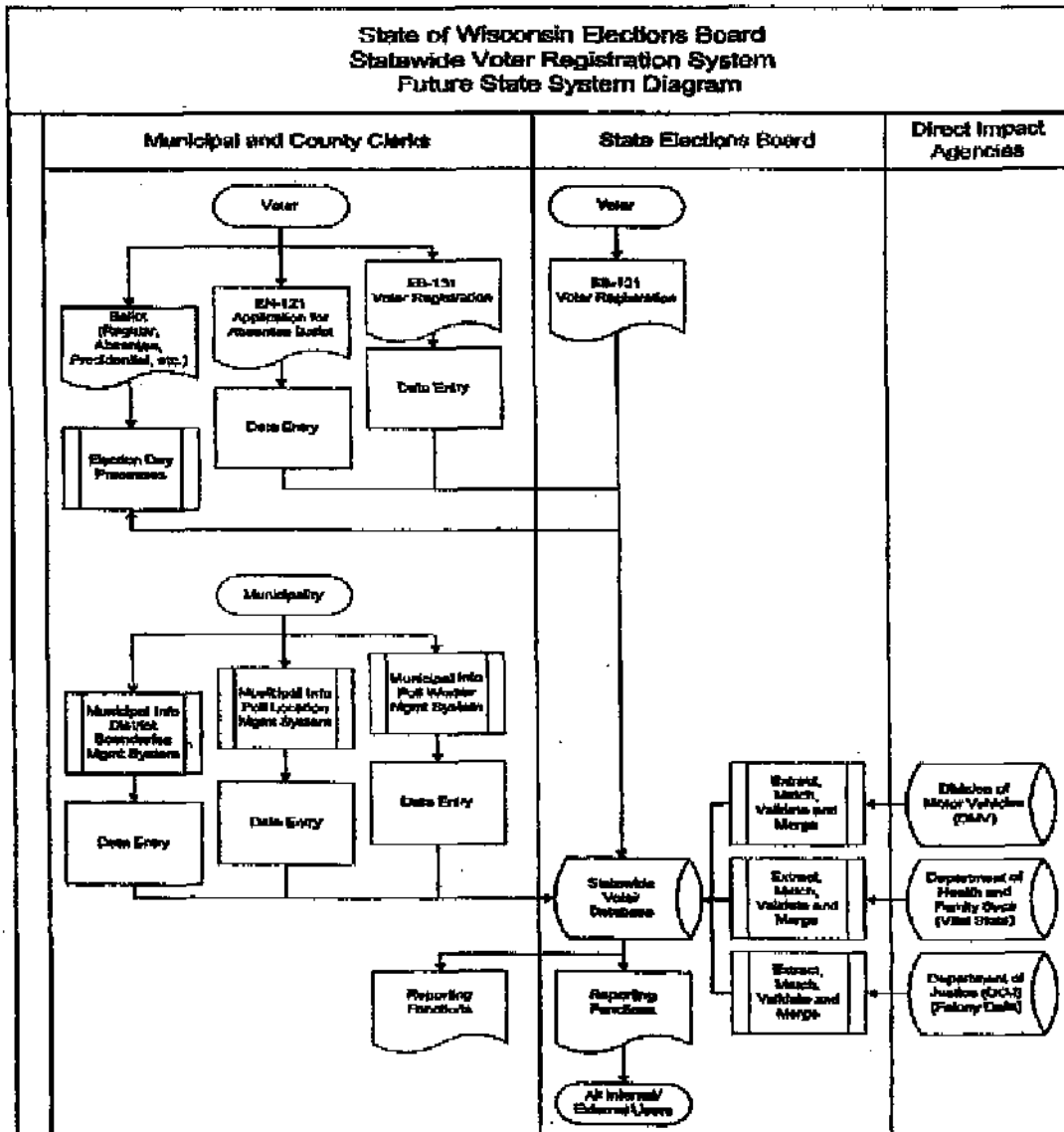


Figure 1. High-Level Statewide Voter Registration Process Map

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., DMFS 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.