Title
Leveraging Administrative Data for Program Evaluations

Journal
Evaluation & the Health Professions, 39(2)

ISSN
0163-2787

Authors
Lorden, Andrea L
Radcliff, Tiffany A
Jiang, Luohua
et al.

Publication Date
2016-06-01

DOI
10.1177/0163278714547568

Peer reviewed

Leveraging Administrative Data for Program Evaluations: A Method for Linking Data Sets Without Unique Identifiers

Andrea L. Lorden[1], Tiffany A. Radcliff[1], Luohua Jiang[2], Scott A. Horel[2], Matthew L. Smith[3], Kate Lorig[4], Benjamin L. Howell[5], Nancy Whitelaw[6], and Marcia Ory[7]

Abstract

In community-based wellness programs, Social Security Numbers (SSNs) are rarely collected to encourage participation and protect participant privacy. One measure of program effectiveness includes changes in health care utilization. For the 65 and over population, health care utilization is captured in Medicare administrative claims data. Therefore, methods as described in this article for linking participant information to administrative data are useful for program evaluations where unique identifiers such as SSN are not available. Following fuzzy matching methodologies, participant information from the National Study of the Chronic Disease Self-Management Program was linked to Medicare administrative data. Linking variables included participant name, date of birth, gender, address, and ZIP code. Seventy-eight percent of participants were linked to their Medicare claims data. Linking program participant information to Medicare administrative data where unique identifiers are not available provides researchers with the ability to leverage claims data to better understand program effects.

Background

The Patient Protection and Affordable Care Act (ACA, 2010) calls for the evaluation of community-based initiatives aimed at improving the health and wellness of the Medicare population. The Stanford Chronic Disease Self-Management Program (CDSMP) is one such community-based Initiative (Lorig et al., 1999). A traditional comprehensive evaluation of such an initiative included self-reported survey responses from program participants (Ory, Ahn, Jiang, Lorig, et al., 2013; Ory, Ahn, Jiang, Smith, et al., 2013). However, the need to understand how interventions such as the CDSMP affect health care utilization and cost drove us to link Medicare administrative claims data to the CDSMP survey data.

Of primary importance when incorporating alternative data sources is to assure that both data sources reflect the same person. Linking data sets in community-based programs could be challenging, because limited identifiers are collected to encourage participation and protect participant privacy. Yet, linking data sets without a single unique identifier can be accomplished as long as a combination of variables creates some level of uniqueness (Hammill et al., 2009; Lawson et al., 2012). For example, millions of people share the same date of birth, thousands of people share the same first name, and hundreds to thousands of people live in the same ZIP code. The unique combination of these pieces of information creates a method known as fuzzy matching (Dubois, Prade, & Testemale, 1988).

The science of data linkage started as the need to remove duplicates from large databases emerged (Winkler, 2006). Record linkage commonly describes the matching of records for the same entity without a single unique identifier and has been identified in the literature under object identification, data cleaning, deduplication, approximate matching, or fuzzy matching (Winkler, 2006). For the purposes of this article, we elect to call the collective methods data linkage. Although data linkage software is available, it tends to be inflexible or financially impractical (Bannister, 2004). Therefore, we explored the literature for relevant information to build our own fuzzy matching program.

We found that Fellegi and Sunter are credited with the probability odds ratio that is the mathematical foundation of data linkage science (Winkler, 2006). More practically, we found that successful fuzzy matching models typically follow a process of data cleaning, blocking, measuring distance, and a decision algorithm (Bannister, 2004; Staum, 2007; Winkler, 1999; Winkler, 2006). In particular, Newcombe's concept of blocking transformed the process of fuzzy matching by systematically limiting the comparison or selection group to the most likely group to include the potential match (Winkler, 2006).

