American Journal of Epidemiology

Vol. 173, No. 9
DOI: 10.1093/aje/kwq482
Advance Access publication:
March 23, 2011

Practice of Epidemiology

Use of a Medical Records Linkage System to Enumerate a Dynamic Population Over Time: The Rochester Epidemiology Project

Jennifer L. St. Sauver, Brandon R. Grossardt, Barbara P. Yawn, L. Joseph Melton III, and Walter A. Rocca*

* Correspondence to Dr. Walter A. Rocca, Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (e-mail: rocca@mayo.edu).

Initially submitted June 10, 2010; accepted for publication December 9, 2010.

The Rochester Epidemiology Project (REP) is a unique research infrastructure in which the medical records of virtually all persons residing in Olmsted County, Minnesota, for over 40 years have been linked and archived. In the present article, the authors describe how the REP links medical records from multiple health care institutions to specific individuals and how residency is confirmed over time. Additionally, the authors provide evidence for the validity of the REP Census enumeration. Between 1966 and 2008, 1,145,856 medical records were linked to 486,564 individuals in the REP. The REP Census was found to be valid when compared with a list of residents obtained from random digit dialing, a list of residents of nursing homes and senior citizen complexes, a commercial list of residents, and a manual review of records. In addition, the REP Census counts were comparable to those of 4 decennial US censuses (e.g., it included 104.1% of 1970 and 102.7% of 2000 census counts). The duration for which each person was captured in the system varied greatly by age and calendar year; however, the duration was typically substantial. Comprehensive medical records linkage systems like the REP can be used to maintain a continuously updated census and to provide an optimal sampling framework for epidemiologic studies.

censuses; cohort studies; data collection; epidemiologic research design; information systems

Abbreviation: REP, Rochester Epidemiology Project.

There is a long tradition of using medical records linkage techniques to create extensive research databases (1). Among English-speaking countries, research databases have been implemented in the United Kingdom (2–6), Australia (7), and Canada (8, 9). However, similar databases have been more limited in the United States because of the lack of a national health system. Only in recent years have attempts been made at the federal level to create publicly accessible databases for research (10). However, these efforts have been hindered by equally strong trends toward strict confidentiality of medical record information (11). Even if national databases become available to investigators, they will lack historical depth and will not be able to answer long-term questions of public health relevance.

By contrast, the Rochester Epidemiology Project (REP) is a rare example of a medical records linkage system in the United States that has almost half a century of activity. The REP has linked and archived the medical records of virtually all persons residing in Olmsted County, Minnesota, for over 40 years, has maintained an electronic index of medical diagnoses and surgical interventions, and has archived all addresses and demographic information over time (12, 13). The REP allows investigators to follow subjects through their outpatient (office, urgent care, or emergency department) and hospital contacts across all local medical facilities, regardless of where the care was delivered and of insurance status. The continuing linkage of medical information also provides a virtually complete enumeration of the Olmsted County population at any point in time (the REP Census). Thus, the REP allows investigators to conduct long-term population-based studies of disease incidence, prevalence, risk and protective factors, outcomes, health services utilization, and cost-effectiveness (1, 12, 13). As of today, however, the linkage of records across

health care providers in Olmsted County is restricted to research applications and cannot be used directly for patient care.

In the present article, we describe the methods used to link medical records from multiple health care institutions to specific individuals and to establish residency in Olmsted County over time. In addition, we provide evidence of the validity of the REP Census.

MATERIALS AND METHODS

Sources of data

In 2008, the health care institutions participating in the REP included the Mayo Clinic and its 2 affiliated hospitals (St. Marys Hospital and Rochester Methodist Hospital), the Olmsted Medical Center and its affiliated hospital (Olmsted Medical Center Hospital), and the Rochester Family Medicine Clinic (all in Rochester, Minnesota) (Figure 1). Historically, the medical records have been primarily in paper form; however, diagnostic codes, surgical procedure codes, and demographic information (including all names, sexes, dates of birth, and Social Security numbers, when available), as well as the physical location of each record, have been entered into a single electronic REP database. In addition, the REP has electronically archived all home addresses of Olmsted County residents over time. Since 1995, electronic medical records have progressively replaced the paper forms. To enumerate the population, we first linked medical records across different health care providers to create a list of unique subjects (person component). Second, we applied residency criteria and imputations to describe the residency status of the subjects over time (time component). Thus, we obtained a complete description of a dynamic cohort of persons over time.

Electronic linkage methods

Because Olmsted County residents often receive medical care from multiple institutions over the course of their lives, it is necessary to link records across institutions to obtain patients' complete history. Adding to the complexity is the fact that patients who received care at Olmsted Medical Center could have multiple records and multiple identification numbers for outpatient and inpatient services. Since its establishment in 1966, the REP has matched the medical records from participating institutions to specific individuals on a study-by-study basis (Figure 1). Matching was done on a probabilistic basis, and investigators had to clarify matches with low scores. This study-by-study probabilistic linkage was time-consuming, expensive, prone to errors, and unsatisfactory for many users of the REP.

Starting in 2002, the process of linking records to individuals was formalized by an initial series of exact matching iterations, as described in Figure 2. The first 9 rounds of matching focused on name variants, sex, and date of birth information. Rounds were repeated using all known names for a person (e.g., including name changes as the result of adoption, marriage, or divorce). The Soundex phonetic filing system of the SAS software package (SAS Institute, Inc., Cary, North Carolina) was used to match names that could be easily misspelled (14). Four additional rounds of matching used Social Security numbers (when available) and 2 internal identification numbers assigned by the Olmsted Medical Center. Finally, subgroups of records with special characteristics (e.g., records with no birth date) were subjected to 7 additional rounds of matching (Figure 2). This exact matching resulted in a primary master file of individuals.

