© Health Research and Educational Trust

DOI: 10.1111/1475-6773.12323

ENHANCED HOSPITAL DISCHARGE DATA

Incorporating the Last Four Digits of Social Security Numbers Substantially Improves Linking Patient Data from De-identified Hospital Claims Databases

James M. Naessens, Sue L. Visscher, Stephanie M. Peterson, Kristi M. Swanson, Matthew G. Johnson, Parvez A. Rahman, Joe Schindler, Mark Sonneborn, Donald E. Fry, and Michael Pine

Objective. Assess algorithms for linking patients across de-identified databases without compromising confidentiality.

Data Sources/Study Setting. Hospital discharges from 11 Mayo Clinic hospitals during January 2008–September 2012 (assessment and validation data). Minnesota death certificates and hospital discharges from 2009 to 2012 for entire state (application data).

Study Design. Cross-sectional assessment of sensitivity and positive predictive value (PPV) for four linking algorithms tested by identifying readmissions and posthospital mortality on the assessment data with application to statewide data.

Data Collection/Extraction Methods. De-identified claims included patient gender, birthdate, and zip code. Assessment records were matched with institutional sources containing unique identifiers and the last four digits of Social Security number (SSNL4).

Principal Findings. Gender, birthdate, and five-digit zip code identified readmissions with a sensitivity of 98.0 percent and a PPV of 97.7 percent and identified postdischarge mortality with 84.4 percent sensitivity and 98.9 percent PPV. Inclusion of SSNL4 produced nearly perfect identification of readmissions and deaths. When applied statewide, regions bordering states with unavailable hospital discharge data had lower rates.

Conclusion. Addition of SSNL4 to administrative data, accompanied by appropriate data use and data release policies, can enable trusted repositories to link data with nearly perfect accuracy without compromising patient confidentiality. States maintaining centralized de-identified databases should add SSNL4 to data specifications.

Key Words. Readmission, posthospital mortality, record linkage, de-identified data

Linkage of data is a fundamental building block for health information exchange. When patient-level data can be linked within and across health care-related data sources, investigators can address important health care policy issues including tracking patients across sites, relating interventions at one site to outcomes recorded at another, identifying posthospital outcomes, and distinguishing repeat procedures on a single patient from multiple procedures each performed on a different patient.

The number of states collecting patient-level health care utilization data is growing. In 2011:

- 48 states and jurisdictions had implemented hospital inpatient reporting systems
- 34 states had ambulatory surgery data reporting
- 31 states had emergency department reporting
- 13 states had implemented or were developing an All Payer Claims Database (APCD) (National Association of Health Data Organizations 2014)

However, in response to frequent troubling reports of personal and financial data compromise, many large state databases are maintained with few, if any, personal identifiers. For example, to comply with HIPAA (Services 2012) and address publicly aired privacy concerns, the Minnesota Hospital Association (MHA) collects no names or other personal identifiers for its statewide hospital discharge dataset.

Attempts have been made to link health care databases with limited personal identifiers. A systematic review of 33 record linkage strategy papers was published by Silveira (Silveira and Artmann 2009). Sensitivities for matching individuals in the papers were between 74 and 98 percent, and specificities were between 99 and 100 percent. The authors concluded that more studies were needed to determine the accuracy of alternative linkage procedures.

Our objective in this paper was to explore how best to link Minnesota hospital discharge data by validating possible linking algorithms on personally identified institutional data and then applying those results to statewide infor-

Address correspondence to James M. Naessens, Sc.D., Mayo Clinic, 200 First Street SW, Rochester, MN 55905; e-mail: naessens@mayo.edu. Sue L. Visscher, Ph.D., Stephanie M. Peterson, B.A., Kristi M. Swanson, M.S., Matthew G. Johnson, M.P.H., and Parvez A. Rahman, M.H.I., are with the Mayo Clinic, Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery, Rochester, MN. Joe Schindler, B.A., and Mark Sonneborn, M.S., are with the Minnesota Hospital Association, St. Paul, MN. Donald E. Fry, M.D., and Michael Pine, M.D., are with the Michael Pine Associates, Chicago, IL.

mation using hospital inpatient information for determining readmissions and adding in death certificates for identifying postdischarge mortalities.