As computational resources grew and evolved, a variety of character string comparison tools emerged (Bannister, 2004; McGowan, 2006; SAS Institute, 2013; Staum, 2007;Winkler, 2006). The difference between two character strings is frequently referred to as the distance between strings. Quantifying the distance in SAS# can be accomplished through SOUNDEX, SPEEDIX, and COMPGED commands. The SOUNDEX command is based on the English language where a score is assigned according to the sound of a string. SPEEDIX compares two strings to assign a distance score based on differences, as does COMPGED. A command called COMPCOST allows the user to set weighting parameters for the COMPGED command. These computational advances allow researchers to develop and implement fuzzy matching algorithms specific to the needs of their data sets (Bannister, 2004).

The gold standard in linking personal information across data sets without a single unique identifier is manually reviewing every available record to assure the correct information is linked to the corresponding individual. However, when attempting to link a list of a hundred individuals with those from a list of several thousands, the use of the gold standard quickly becomes impractical. Creating a systematic way to reliably link survey information to Medicare claims data is feasible when a few personal variables are available in both data sets. Fundamental skills to accomplish this include (1) detailed understanding of underlying data issues related to name and address information and (2) the ability to parse complex character strings.

Underlying data issues relevant to our study and the use of Medicare claims data translate to the use of other surveys and claims data linkage. For example, common issues that create difficulties in data linkage include recall bias for survey responses, and data entry–related issues such as typographical errors, or transposed letters, digits or fields of information. Other issues more specific to Medicare claims data that can hamper data linkage include a Medicare beneficiary listing a child or other guardian's address as their primary contact information, even if the child or guardian lived in another community or state. Additionally difficulties arise when a beneficiary relocates, sometimes within the same area, but to a different ZIP code.

Another potential anomaly could arise if changes in ZIP code definitions result in a ZIP change although the beneficiary has not moved. Another important consideration when looking to achieve a high rate of linkages includes the proportion of the survey population thought to be in the claims data. In linking Medicare claims data, the national average enrollment for eligible elders was approximately 95% of individuals aged 65 years and over during 2010 and 2011. We felt it is reasonable to assume that the 5% not enrolled were newly eligible or near 65 years of age. In 2010, this translated to approximately 15% of individuals between ages 65 and 69 years that potentially were not yet enrolled in Medicare. Therefore, it was important to consider the effect of those near 65 years and not yet enrolled may have on the ability to link survey participants to corresponding Medicare claims data.

Issues related to long character strings and numeric data include transposed values, compressed strings, and several character strings that share the same meaning but are not measured as similar by computerized string comparison tools (Bell & Sethi, 2001; Staum, P., 2007). Examples include transposed characters or numbers during data entry, or a forgotten or additional space between names. Much confusion can also arise over abbreviations. Should ‘‘St.’’ be interpreted as street or saint? Acceptable forms of the word street in an address field include St., Str., or street. As you can see, the data anomalies that arise can be numerous. In generating a systematic computerized method to link program participants to Medicare claims data, an understanding of the contents for each linking variable is important.

The motivation for linking administrative data to survey data is to facilitate the comprehensive evaluation of community-based program effects on utilization and cost. The research team had a collective knowledge base regarding data linkage that was sufficient to develop a reliable and systematic approach that was not over burdensome. In sharing this application of our computerized systematic method for linking program participant information to Medicare claims data, other researchers with limited expertise in data linkage can benefit by leveraging large data sets such as Medicare claims information to assess, evaluate, and report the effects of community-based programs on health care utilization and cost.

Method

Data

Chronic disease self-management program participant data. The Stanford CDSMP aids individuals with chronic diseases to develop self-management skills that improve health status through an evidence-based disease prevention model in community-based settings (Lorig et al., 1999; Ory, Ahn, Jiang, Lorig, et al., 2013). With the nationwide implementation of the CDSMP, the National Council on Aging (NCOA), the Stanford Patient Education Research Center, and the Texas A&M Program on Healthy Aging worked collaboratively to recruit participants from 22 implementation sites across 17 states for a longitudinal study of the CDSMP (Ory, Ahn, Jiang, Lorig, et al., 2013; Ory, Ahn, Jiang, Smith, et al., 2013). Written informed consent was obtained from 1,170 individuals to collect and use survey data about health status, health care utilization, and other self-reported health care measures relevant to a participant’s chronic conditions (Ory, Ahn, Jiang, Lorig, et al., 2013). Part of the data included the participant’s name, mailing address, state, ZIP code, birth date, and gender. When the original longitudinal study was extended to include evaluation with Medicare claims data, the need to obtain consent for linking Medicare information to survey data caused the final sample size to dwindle to 267 participants. While this limited the generalizability of some study results, 267 participants were sufficient to develop a method for linking CDSMP program and survey information to Medicare claims data.