Subsequently, the REP was continually updated through the matching of newly generated records by using the same criteria (Figure 2). Currently, all REP data are stored in a Sybase database (Dublin, California) and a SAS structured query language function (SQL) is used to match new records with records stored in the existing database (PROC SQL). Uncertain matches are verified by the REP

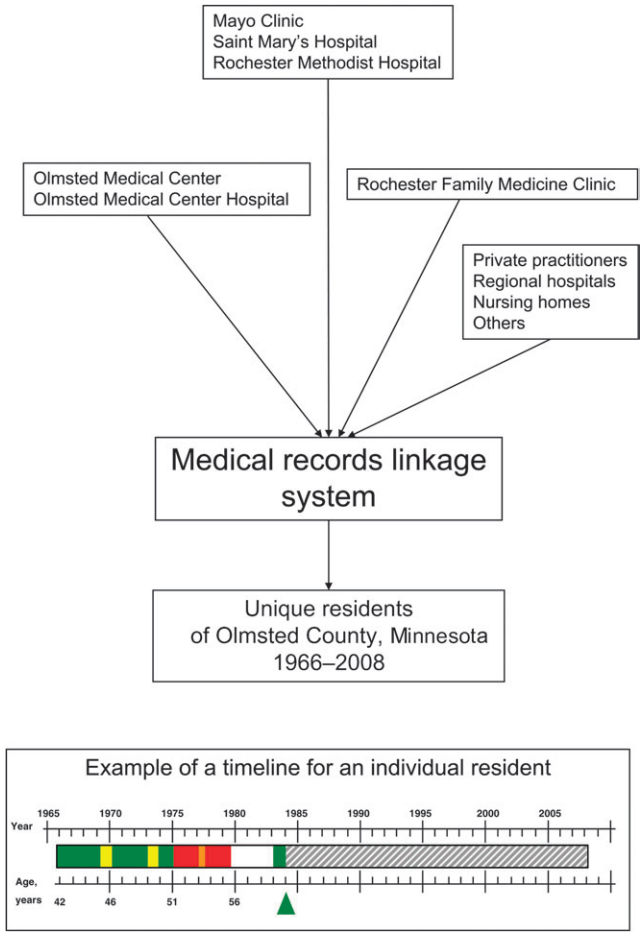

Figure 1. Steps involved in linking medical records from multiple care providers to unique residents of Olmsted County, Minnesota, 1966–2008. The lower part of the figure shows an example of a timeline for a person who entered the system in 1966 at the age of 42 years, lived in and out of Olmsted County for a number of years, and died in 1984 at age 60 years while residing in the County (green arrow).