METHODS

Data Sources

The following data sources were used to create and validate alternative algorithms (Figure 1):

1. Mayo-MHA discharge data: Hospital discharge claims submitted to MHA from 11 Mayo Clinic Health System and Mayo Clinic Rochester hospitals from January 2008 through September 2012. The 376,928 records included site-specific patient identifiers, discharge date, and discharge status.
2. MPII: Mayo Patient Identification Index, an operational repository which provides linkage to a unique Mayo patient number (approximately 8 million) across all Mayo sites. The index also included site-specific patient identifiers, demographic data including address and Social Security number (SSN), plus information on known deaths and death dates.
3. MN DC: Death certificates (184,604 records) from the state of Minnesota with date of death from 2008 to 2012. The data included decedent name, SSN, and address with five-digit zip codes which we converted to nine-digit zip codes with publicly accessible software, Code1-Plus from PitneyBowes (Bowes).
4. MHA discharge data: Hospital discharge claims for all Minnesota hospitals for January 2009 through September 2012. The 2,325,814 records included birthdate, sex, zip code (37 percent were nine-digit), hospitalization dates and discharge status. A subset of 2,155,836 records through June 2012 comprised the index dataset; the remaining records were only used to identify potential readmissions.

A gold standard readmission discharge dataset was created by adding the unique patient identifiers from MPII to the Mayo-MHA discharge data linked by site-specific patient identifiers. A gold standard death dataset was created by adding 355 previously unknown deaths occurring within 30 days of discharge to MPII by matching to MN DC on patient name, date of birth, and SSN. Perfect matches on two of the three criteria were considered part of the gold standard deaths.

Figure 1: Data Used to Validate and Apply Matching Algorithms for Identifying Posthospital Outcomes: (A) Validation, (B) Application

A Validation

B Application

Notes: 1—Match to create gold standard for readmissions; 2—Match to create gold standard for post-hospital deaths; 3—Assess algorithms for readmissions; 4—Assess algorithms for mortality (B) Application. *MPII: Mayo Patient Identification Index; **MN DC: Minnesota Death Certificates; MHA: Minnesota Hospital Association.

Algorithms

Since our goal was to determine how best to link Minnesota hospital discharge data, our algorithms were limited by the variables in MHA discharge data. However, because we had patients’ complete SSN in both the MPII and MN DC, we also tested the value of including the last four digits of the SSN as a potential anonymous linking factor. Matching algorithms were compared on their sensitivity (ability to identify “true matches”) and positive predictive value (PPV) (percent of algorithm matches which were correct). Deterministic algorithms used to link records included:

1. Gender, date of birth, and five-digit zip code
2. Gender, date of birth, and nine-digit zip code (using five-digit values where a nine-digit code was unavailable)
3. Gender, date of birth, and the last four digits of SSNs (missing SSNs were considered unmatched)
4. Algorithm 3 plus substituting five-digit zip codes for missing SSNs

Validation

We tested the algorithms against the Mayo-MHA discharge data to determine readmissions within 30 and 90 days of discharge to the same or another hospital. Transfers to another hospital were also included to increase analytical power. These identified events were then compared to true readmissions/transfers identified in the gold standard readmission discharge dataset to calculate sensitivity and PPV. Each hospital discharge not ending in death was considered as a potential readmission. The basis for readmission rates and for sensitivity was whether any subsequent hospital discharge was linked to an index discharge within the requisite timeframe. The basis for the number and percent of false-positive readmissions considered all matches within the timeframe and was not restricted to a single hospital discharge.

We also tested the algorithms against matching the MN DC with the Mayo-MHA discharge data to identify deaths within 30 days of discharge. These identified events were compared against the gold standard death dataset to measure sensitivity and PPV of the algorithms for 30-day postdischarge mortality.

Application

Matching algorithms were applied to MHA discharge data and to MN DC to estimate 30-day readmission rates and 30-day mortality rates by region in Minnesota. As Algorithm 1 was the only method that could be currently applied to the MHA discharge data, we calculated the readmission rate by applying Algorithm 1 to the data. For these estimates, we excluded hospital-to-hospital transfers. We estimated the impact of using Algorithm 4 which includes the last four digits of SSN through a simple model by applying the improvement in sensitivity and PPV of Algorithm 4 over Algorithm 1 to the previous estimates. Similar estimates were made for mortality rates.