Institutional review board (IRB) approval was received from Stanford and Texas A&M IRBs for the program evaluation with the Texas A&M approval allowing for linking of data.

Center for Medicare and Medicaid Services Vital Status File with names and addresses. The Center for Medicare and Medicaid Services maintains a number of administrative claims files that contain utilization information for Medicare beneficiaries including hospital, outpatient, physician, laboratory, and pharmaceutical care. To assure the security and privacy of beneficiary information when it is used for research purposes, each beneficiary's Social Security Number (SSN) is encrypted then removed along with other personal identifiers from administrative records. The encrypted beneficiary identification number (BIN) can then be used to link as necessary across multiple data sets. To enable extraction of CDSMP participant data from the Medicare claims files, the Vital Status File with names and addresses was used to identify the correct BIN for each participant. The variables in the Vital Status File relevant to linking CDSMP participants to their Medicare utilization data included encrypted BIN, beneficiary name, beneficiary date of birth, beneficiary gender, beneficiary mailing address, beneficiary state, and beneficiary ZIP code. Available through the Research Data Assistance Center, we recommend ample lead time to acquire the Vital Status File due to the highly sensitive nature of the data and the rigor of the process associated with obtaining personal health information. Additionally, the Vital Status File is a current file with limited or no ability to access historical cross-sections of the Medicare population. To assure information is sufficiently accurate for linking survey participants, either the Vital Status File should be abstracted as close to the intervention time frame as possible, or the survey participant address information should correspond as closely as possible to the time the Vital Status File was abstracted.

To limit the number of beneficiaries and make comparisons more manageable, we only requested Medicare beneficiary information from ZIP codes where a participant lived or a CDSMP participating site held its classes. Although this narrowed to Medicare beneficiaries living in 170 ZIP codes, the resulting data set represented over 724,000 Medicare beneficiaries, for an average of approximately 4,300 beneficiaries per ZIP code.

Linking CDSMP Participant Information to Medicare Encrypted BIN

To accomplish accurate matching in an efficient manner, a multistage computerized process was developed using SAS 9.3# with a manual review of the linking process for validation purposes. While other data management software can be used, the process described here will specify SAS# commands for interpretation of string parsing techniques.

Following a block-based fuzzy matching algorithm to minimize resource use and maximize the number of CDSMP participants linked to their corresponding Medicare BIN, the multistage process included data standardization, blocking, field score assignment, field score evaluation, and match selection. Figure 1 contains a logic flow diagram of the process described.

Data standardization. Standardization of matching variables required several techniques related to parsing text from complex strings of character data. The first step for all character fields was to capitalize all words and remove excess blank characters before, within, and after the character string. Based upon the assumption that theMedicare beneficiary information was more accurate due to benefit or payment distributions, the CDSMP participant information for the first and last name fields was further parsed to extract additional information such as titles, suffixes, or middle initials. Name variables created for the CDSMP consent file included first name, last name, title, middle name, second or alternate first name, middle initial, and suffix. The Medicare Vital Status File name fields were standardized to first and last name.