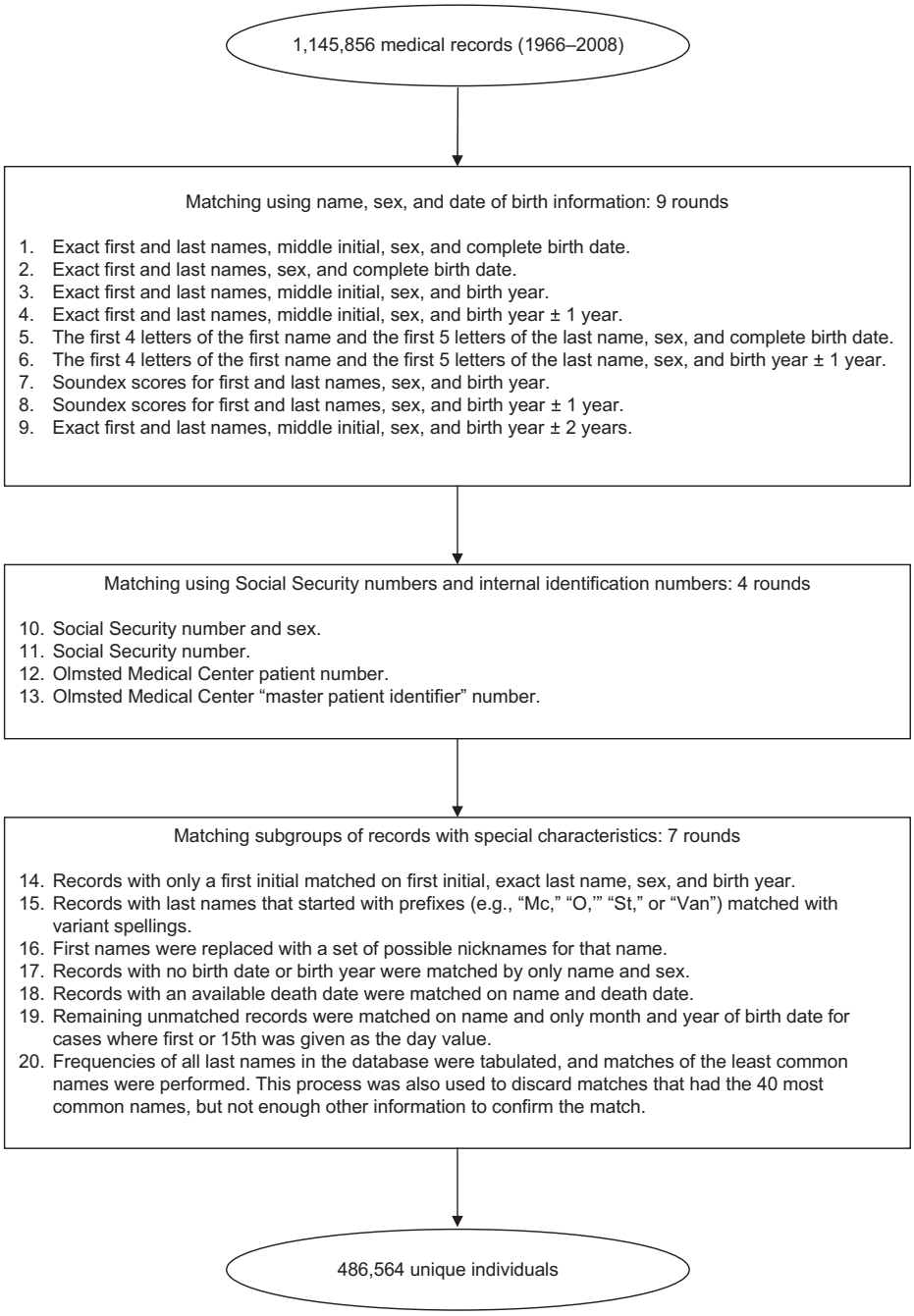

Figure 2. Procedures used to link multiple medical records to single individuals in the Rochester Epidemiology Project from 1966 to 2008. The 20 rounds of exact matching are presented in hierarchal order from the most perfect match to the least perfect match.

technical team. In addition, if any error in matching is discovered by any user of the system, the REP technical team investigates it manually and corrects it when needed.

Manual checks

Although our electronic linkage methods use the full name, sex, and birth date of each patient, the medical records frequently contain extensive additional data (e.g., spouse's name) that can be used to verify the electronic matching. Manual verification of new records is routinely performed when one of the following problems is encountered: 1) missing first or last names; 2) discordant birth dates; 3) discordant middle names; 4) first name listed only as ''baby boy'' or ''baby girl''; or 5) last names that are common in Olmsted County. Additional manual checks were completed on

Table 1. Residency Rules Used to Enumerate the Population of Olmsted County, Minnesota, Rochester Epidemiology Project, 1966–2008

Included Subjects	Excluded Subjects
Persons who received at least 1 medical diagnosis in the system between January 1, 1966, and December 31, 2008, with a corresponding residential Olmsted County address on the date of diagnosis or within ± 1 year of the date of diagnosis.	Persons with addresses in Olmsted County who lived in homes, schools, hospitals, or wards for the physically or mentally disabled, or mentally ill; in drug/alcohol recovery facilities; in juvenile institutions; or in area hotels and motels, with no previous residential Olmsted County address.
Persons residing in an Olmsted County nursing home, sanitarium, or state hospital for at least 12 months.	Inmates of the Rochester Federal Medical Center (prison).
College students attending college at a campus in Olmsted County or returning from college and residing in Olmsted County.	Persons residing in a nursing home, sanitarium, or state hospital for <12 months with no previous residential Olmsted County address[a].
Persons with a current "care of" address but with a residential Olmsted County address before the "care of" address.	
People who move between multiple residences during the year ("snowbirds"), as long as they maintain an Olmsted County address at the date of health care visits within the system.	
Retired nuns living at the Assisi Heights Convent located in Rochester, Minnesota.	

[a] Persons who resided in an institution for <12 months were considered residents if they moved to the institution from their home in Olmsted County.

random samples of records that failed to match electronically with any other record for a series of specified reasons (e.g., the subject had only a birth year listed or had no first name recorded).

Validation of the linkage methods

To validate our linkage methods, we obtained 2 age-stratified random samples of patients who visited at least 1 REP care facility during 1985 ($n = 200$) or 2005 ($n = 200$). For each person, we manually reviewed all records to determine whether there were any other potential names or name spellings, to verify or obtain dates of birth, and to obtain Social Security numbers and all possible addresses. These data were then used to determine whether any record was incorrectly linked to a given person or whether any other records in the REP database matched the person of interest. We calculated the proportion of incorrect linkages (overinclusion of records and false positives) and missed linkages (underinclusion of records and false negatives) and compared results across strata by age, sex, and calendar year (1985 and 2005).

Establishment of Olmsted County residency and the REP Census

Before 2008, the residency of any subject in Olmsted County at a given point in time was determined on a study-by-study basis through a labor-intensive process that involved manual review of all addresses. In 2008, however, we constructed and released the first formal Olmsted County census based on the medical records linkage system (the REP Census). Dates of residency in Olmsted County were established by using the addresses associated with dates of medical visits. Individuals were considered residents if they had an Olmsted County address at the time of a medical visit between 1966 and 2008. Additionally, specific rules were established for persons residing in nursing homes, colleges, motels, or incarceration facilities and those in other special situations. These rules were tailored to the local population but were similar to those used by the US Census Bureau and are summarized in Table 1 (15).

Graphic displays of residency data were then created for each person who resided in Olmsted County at any point between 1966 and 2008 (REP timelines). The timelines indicated the dates during which residency in (green bars) or outside of (red bars; Figure 1) Olmsted County could be confirmed. Residency was assumed for 1 year before and 1 year after any medical contact for subjects aged 3 years and older. Additionally, residency was imputed for women who had 2 contacts as an Olmsted County resident separated by up to 3 years (yellow bars), and nonresidency was imputed for women who had 2 contacts while residing outside of Olmsted County separated by up to 3 years (orange bars). The corresponding gap for men was 4 years. The rules for these imputations were derived empirically from age- and sex-specific patterns of medical contacts. For children aged less than 3 years, residency was assumed only for 6 months before and 6 months after any medical visits. Imputations of residency between visits were again up to 3 years for girls and 4 years for boys. Finally, the timelines included arrows indicating births and deaths that have occurred since 1966. An updated version of the REP Census is released annually, and all historical REP Censuses are permanently archived.