RESULTS

Readmissions

The gold standard readmission discharge data contained 6,238 hospitalizations ending in death and 370,695 live discharges with potential readmissions or transfers. To allow for at least 3 months of follow-up, 355,638 discharges through June 2012 were classified as index discharges for potential readmission or transfer. Within the MHA discharge data, most hospitals use a site-specific patient identifier. While total readmissions are of interest to the MHA for measuring hospital quality, readmissions to other hospitals and transfers are of particular interest when testing the algorithm because their identification requires more than a site-specific identifier making them more difficult to detect and currently unknown statewide. Within the gold standard readmission discharge data, we identified 52,212 readmissions/transfers within 30 days (some discharges were followed by multiple readmissions within 30 days), of which 6,516 (12.5 percent) were readmitted or transferred to a hospital other than the one from which they were discharged. A 90-day follow-up period produced 78,326 (22.0 percent) readmissions, of which 8,896 (11.4 percent) were readmitted/transferred to another hospital.

Table 1 provides sensitivities and PPVs of the four algorithms for identifying readmissions. Sensitivities and PPVs for 90-day readmission were slightly lower than for 30-day readmission for Algorithms 1 and 2 and were unchanged or slightly increased for Algorithms 3 and 4.

Several other findings were observed when further examining algorithm mismatches. Two false-positive matches on Algorithms 1 and 2 appeared to be twins as they shared the same birthdate and address. The largest reason ($n = 1,036$, 95.7 percent) for not finding true readmissions with Algorithms 1 and 2 was due to different zip codes between hospitalizations. A large portion of these patients ($n = 203$) were discharged to skilled nursing facilities. For Algorithms 3 and 4, of the 7 percent of true readmissions missing SSNs, 32.2 percent were under age 1 with an additional 20.7 percent between ages 1 and 5.

In applying algorithms to Minnesota hospitals, only 37 percent of discharges had a complete nine-digit zip code and the last four digits of SSNs are not currently reported. Therefore, we were able to use only Algorithm 1 which matched on date of birth, gender, and five-digit zip code to identify readmissions to Minnesota hospitals other than those from which a patient was discharged. These discharges were added to the known 9.5 percent of 30-day readmissions back to the same institution to calculate an overall readmission rate resulting in a statewide rate of 11.9 percent. We estimated that apply-

Table 1: Summary on Readmission Sensitivity and Positive Predictive Value for Mayo-MHA Discharge Data

<i>Alg No.</i>	<i>Description</i>	<i>Sensitivity (%)</i>	<i>PPV (%)</i>	<i>No. of False Negatives</i>	<i>No. of False Positives</i>
(A) 30-day readmissions ($N = 52,212$ true readmissions of 355,638 discharges)					
1	DOB, sex, 5-digit zip code	95.5	97.7	295	150
2	DOB, sex, 9- or 5-digit zip code	94.1	98.4	386	101
3	Last 4 SSN digits, DOB, sex	95.7	99.9	283	3
4	DOB, gender, last 4 SSN or 5-digit zip code	98.8	99.3	77	43
(B) 90-day readmissions ($N = 78,326$ true readmissions of 355,638 discharges)					
1	DOB, sex, 5-digit zip code	94.8	96.8	461	276
2	DOB, sex, 9- or 5-digit zip code	93.2	97.7	604	193
3	Last 4 SSN digits, DOB, sex	96.1	99.9	346	7
4	DOB, gender, last 4 SSN or 5-digit zip code	98.9	99.3	96	64

Alg, algorithm; DOB, date of birth; Mayo, Mayo Clinic System; PPV, positive predictive value; SSN, Social Security number.

ing Algorithm 4 would increase the identification of readmissions resulting in an overall 30-day readmission rate of 12.3 percent, while dropping the number of false-positive cases from 1,209 to 430. Post-discharge admissions to Minnesota hospitals other than those from which a patient was originally discharged as calculated from Algorithm 1 are displayed by region in Figure 2. The map titled “Transfers included” includes both transfers and acute care readmissions. As our goal was finding the same patient at different institutions, combining transfers and readmissions in assessments of the effectiveness of the matching algorithms increased the sample size without biasing analyses. However, in assessing actual readmission rates, it is best to use patients’ discharge dispositions to exclude transfers from one acute care hospital to another, as most transfers are “elective.” The map titled “Transfers not included” shows regional rates after excluding transfers. Individual hospital 4-year average readmission rates were as high as 15.8 percent.