Date of birth was parsed into three fields for both files: year, month, and day of birth. The choice to parse the date of birth in this manner provided an opportunity to review the file for data entry errors such as transposed month and day. For the address fields, Perl regular expressions (PRX) commands were used for parsing (McGowan, 2006). Perl regular expressions and the PRX commands allow the analyst to create flexible string structures that parse variations in the address field such as post office box versus a street number and street name or both (McGowan, 2006). The CDSMP data had fewer variations in address fields, while the Medicare Vital Status File mailing address field contained substantial variation in format. The variation in format created limits in the ability to parse out all beneficiary address fields.

Name, date of birth, and address variables in both data sets were populated with valid information. Had any variable been missing or contained invalid data, it would have been excluded from the evaluation.

Blocking. The exact linkage of a CDSMP participants' information to Medicare beneficiary information would theoretically mean comparing the information from each participant's survey to each Medicare beneficiary's information to determine if the two records referred to the same person. Since the time required for a one-to-all comparison could be prohibitive, creating blocks of Medicare beneficiaries on gender and ZIP code enabled a one-to-many comparison. This maximized the opportunity to find the appropriate Medicare beneficiary information in an efficient manner.

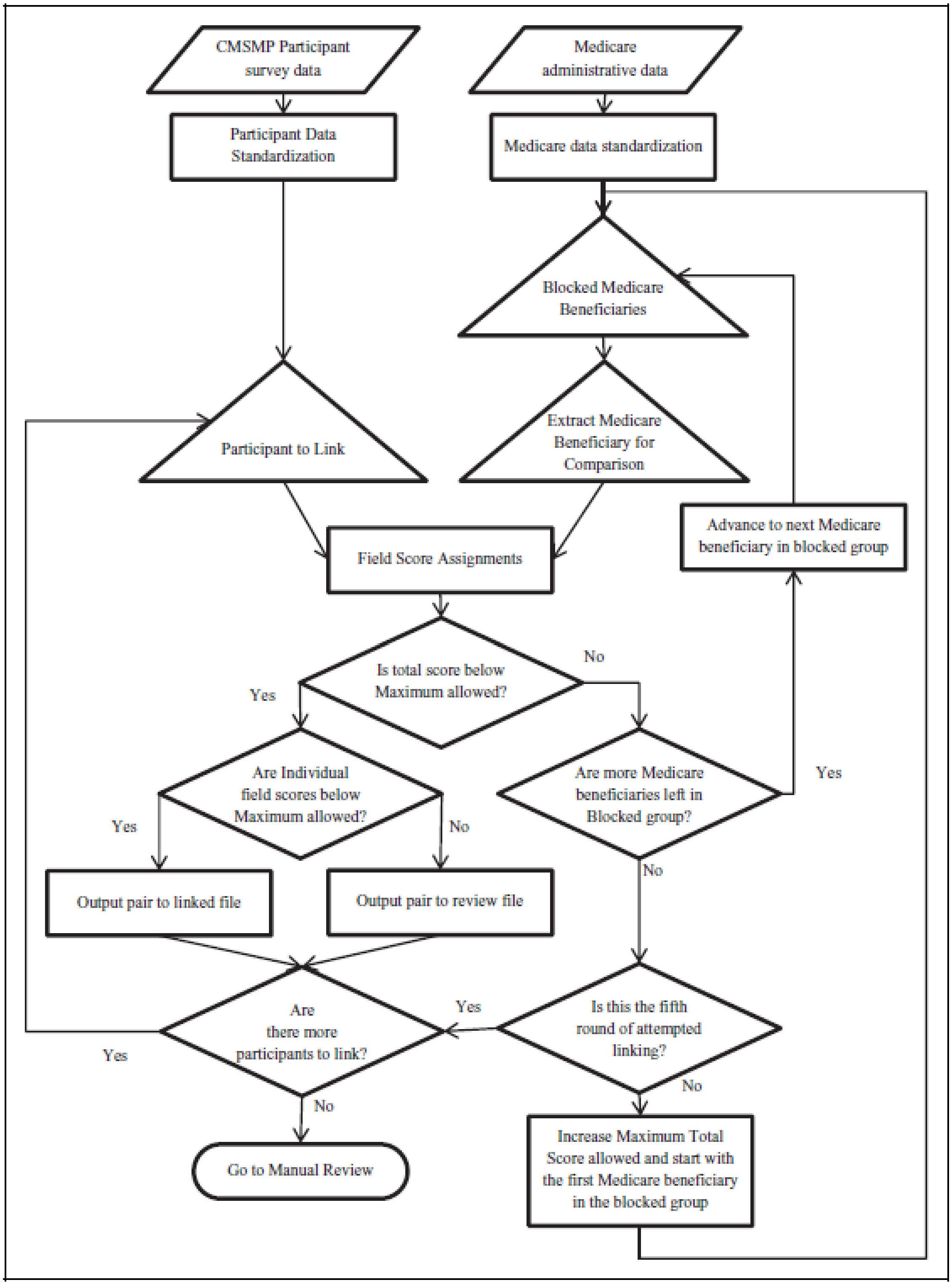

Figure 1. Logic flow for participant linking to Medicare beneficiary identification number.

Field score assignment. To compare the strength of matching for Medicare beneficiary record with each CDSMP survey record containing the same variables, a scoring system was developed. Four scores were generated when first name, last name, date of birth, and address fields were compared. Each score or generalized edit distance (GED) measured by the COMPGED command in SAS#, calculated the difference between two character strings such as CDSMP first name and Medicare beneficiary first name (SAS Institute, 2013). The amount of change the first string must make to match the second string is known as the distance between strings. The COMPGED command uses the second string to recreate the first string one character at a time (SAS Institute, 2013). Depending on the change required, the COMPGED command assigns a GED between 0 and 200 for each character in the string. The more characters requiring change, the larger the distance scored. Therefore, for our purposes, a smaller distance or smaller score translated to a better match between participant and Medicare beneficiary.