Validation of the REP Census enumeration

Two groups of Mayo Clinic investigators have conducted reliability studies of residency status as determined by the REP Census. Each group selected a sample of individuals from the REP population and manually reviewed the medical records to determine residency on a particular date. The results from the manual review were then compared with the results obtained electronically from the REP Census.

In addition, we compared the REP Census enumeration with the US Census enumerations for April 1 of 1970, 1980, 1990, and 2000 (15–21). The capture rate was computed by dividing the number of subjects enumerated using the REP Census by the number published by the US Census Bureau for specific age, sex, and calendar year strata. Finally, we compared the mortality rates (probability of dying within 1 year of a given reference age) (22) obtained using the REP data with those obtained using publicly available vital statistics (23). In both calculations, we used smoothing methods that have been described elsewhere (24).

Impact of the Minnesota confidentiality law

In 1996, the state of Minnesota passed a new law requiring each patient to provide a general written authorization if they chose to allow researchers permission to review their medical records for research (25–28). All health care providers affiliated with the REP established procedures to comply with the law for all subjects attended to after January 1, 1997. Two attempts to obtain a research authorization from each participant were made in writing (via the US mail) with at least 60 days between attempts. If the patient gave explicit authorization or did not respond after these 2 attempts, then the record was considered accessible for research purposes. Authorization for review of existing medical record data was also implied for patients who did not see a medical professional after January 1, 1997. The authorization does not expire, but it can be revoked and covers any research project that has received institutional review board approval. The impact of the law on participation of patients in passive medical record research was studied at the Mayo Clinic and Olmsted Medical Center (26, 27). In addition, we computed the participation rate for subjects in Olmsted County who were included in our census from 1998 through 2007 (10 years after the introduction of the law).

RESULTS

Linkage results

Overall, 1,145,856 medical records were available for residents of Olmsted County who had visited a local health care provider at least once between January 1, 1966, and December 31, 2008 (Figures 1 and 2). These records were electronically matched through the computerized algorithms described in Figure 2 to 486,564 unique individuals (the REP Census population), for a median of 2 records per person (range, 1–23 records). A total of 215,127 individuals (44%) had only 1 medical record (i.e., received care from only 1 facility), whereas 271,437 individuals (56%) had 2 or more records in the system (i.e., received care from multiple facilities).

Validity of the linkage methods

The sample of 400 Olmsted County residents who had visited an Olmsted County medical provider in 1985 (200 subjects) or in 2005 (200 subjects) yielded a total of 1,319 medical records, with a median of 3 records per person (range, 1–11 records). The median number of records did not differ by age or sex (Table 2). Among these 400 people, 10 had at least 1 incorrect record included (rate of overinclusion = 2.5%; 95% confidence interval: 1.0, 4.0). The proportion of incorrect matches did not differ by age or sex (Table 2). Additionally, 5 individuals were missing at least 1 record (underinclusion rate = 1.3%; 95% confidence interval: 0.2, 2.4). Again, the proportion of individuals missing records did not differ by age or sex (Table 2). Finally, neither the overinclusion nor the underinclusion percentages differed by sampling year. In both 1985 and 2005, a total of 5 individuals out of 200 had at least 1 incorrect record included (2.5%). In 1985, a total of 4 individuals out of 200 were missing at least 1 record (2.0%), whereas in 2005, 1 individual out of 200 was missing at least 1 record (0.5%).

Validity of the census enumeration

Two Mayo Clinic research groups compared the residency status determined by the REP Census on a given index date with that obtained by manual review of the records. The first study included 201 individuals and the second study included 447 individuals. The REP Census did not have residency data for 11 individuals (5.5%) from the first study or 12 individuals (2.7%) from the second study (total of 23; 3.5%). Manual review of the medical records of these individuals indicated that 22 of the 23 subjects were nonresidents at the index date (95.7%), thus confirming that the REP Census was valid. Among the remaining 625 individuals for whom we had both REP Census and manual-review residency information, agreement between the 2 sources was 96.8% (Table 3).

The REP Census population estimates were also compared with US Census estimates for the Olmsted County population in 1970, 1980, 1990, and 2000. Overall, the REP population counts at comparable points in time were slightly higher than those reported by the US Census (1970, 104.1%; 1980, 103.5%; 1990, 102.4%; and 2000, 102.7% of the US Census counts). As shown in Figure 3, however, these comparisons varied by age and sex. For example, the REP consistently overcounted individuals aged 20–29 years compared with the US Census (men, 110.5%–122.8% of US Census counts; women, 121.4%–138.9% of US Census counts). By contrast, the REP consistently undercounted individuals aged 40–69 years compared with the US Census (<10%; Figure 3). Over time, the REP Census estimates have moved closer to the US Census estimates (data not shown). In addition, the mortality rates computed using the REP data were similar to those computed using publicly available vital statistics data for both men and women and for all ages (Figure 4).