Death within 30 Days of Discharge

Some of the data sources were inconsistent with respect to information on deaths. While the information on deaths in hospital was well captured, follow-up death information was less consistent. At Mayo hospitals, there were 6,009 (1.7 percent) discharges in 4½ years of Mayo-MHA data with death as a disposition code. Of all the hospital deaths known from MMPI and Mayo-MHA

Figure 2: Applying Readmission Algorithm 1 to Minnesota Hospital Discharges by Region, January 2009 through June 2012; Readmission Rate to Other Hospitals with and without Including Hospital Transfers

data, 5,973 (99.4 percent) were found in MN DC. Fifty-eight hospital deaths miscoded on the discharge abstracts were verified and subsequently corrected in the gold standard death data (0.9 percent of all hospital deaths). Meanwhile, an additional 8,740 (2.4 percent) discharges in Mayo-MHA data were known to have died within 30 days of discharge from information in the MPII. Three hundred fifty-five additional deaths within 30 days of hospital discharge were found in MN DC and added to the gold standard (4.1 percent of 30-day deaths). However, 1,447 deaths were known in MPII data but not found in MN DC. Of these 1,447 deaths, almost all the missing death certificates were for non-Minnesota residents: 98.8 percent of deaths of Minnesota residents were documented in MN DC, but only 24.6 percent of deaths of non-Minnesota residents were documented in MN DC. For this reason, subsequent analyses were limited to Minnesota residents.

There were 260,342 discharges in the Mayo-MHA discharge data with a Minnesota zip code. There were 349 deaths within 30 days of discharge found in MN DC but not in MPII data, while only 81 were found in MPII but not in MN DC. There were 6,502 deaths found in both sources. All 6,932 discharges with a death recorded in either source were included in our gold standard for 30-day mortality.

Table 2 provides sensitivities and PPVs of each algorithm for detecting death within 30 days of hospital discharge. Of 193 known deaths not identified by Algorithm 4, 28 did not have SSNs available in the MPII. Of the 38 patients who matched on the last four digits of the SSN, 33 did not match on birthdate, four did not match on sex, and one did not match on the date of death. Fifty-two patients matched on all variables except last four digits of the SSN, and seven did not match on both SSN and date of death.

When applied across the state to 2,155,836 MHA discharge data index hospitalizations from January 2009 through June 2012 at all Minnesota hospitals, Algorithm 1 identified 40,026 deaths within 30 days of a hospital discharge (1.9 percent of eligible discharges). We estimated that applying Algorithm 4 would increase the identification of deaths by 14 percent, resulting in an overall 30-day mortality rate of 2.1 percent, while dropping the number of false positive cases from 440 to 92.

DISCUSSION

In this era of health care reform and pay-for-value, increasing attention is being placed on quality of care, particularly outcome measures such as 30-day

Table 2: Summary on 30-day Mortality Sensitivity and Positive Predictive Value for Mayo-MHA Discharge Data Using Only Minnesota Residents ($N = 6,932$ Deaths of 260,342 Discharges)

<i>Alg No.</i>	<i>Description</i>	<i>Sensitivity (%)</i>	<i>PPV (%)</i>	<i>No. of False Negatives</i>	<i>No. of False Positives</i>
1	DOB, sex, 5-digit zip code	85.0	98.9	1,038	66
2	DOB, sex, 9- or 5-digit zip code	72.8	99.7	1,886	16
3	Last 4 SSN digits, DOB, sex	96.1	99.9	270	2
4	DOB, gender, last 4 SSN or 5-digit zip code	97.2	99.8	193	16

Alg, algorithm; DOB, date of birth; MHA, Minnesota Hospital Association; PPV, positive predictive value; SSN, Social Security number.

readmissions and 30-day mortality. In Minnesota and in many other current statewide databases, reports on readmissions are only able to report on readmissions returning to the same facility. In this study we examined four matching algorithms enabling the identification of 30-day and 90-day readmissions and 30-day posthospital death from de-identified Minnesota hospital discharges. We saw that Algorithm 4 which includes the last four digits of SSN performed better than the alternatives. When examining the results of identifying readmissions and posthospital deaths for all of the Minnesota data based on Algorithm 1: date of birth, sex, and zip code, we found a 30-day readmission rate for the entire state of 11.9 percent for the entire 4-year period. The previously known rate of 30-day readmissions to the hospital from which patients were discharged was 9.5 percent. The proportion of 79.8 percent of all 30-day readmissions returning to the same hospital is similar in Minnesota to that found in California (Davies et al. 2013). Furthermore, we estimated that 1.9 percent of patients died within 30 days of hospital discharge.

We selected this algorithm for our statewide analyses because it performed reasonably well when compared with our three alternative algorithms and, as hospitals currently do not report last four digits of SSNs and only 37 percent report nine-digit zip codes, this was the only algorithm we could run without collecting additional data. Furthermore, the generalizability of these findings is enhanced in that many, if not all, state discharge data systems include these three fields necessary to perform this matching.