To maximize potential matches and account for anomalies in the data, combinations of CDSMP participant survey information were compared to the corresponding Medicare Vital Status record fields. For example, Medicare beneficiary first name was compared to CDSMP participant survey first name, alternate first name, first name plus second first name, and first name plus middle initial. Since the best combination of first name fields could vary from record to record, the minimum distance generated by the combinations was selected to reflect the difference between the Medicare beneficiary information and the CDSMP participant information. Similar comparison processes were followed for last name, date of birth, and address. Finally, a total score was calculated by adding together the GEDs from the four fields.

Score evaluation and match selection. To determine whether the administrative record for a Medicare beneficiary matched the CDSMP participant survey, the total score was examined first. To be considered linkable, the total score had to fall under a predetermined maximum. If the total distance was less than the maximum score allowed, the pair could be linkable. Individual fields were then evaluated and also had predetermined field maximums. If the distances for first name, last name, or date of birth were all below the predetermined maximum field distance of 500 for first and last name and 0 for date of birth, the Medicare/CDSMP pair was sent to an approved linkage file. If any of the first name, last name, or date of birth field distances were greater than the predetermined individual field score maximum, the pair was sent to a file for manual review. Due to substantial variation associated with the larger address field, it was not evaluated individually.

To assure the best possible linkages, the initial maximum total score was set low enough to assure only exact matches would be selected for the approved linkage file. We tested different thresholds for the total score and found an initial threshold of 1,000 for total score allowed for typographic or data entry errors without sacrificing accuracy of the linkage. Four additional rounds of comparisons followed, with each round raising the maximum total score by 200 points to allow for less strict or fuzzier linkages. The predetermined maximum for individual field distances was not changed for subsequent rounds of linking comparisons. After each round, CDSMP participants and Medicare beneficiaries that were considered linked or potentially linkable were removed from the respective data files to avoid duplicate linking.