Table 2. Distribution of Subjects With Incorrect Linkage of Medical Records by Age and Sex (400 Individuals)[a], Rochester Epidemiology Project, 1966–2008

Age, years	No. of Subjects	Median No. of Records per Subject (25%, 75% Percentile)	Overincluded[b]			Underincluded[c]		
			No. of Subjects	%	95% CI	No. of Subjects	%	95% CI
Men								
0–19	45	3 (2, 4)	1	2.2	0.0, 6.5	0	0.0	
20–39	42	3 (2, 4)	0	0.0		2	4.8	0.0, 11.3
40–59	47	4 (2, 4)	2	4.3	0.0, 10.1	0	0.0	
≥60	44	3 (2, 4)	1	2.3	0.0, 6.7	0	0.0	
Total	178	3 (2, 4)	4	2.3	0.1, 4.5	2	1.1	0.0, 2.6
Women								
0–19	49	3 (1, 4)	0	0.0		0	0.0	
20–39	62	3 (2, 5)	4	6.5	0.4, 12.6	0	0.0	
40–59	51	3 (2, 4)	1	2.0	0.0, 5.8	0	0.0	
≥60	60	3 (2, 4)	1	1.7	0.0, 5.0	3	5.0	0.0, 10.5
Total	222	3 (2, 4)	6	2.7	0.6, 4.8	3	1.4	0.0, 2.9
Total								
0–19	94	3 (1, 4)	1	1.1	0.0, 3.2	0	0.0	
20–39	104	3 (2, 4)	4	3.9	0.2, 7.6	2	1.9	0.0, 4.5
40–59	98	3 (2, 4)	3	3.1	0.0, 6.5	0	0.0	
≥60	104	3 (2, 4)	2	1.9	0.0, 4.5	3	2.9	0.0, 6.1
Total	400	3 (2, 4)	10	2.5	1.0, 4.0	5	1.3	0.2, 2.4

Abbreviation: CI, confidence interval.
[a] The overall random sample included 2 subsamples recruited in 1985 ($n = 200$) and 2005 ($n = 200$).
[b] Subjects with at least 1 record incorrectly included.
[c] Subjects with at least 1 record missing from the linkage.

Length of time in the system

Web Figure 1 (available at http://aje.oxfordjournals.org/) shows the distribution of the median length of time for which subjects were captured in the system on April 1 of 1970, 1980, 1990, and 2000. The duration varied greatly by age and calendar year. For example, individuals who were aged 0–9 years in 1970 had a median of 3.5 years of medical information available before that date and 23.1 years of follow-up after that date. By contrast, individuals who were aged 70–79 years in 1970 had a median of 19.2 years of information available before that date and 8.7 years of follow-up after that date. Medical information before 1966 was available electronically only through approximately 1950 because diagnostic codes were not fully captured electronically before 1950. We emphasize that further historical information is available before 1950 but requires manual review of the paper medical dossier.

Table 3. Agreement of Residency Status Obtained from the REP Census Versus Manual Medical Record Review, Rochester Epidemiology Project, 1966–2008

Study	Total No. of Subjects	Subjects With No REP Data		Subjects With REP Data		Agreement[a]				Disagreement[b]				% Agreement	95% CI
						+/+		–/–		+/–		–/+			
		No.	%	No.	%	No.	%	No.	%	No.	%	No.	%		
First study	201	11	5.5	190	94.5	183	96.3	2	1.0	5	2.6	0	0.0	97.4	95.1, 99.6
Second study	447	12	2.7	435	97.3	277	63.7	143	32.9	11	2.5	4	0.9	96.6	94.8, 98.3
Combined sample	648	23	3.5	625	96.5	460	73.6	145	23.2	16	2.6	4	0.6	96.8	95.4, 98.2

Abbreviations: CI, confidence interval; REP, Rochester Epidemiology Project.
[a] Positive agreement is shown as +/+, and negative agreement is shown as –/–.
[b] Numbers given under the +/– column represent those persons for whom the medical records linkage system indicated Olmsted County residency, but the residency was not confirmed by manual review. Numbers given under the –/+ column represent those persons for whom the manual review indicated Olmsted County residency but the medical records linkage system did not agree.

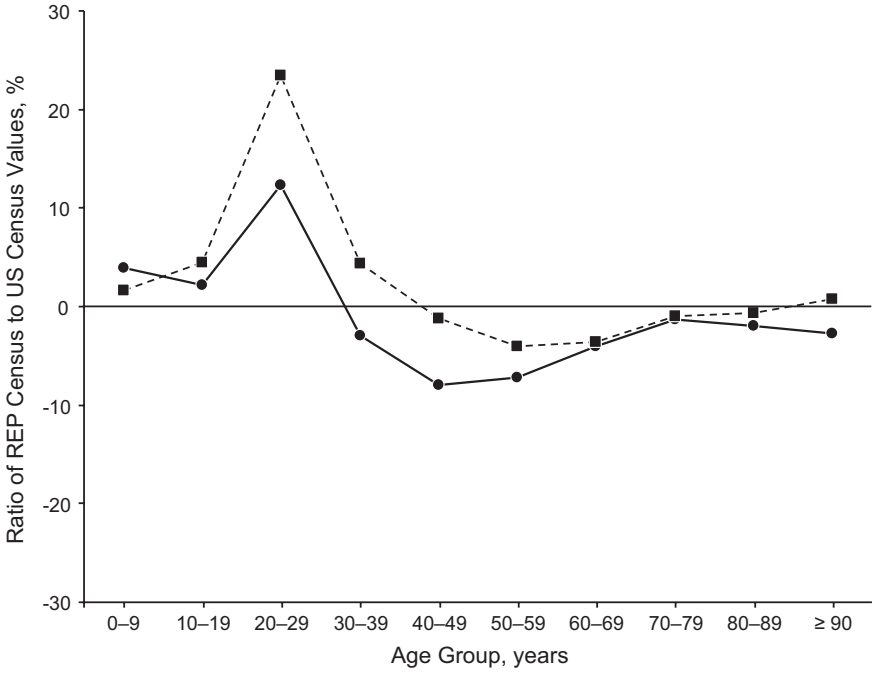

Figure 3. Age- and sex-specific capture rate by the Rochester Epidemiology Project (REP) medical records linkage system compared with US Census data (median capture rate in 1970, 1980, 1990, and 2000). Data from men (solid line, circle points) and women (dashed line, square points) are shown separately. The 0% line corresponds to perfect agreement between the system and the US Census. Values plotted above the 0% line indicate that the REP counted more persons than the US Census; values plotted below the 0% line signify that the REP counted fewer persons than the US Census.