In trying to capture more accurate patient-identifying information, organizations face many privacy and security concerns, as well as state and federal regulations. Therefore, we selected to use Algorithm 1. However, in addition to focusing only on existing data, we also considered two potential enhancements to improve linkage: expanding zip codes to nine digits and using the last four digits of the SSN. Procedures to identify nine-digit zip codes from residential addresses are readily available. The use of the last four digits of the SSN as a relatively anonymous verification tool for identification is widely used in commerce. Of the four algorithms we evaluated, Algorithm 4, which used date of birth, gender, four-digit SSNs, and five-digit zip codes when four-digit SSNs were unavailable, appeared to have the best overall sensitivity and PPV. If the last four digits of SSN were available, we estimate that the number of identified readmissions would increase by 3.5 percent statewide, while the number of identified posthospital deaths would increase by 14.4 percent while dropping the number of false-positive cases by 64.4 percent for readmissions and by 79.2 percent for deaths.

Including the last four digits of the SSN reduced the number of missed true readmissions and posthospital deaths and actually also reduced the number of false positives. Nine-digit zip codes allowed more precise identification of patient residence; however, we saw a sizeable number of patients changing their address within 30 days of hospital discharge (e.g., they moved to skilled nursing or hospice facilities or to the home of a close friend or relative), particularly among patients discharged to extended care facilities. Missing SSNs were most common for patients under age 5; the majority occurred in patients under age one. Based on this research, we recommended that Minnesota hospitals report, as part of their submission of discharge abstracts to MHA, both the last four digits of SSNs and nine-digit zip codes.

Our study does have several limitations. Our gold standard database is based on the multiyear experience of readmissions and mortality following hospital discharge at community hospitals and a national referral center in a relatively rural area in the Midwest United States. In particular, zip code may serve as a better surrogate for patient identification in rural areas than in more densely populated areas, which may experience higher rates of false positive linkages. Our assessment of linkage algorithms relied on the use of PPVs rather than specificity. This could be viewed as a limitation in that PPVs are influenced by the underlying rates of readmission and posthospital death, whereas specificity is not. However, knowing the proportion of true-positive screens is more useful for comparing linkage algorithms. Some states do not have site-specific patient identifiers available; however, we found no evidence that algorithm performance would be worse for readmissions within a hospital than they were for readmissions between hospitals.

Without the ability to accurately link patient records across institutions and with death data, regional differences in out-of-hospital deaths and readmissions cannot be determined. Limitation in access to out-of-state data will remain a challenge for statewide outcome measurement.

CONCLUSION

Inclusion of the last four digits of patients' SSNs to de-identified health care claims can enable trusted data repositories to encrypt linked patient-level data within and across health-related databases with nearly perfect accuracy and without compromising patient confidentiality. All states that maintain centralized claims databases containing de-identified data should attempt to add the

last four digits of patients' SSNs to their data specifications. Even with nearly perfect linkage, care must still be taken to account for loss of vital information that may not be provided from routinely tapped sources of data.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This study was supported by Agency for Research and Quality Grant 1R01HS020043 awarded to the Minnesota Hospital Association with a subaward extended to the Mayo Clinic. We acknowledge the contributions of Jaclyn Roland, Minnesota Hospital Association, for data extraction support and Sara K. Hobbs, Caroline Plank, and Teresa Koski for technical and editorial support, as well as the anonymous reviewers for numerous helpful suggestions. None of the authors reported a conflict of interest with respect to this project.

Disclaimers: None.

REFERENCES

- Bowes, P. Code-1 Plus (release.)
- Davies, S. M., O. Saynina, K. M. McDonald, and L. C. Baker. 2013. "Limitations of Using Same-Hospital Readmission Metrics." *International Journal for Quality in Health Care* 25 (6): 633–9.
- National Association of Health Data Organizations. 2014. "Data System Tech Resources" [accessed on October 2, 2014] Available at https://www.nahdo.org/data_resources
- Silveira, D. P., and E. Artmann. 2009. "Accuracy of Probabilistic Record Linkage Applied to Health Databases: Systematic Review." *Revista de Saude Publica* 43 (5): 875–82.
- US Department of Health & Human Services. 2012. "Summary of the HIPAA Privacy Rule" [accessed on February 20, 2015]. Available at <http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.