Linkage review. Linkage review consisted of a line-by-line review of the accepted linkage file and line-by-line review of potential linkages from the review file. A comparison between Medicare beneficiary information and CDMSP participant information of last name, first name, date of birth, and address was evaluated to determine whether linkages were appropriately assigned to linked or review files. Within the review file, all linking fields were compared to identify whether the linking algorithm was effective in linking as the maximum cumulative score was elevated. Since there were 267 CDSMP survey participants, all linkages were reviewed to assess the ability of the linking algorithm to accurately link data and assign suspect matches to review. The small number of CDSMP survey respondents enabled 100% manual review of each step in the process.

Results

Of the 196 linkages assigned to the linked file, all were assessed to be accurate and correct linkages between CDMSP participants and the Medicare vital status beneficiary data (Table 1). Of the 64 linkages assigned for review, 7 were determined to be accurate linkages. The most pairs were assigned to the linked file in the first round of linking, with no pairs assigned in the fifth round. For the review file, the most pairs were assigned for review in the first round and the least in the last round. During the final program debugging and testing process, five additional participants' BINs were linked. As the final form of the computerized method was used, and we confirmed these five participant/Medicare BIN pairs were correct linkages, we included them in the final total linkages of 208, resulting in 78% of program participants being linked to their BINs.

The pairs sent to review were easily identified as linkable or not linkable, and the individual field distances for first name, last name, and date of birth were essential in detecting incorrect linkages during the automated process. Of the participant/Medicare BIN pairs sent to the review file that were confirmed to be linkable, six were reviewed for date of birth discrepancies and one had maiden name and last name that were not separated by a blank character in the Medicare data.

Table 1. Linking Assignment After Each of Five Rounds.

	Assigned to linked file		Assigned to review file			
					Reason for review	
Linking round	*n* (%)	Assessed as correct	*n* (%)	Assessed as correct	Last name	Date of birth
Round 1	186 (95)	186	21 (33)	6	1	5
Round 2	5 (3)	5	10 (16)	1		1
Round 3	3 (2)	3	12 (19)	0		
Round 4	2 (1)	2	12 (19)	0		
Round 5	0 (0)	0	9 (14)	0		

Discussion

The ACA mandate to assess the effect of community-based prevention and wellness programs on the Medicare population will drive the need to measure changes in health care utilization and cost associated with the implementation of such programs. For Medicare beneficiaries, Medicare claims

information is a valuable repository of comprehensive health care utilization and cost information. In the past, one barrier to using Medicare claims data in community-based program evaluations has been the lack of a method to reliably link Medicare beneficiary information to program participant information when a unique identifier such as an SSN is not present in both data sets.

However, as demonstrated in this article, linking information from program participants to their Medicare beneficiary information can be accomplished when a unique identifier such as SSN is not available. When name, address, and date of birth are available, a computerized systematic approach can be applied with great success. It is important to note that this process could be further refined. For example, once a pair had a distance score below the maximum allowed, both were removed from further consideration. Removal eliminates the chance of duplicate matching. However, we found that removal before comparing to all Medicare beneficiaries in the blocked group left at least five participants unlinked that had linked in the final test of the program development. While coding modifications such as assignment of linking scores for all Medicare beneficiaries in the block then selecting the lowest score are possible, the question remains as to the benefit gained. Much depends upon the number of beneficiaries in the block, and the likelihood the survey participant is enrolled in Medicare.

Some might say a 78% linkage rate seems low. However, we can think of at least four possible reasons for a match rate of less than 100%. First, noting that 95% of individuals aged 65 and older enrolled in Medicare in 2011, we might assume that the majority of unenrolled individuals are newly eligible. This translates to 15% of individuals between ages 65 and 69 who may not be enrolled in Medicare. Our study included 65 individuals less than 69 years. One would therefore expect at least 10 of those individuals in the CDSMP sample were not enrolled in Medicare.

Second, due to extremely large amount of data in the Medicare claims files, we elected to use ZIP code as our initial blocking variable as both data sets contained ZIP code. However, not all beneficiaries' address information matches the beneficiary's residential address. Some beneficiaries use an authorized representative's information for correspondence. We are uncertain the number of beneficiaries who elect to list an authorized representative's address instead of the beneficiaries' residential address.