Because of the historical span of the REP, the median length of follow-up for all age-groups became shorter as the index calendar year moved from 1970 through 2000. For example, a person aged 60–69 years in 1970 had a median of 15.6 years of follow-up, whereas a person of the same age in 2000 had only 9.0 years of follow-up. By

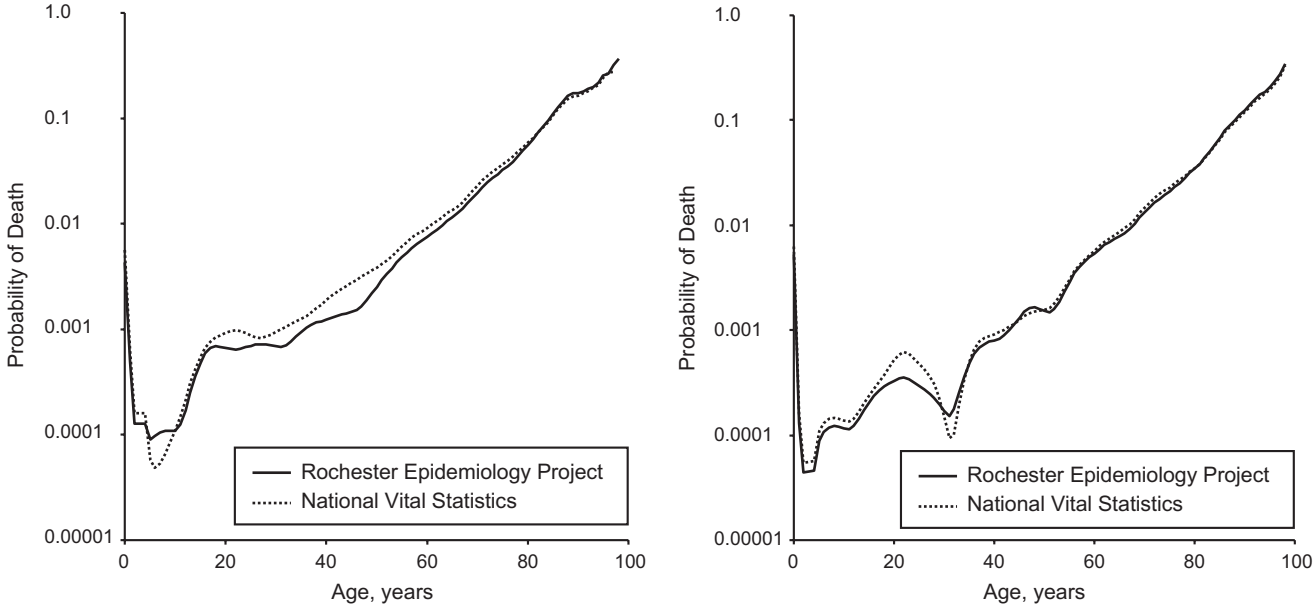

Figure 4. Age- and sex-specific mortality rates for Olmsted County in 2000 computed using Rochester Epidemiology Project data (solid line) compared with rates derived from national vital statistics data (dotted line). Mortality rates are shown separately for men (left panel) and women (right panel).

contrast, as the index date moved closer to the present, the median number of years of information available before the date increased. For example, a person aged 60–69 years in 1970 had 18.9 years of information available before that date, whereas a person of the same age in 2000 had 39.5 years of information before that date.

Impact of the Minnesota confidentiality law

In 2 studies of the impact of research authorization on research, 97% of 2,463 individuals seen from 1994 through 1996 provided authorization at the Mayo Clinic, whereas 96% of the 15,997 patients seen in January or February 1997 provided authorization at Olmsted Medical Center (26, 27). Subjects who did not respond to 2 requests were considered to have provided authorization. In the Mayo Clinic study, the refusal rate was higher for women, younger subjects, local subjects, and subjects with prior sensitive diagnoses (e.g., mental disorders) (27). A total of 90.7% of the subjects residing in Olmsted County from 1998 through 2007 gave authorization to all health care providers, an additional 7.2% gave authorization to at least 1 health care provider, and only 2.1% denied authorization to all health care providers included in the REP.

DISCUSSION

In the present article, we have described our use of a medical records linkage system to enumerate a dynamic population over more than 40 years (1966–2008). The REP links medical information from multiple care providers to single individuals, establishes residency in Olmsted County, and provides the length of residency in this community. This process has primarily involved matching records by using computer algorithms but has been augmented by routine manual verification of questionable matches.

Validation of our matching process through a manual verification of 400 randomly selected individuals (with 1,319 medical records) suggested that only a small proportion of the individuals in the database have incorrect record inclusions (2.5%). Similarly, only a small proportion of individuals had missed records that should have been matched to them (1.3%). These data suggest that our linkage methods have high sensitivity (ability of the REP to correctly link records that belong to the same person) and excellent specificity (ability of the REP to correctly exclude the linkage of records that do not belong to the same person). However, records with very different name spellings or records with other names that were not already linked would have been difficult to identify with our validation process. The only way for us to completely identify all missing records would be to hand-check approximately 1.1 million records. Because this was not feasible, we may have overestimated the sensitivity of the system.

Our results were comparable to those of other investigators who used similar methods. Using probabilistic linkage techniques, Dean et al. (29) had false-positive linkage rates of 2.2%–4.7% (overinclusion) when linking emergency medical service data to hospital discharge data. Victor et al. (30) achieved a sensitivity of 92% and a specificity approaching 100% by applying both exact and probabilistic matching techniques to several commercial insurance claim databases. However, our matching results were not as complete as those of the Western Australian Health Services Research Linked Database, in which 7 million records from 6 core data sets were matched with a false-positive linkage rate of 0.1% and a false-negative linkage rate of 0.1% (7). Our inability to achieve this level of linkage precision might have been due to the incomplete demographic information available in some of our older records.

Before 2008, the validity of the REP Census enumeration was supported by 2 studies (12, 31). In a study conducted in the 1980s, the REP enumeration was found to be virtually complete compared with the results from a random digit dialing telephone survey and with a list of all residents in nursing homes and senior citizen complexes in the city of Rochester (12, 32). In 2005, the REP enumeration was also found to be complete compared with a commercial list of noninstitutionalized Olmsted County residents who were 18 years of age or older. In particular, 6,723 of 6,996 (96.1%) subjects from the commercial list were correctly matched to a person in the REP Census (31).

After 2008, the 2 independent studies reported here showed that the residency status from the REP Census agreed with information obtained through hand-review of the complete medical records 96.8% of the time. Additionally, in virtually all cases where the REP lacked residency information, the person was not a resident of Olmsted County on the date of interest. Further, the REP captured approximately the same number of individuals residing in Olmsted County as expected from the US Census data from 1970 to 2000. However, the capture rate varied by age, sex, and calendar year. In the youngest age group (aged 0–9 years), the REP captured up to 4.6% more individuals than expected. Because children aged 0–4 years are seen frequently for routine care (33), the REP captured children who resided in Olmsted County for only a brief period of time and who were missed in the periodic US Census.

Since 1970, the REP has also counted 10.5%–38.9% more individuals than expected in the 20–29 year age group. This overcounting may have happened because the REP includes young adults who are full-time students and are covered by their parents' health insurance plans up to age 25 years, regardless of where they live. By contrast, the US Census counts young adults as residents of the location where they are living at the time of the census.

Among individuals aged 40–69 years, the REP estimates have been slightly lower than expected since 1970 (but within 10% of the US Census counts). This undercounting was more pronounced in men and could have happened because some subjects contact medical facilities infrequently in their adult life. The capture rate has also changed over time. In general, the REP population counts are more similar to the US Census estimates in 2000 than in 1970, suggesting a progressive convergence of the 2 methods over time. The validity of the REP Census enumeration is also confirmed by a comparison of age- and sex-specific mortality rates derived from REP data with those derived from national vital statistics data (22).

The length of time subjects are covered by the medical records linkage system before or after an index calendar year has important implications for the design of case-control studies, in which the time before disease onset is important, and cohort studies, in which the time after a given exposure has occurred is important (1, 34). The long-term follow-up of individuals makes the REP optimal for conducting historical cohort studies, even decades after an exposure has occurred (35, 36). This rich information also makes it possible to examine early life exposures in case-control studies without relying on self-reporting and memory (37, 38).

Currently, REP studies that involve only review of existing medical records can be conducted without obtaining study-specific written informed consent if the investigators obtain a Health Insurance Portability and Accountability Act waiver from the Mayo Clinic and Olmsted Medical Center institutional review boards (39). Waiver of informed consent is provided because subjects have signed the Minnesota state research authorization and because obtaining written consent for each specific study would be almost impossible (the study may span decades and include thousands of subjects), could potentially cause harm to the patient (inform the patient of events or diagnoses of which he/she is not aware), and would pose a respondent burden (repeated mail contacts by different investigators). This practice is consistent with the recommendation by the Council for International Organizations of Medical Sciences (1, 40–42). The general Minnesota research authorization and the Health Insurance Portability and Accountability Act waiver have allowed investigators using the REP to conduct studies with high participation rates.

Comprehensive medical records linkage systems like the REP can be used to maintain a continuously updated census of the population over time. Our experience could guide other investigators in designing medical records linkage projects. In addition, this article provides background information to readers of studies that are based on the REP.

ACKNOWLEDGMENTS

Author affiliations: Division of Epidemiology, Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, Minnesota (Jennifer L. St. Sauver, Barbara P. Yawn, L. Joseph Melton III, Walter A. Rocca); Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, Minnesota (Brandon R. Grossardt); Department of Neurology, College of Medicine, Mayo Clinic, Rochester, Minnesota (Walter A. Rocca); and Department of Research, Olmsted Medical Center, Rochester, Minnesota (Barbara P. Yawn).

This work was supported by the National Institutes of Health (grants R01 AR030582 and R01 AG034676).