Third, also related to blocking on ZIP code, certain individuals may be excluded from the initial data pull due to relocation. Individuals who move a short distance effectively live in the same area but may have moved into a different ZIP code. In urban and suburban areas where the majority of people live, this is more likely to occur. Since the Vital Status File is accurate at the time it is pulled, it is important for researchers to gain access to the Vital Status File as close as possible to the time survey participant address information is collected.

Although blocking on ZIP code presented several challenges in linking survey participants to their Medicare beneficiary data, other available variables such as year of birth or state did little to reduce the pool of potential matches in a meaningful way. Even when limited to the same ZIP code, the number of beneficiaries available for comparison was approximately 4,300. Since the purpose of a blocking variable is to limit the population of potential matches to a group of the most likely matches, we felt ZIP code was the best way to limit the number of Medicare beneficiaries while ensuring the greatest likelihood of including the potential match in the sample.

Finally, the reality of working with administrative data is an expectation for some imperfection and inaccuracies. Even in our review of exact matches, almost none had scores of zero. While the systematic approach attempts to account for spelling and formatting errors in the data, it is unlikely to capture all such anomalies.

Since we were able to link 78% of the CDSMP survey participants to their corresponding Medicare BIN, we question the value in allocating additional computational resources that would likely produce few additional linkages. For example, the time required to sort a large data set can be substantial and varies by the program, number of data elements, and commands used. These computational issues become relevant when linking a large number of program participants to a large beneficiary file such as those related to Medicare. Potential benefit may be realized from the additional time and coding efforts when the sample size needed for a study is closer to the number of participants available for linking, or the simplified linking process creates systematic bias in the sample of beneficiaries. Even in this case, consideration should be given to the proportion of the survey participant population likely to be in the claims data.

Also, issues related to parsing data are important to consider. For example, during data standardization, substantial variation in the formatting of the Medicare beneficiary mailing address was found. This proved difficult to parse resulting in loss of comparable data for some

Medicare beneficiaries. The variation in the address field also introduced scoring challenges. In establishing an appropriate scoring method, we were required to review the linking sequence several times testing different strategies. Ultimately, a combination of total score and individual field distances resulted in the iterative process that allowed the maximum opportunity to find appropriate linkages.

Finally, although a larger number of consented participants would have been desirable for other analysis, it is possible to develop a reliable linking program with a sample of approximately 100 to 250 participants. Initial testing with a sample of participants illuminates the most common data anomalies related to parsing names and addresses. It also allows for review of each phase to assess the accuracy of linkages. While almost no CDSMP participant/Medicare BIN pairs had a total distance of zero when reviewed, accuracy of the linkage was unquestionable for a large proportion of the linkages.

The manual review of linkages depends on the purpose of the analysis and the size of the linked survey participants. In the case where clinical accuracy is needed, a 100% manual review would be necessary to assure patient safety. For our purposes and most other researchers performing secondary analysis, a 100% validation of linkages is not necessary. We performed a 100% validation to assure the method was accurately linking survey participants to their corresponding Medicare BIN, and because our sample size was small. If the number of linkages increased beyond the feasibility of manual review, we suggest a two-stage process. First, a 100% manual review of the process for approximately 200 individuals were performed. This allows researchers to identify systematic flaws in the linkage process. Second, once the entire participant population has been linked, we suggest a random sample of linkages be validated through manual review. Sample size for manual review should vary according to the number of survey participants and in accordance with measures of statistical assurance.

Success or refinement of this approach in other research will depend upon the quality and completeness of the survey data research teams possess. From our experience, development of a computerized systematic method for linking survey data to health utilization data was an efficient and effective use of resources. In addition to eliminating recall bias related to health care utilization, the Medicare administrative data allowed us to examine costs related to changes in utilization. By incorporating the cost analyses, researchers can achieve a more complete look at the effect of community-based initiatives such as the CDSMP.

Authors' Note

The contents of this article are solely the responsibility of the authors and do not necessarily reflect the official views of the the National Council on Aging (NCOA) or the Centers for Medicare and Medicaid Services.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Kate Lorig receives royalties from Bull Publisher.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This publication was made possible by funding through Contract: HHSM-500-2011-00088C to the National Council on Aging (NCOA).

References

Bannister, W. (2004). *Matching without unique identifiers—Fuzzy matching applications with SAS 1 software.* Cary, NC: SAS1 Institute.

Bell, G. B., & Sethi, A. (2001). Matching records in a national medical patient index. *Communications of the ACM*, 44, 83–88.

Dubois, D., Prade, H., & Testemale, C. (1988). Weighted fuzzy pattern matching. Fuzzy Sets and Systems, 28, 313–331.

Hammill, B. G., Hernandez, A. F., Peterson, E. D., Fonarow, G. C., Schulman, K. A., & Curtis, L. H. (2009). Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *American Heart Journal*, 157, 995–1000.

Lawson, E. H., Ko, C. Y., Louie, R., Han, L., Rapp, M., & Zingmond, D. S. (2012). Linkage of a clinical surgical registry with Medicare inpatient claims data using indirect identifiers. *Surgery*, 153, 423–430.

Lorig, K. R., Sobel, D. S., Stewart, A. L., Brown, B.W., Jr, Bandura, A., Ritter, P., . . . Holman, H. R. (1999). Evidence suggesting that a chronic disease self-management program can improve health status while reducing hospitalization: A randomized trial. *Medical Care*, 37, 5–14.

McGowan, K. (2006). *Perl regular expressions in SAS* (No. AP09-2006). Durham, NC: Constella Group.

Ory, M. G., Ahn, S., Jiang, L., Lorig, K., Ritter, P., Laurent, D. D., . . . Smith, M. L. (2013). National study of chronic disease self-management six-month outcome findings. *Journal of Aging and Health*, 25, 1258–1274.

Ory, M. G., Ahn, S., Jiang, L., Smith, M. L., Ritter, P. L.,Whitelaw, N., & Lorig, K.(2013). Successes of a national study of the chronic disease self-management program: Meeting the triple aim of health care reform. *Medical Care*, 51, 992–998.

Patient protection and Affordable Care Act, 42 U.S.C.300u-14 (2010).

SAS Institute, I. (2013). COMPGED function. Retrieved August 10, 2012, from http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a002206133.htm

Staum, P. (2007). Fuzzy matching using the COMPGED function. *Proceedings from the Northeast SAS Users Group*. Retrieved from http://www.nesug.org/Proceedings/nesug07/ap/ap23.pdf

Winkler, W. E. (1999). The state of record linkage and current research problems. *Statistical Research Division, US Census Bureau* (Research Report). Washington, DC. Retrieved from http://www.census.gov/srd/papers/pdf/rr99-04.pdf

Winkler, W. E. (2006). Overview of record linkage and current research directions. Research report series (Statistics #2006-2). *Statistical Research Division*, *US Census Bureau,* Washington, DC.

[1]Department of Health Policy and Management, School of Public Health, Texas A&M Health Science Center, College Station, TX, USA

[2]Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M Health Science Center, College Station, TX, USA

[3]Department of Health Promotion and Behavior, College of Public Health, University of Georgia, Athens, GA, USA

[4]Department of Medicine/Immunology & Rheumatology, Stanford University, Stanford, CA, USA

[5]Centers for Medicare and Medicaid Innovation and Strategic Planning, Centers for Medicare and Medicaid Services, Baltimore, MD, USA

[6]Center for Healthy Aging, National Council on Aging, Washington, DC, USA

[7]Department of Health Promotion and Community Health Sciences, School of Public Health, Texas A&M Health Science Center, College Station, TX, USA

Corresponding Author:

Andrea L. Lorden, Department of Health Policy and Management, School of Public Health, Texas A&M Health Science Center, 1266 TAMU, College Station, TX 77843, USA. Email: lorden@sph.tamhsc.edu