The authors thank Cindy Crowson and Sara Farmer for permission to include data from their studies comparing hand-checking of residency status to data electronically available through the REP Census. They also thank Barbara J. Balgaard for preparing the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Porta MS. International Epidemiological Association. *A Dictionary of Epidemiology.* 5th ed. Oxford, United Kingdom: Oxford University Press; 2008.
2. Acheson ED. The Oxford record linkage study: a review of the method with some preliminary results. *Proc R Soc Med*. 1964;57(4):269–274.
3. Gill L, Goldacre M, Simmons H, et al. Computerised linking of medical records: methodological guidelines. *J Epidemiol Community Health*. 1993;47(4):316–319.
4. Kendrick S, Clarke J. The Scottish record linkage system. *Health Bull (Edinb)*. 1993;51(2):72–79.
5. Kendrick SW, Douglas MM, Gardner D, et al. Best-link matching of Scottish health data sets. *Methods Inf Med*. 1998;37(1):64–68.
6. Walley T, Mantgani A. The UK General Practice Research Database. *Lancet*. 1997;350(9084):1097–1099.
7. Holman CD, Bass AJ, Rouse IL, et al. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health*. 1999;23(5):453–459.
8. Roos LL, Soodeen R-A, Jebamani L. An information-rich environment: linked-record systems and data quality in Canada. Presented at Statistics Canada Symposium 2001—Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Hull, Canada, October 17–19, 2001.
9. Roos LL, Menec V, Currie RJ. Policy analysis in an information-rich environment. *Soc Sci Med*. 2004;58(11):2231–2241.
10. Conway PH, VanLare JM. Improving access to health care data: the Open Government strategy. *JAMA*. 2010;304(9):1007–1008.
11. Nattinger AB, Pezzin LE, Sparapani RA, et al. Heightened attention to medical privacy: challenges for unbiased sample recruitment and a possible solution. *Am J Epidemiol*. 2010;172(6):637–644.
12. Melton LJ III. History of the Rochester Epidemiology Project. *Mayo Clin Proc*. 1996;71(3):266–274.
13. Kurland LT, Molgaard CA. The patient record in epidemiology. *Sci Am*. 1981;245(4):54–63.
14. Fan Z. Matching character variables by sound: a closer look at Soundex function and Sounds-Like Operator (=*). Presented at SUGI (SAS Users Group International) 29, Montreal, Canada, May 9–12, 2004.
15. US Census Bureau, Population Division. *Plans and Rules for Taking the Census, Residence Rules*. Washington, DC: US GPO; 2000.
16. US Census Bureau. *Census of Population: 1970. Vol. I, Characteristics of the Population, Part 25, Minnesota*. Washington, DC: US GPO; 1973.
17. US Census Bureau. *1980 Census of Population. Volume I: Characteristics of the Population*. Washington, DC: US GPO; 1982.
18. US Census Bureau. *1980 Census of Population: General Social and Economic Characteristics*. Washington, DC: US GPO; 1983.
19. US Census Bureau. *1990 Census of Population and Housing. Summary Tape File 1*. Washington, DC: US GPO; 1990.

20. US Census Bureau. *Census 2000*. Washington, DC: US GPO; 2000.
21. US Census Bureau. *Current Population Reports, Series P-60, No. 175, Poverty in the United States: 1990*. Washington, DC: US GPO; 1991.
22. Siegel JS, Swanson DA, Shryock HS. *The Methods and Materials of Demography.* 2nd ed. Amsterdam, the Netherlands: Elsevier; 2004.
23. Minnesota Department of Health. *Minnesota Vital Statistics Interactive Queries*. St. Paul, MN: Minnesota Department of Health; 2010. (https://pqc.health.state.mn.us/mhsq/frontPage.jsp). (Accessed May 13, 2010).
24. Anderson RN. A method for constructing complete annual U.S. life tables. *Vital Health Stat 2*. 2000;(129):1–28.
25. Melton LJ III. The threat to medical-records research. *N Engl J Med.* 1997;337(20):1466–1470.
26. Yawn BP, Yawn RA, Geier GR, et al. The impact of requiring patient authorization for use of data in medical records research. *J Fam Pract*. 1998;47(5):361–365.
27. Jacobsen SJ, Xia Z, Campion ME, et al. Potential effect of authorization bias on medical record research. *Mayo Clin Proc.* 1999;74(4):330–338.
28. Access to Health Records. *Minn Stat §144.335*. 2005.
29. Dean JM, Vernon DD, Cook L, et al. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. *Ann Emerg Med*. 2001;37(6):616–626.
30. Victor TW, Mera RM. Record linkage of health care insurance claims. *J Am Med Inform Assoc.* 2001;8(3):281–288.
31. Beebe TJ, Ziegenfuss JY, St. Sauver JL, et al. HIPAA authorization and survey nonresponse bias. *Med Care*. 2010. In press.
32. Phillips SJ, Whisnant JP, O'Fallon WM, et al. A community blood pressure survey: Rochester, Minnesota, 1986. *Mayo Clin Proc*. 1988;63(7):691–699.
33. American Academy of Pediatrics. *Recommendations for Preventive Pediatric Health Care*. Elk Grove Village, IL: American Academy of Pediatrics; 2008.
34. Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. 2nd ed. Sudbury, MA: Jones and Bartlett Publishers; 2007.
35. Rocca WA, Grossardt BR, de Andrade M, et al. Survival patterns after oophorectomy in premenopausal women: a population-based cohort study. *Lancet Oncol*. 2006;7(10):821–828.
36. Elbaz A, Bower JH, Peterson BJ, et al. Survival study of Parkinson disease in Olmsted County, Minnesota. *Arch Neurol*. 2003;60(1):91–96.
37. Savica R, Grossardt BR, Carlin JM, et al. Anemia or low hemoglobin levels preceding Parkinson disease: a case-control study. *Neurology.* 2009;73(17):1381–1387.
38. Savica R, Carlin JM, Grossardt BR, et al. Medical records documentation of constipation preceding Parkinson disease: a case-control study. *Neurology.* 2009;73(21):1752–1758.
39. Office for Civil Rights, US Department of Health and Human Services. Standards for privacy of individually identifiable health information. Final rule. *Fed Regist*. 2002;67(157):53181–53273.
40. Council for International Organizations of Medical Sciences, World Health Organization. *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Geneva, Switzerland: World Health Organization; 2002.
41. Council for International Organizations of Medical Sciences, World Health Organization. *International Ethical Guidelines on Epidemiological Studies*. Geneva, Switzerland: World Health Organization; 2009.
42. Hansson MG. Need for a wider view of autonomy in epidemiological research. *BMJ*. 2010;340:c2335.