SAMPLE FRAME DEDUPLICATION IN THE WORLD TRADE CENTER HEALTH REGISTRY: MINIMIZING OVERCOVERAGE AND COST[1]

Joe Murphy, Paul Pulliam, and Randolph Lucas: RTI International

KEY WORDS: deduplication, sample building, coverage

Summary

The World Trade Center (WTC) Health Registry is designed to assess the health effects of the WTC disaster of September 11, 2001. It will follow those exposed to dust and fumes on 9/11 and in the ensuing weeks as the fires burned. Persons who may enroll in the Registry include those who were in lower Manhattan on 9/11; residents and school children south of Canal Street; and persons involved in rescue, recovery, or clean-up at the WTC site or Staten Island Recovery Operations between September 11, 2001 and June 30, 2002. The sample frame includes people from several hundred potentially overlapping list sources of individuals. To avoid overcoverage, list entries are systematically deduplicated using an algorithm to identify likely duplicates. Indeterminates are manually reviewed to assure that the same individual was not included in the sample more than once. This paper describes the process of deduplication and assesses the resulting increase in quality and reduction in cost and respondent burden.

Introduction

The World Trade Center Health Registry is a database for tracking persons who were exposed to the WTC disaster on September 11, 2001. It is being conducted by the New York City Department of Health and Mental Hygiene (NYCDOHMH) and the Agency for Toxic Substances and Disease Registry (ATSDR). Data collection is being conducted by RTI International. The primary mode of collection is telephone interviewing.

The purpose of the Registry is to evaluate potential short and long term physical and mental health effects of the disaster. People are assigned to defined exposed populations with the expectation that after baseline recruitment they would be followed for up to twenty years. Exposed groups are broadly defined based on proximity to the WTC disaster and its aftermath. The Registry includes persons who were downtown (South of Chambers Street in Manhattan) on the morning of September 11, 2001 and who may have been present during the collapse of the two towers and the subsequent dust/debris cloud; rescue, recovery, and clean-up workers who worked on the pile or its vicinity in the days and weeks following the disaster; residents who lived in the surrounding area around the WTC disaster site (South of Canal Street in Manhattan); and school children and staff in schools in downtown Manhattan (South of Canal Street).

The broadly defined exposure groups were separated into high priority and low priority exposure populations. High priority exposed persons are defined by those who have relatively high levels of exposure, estimable denominator, and a greater chance of being located; this Group is referred to as Group 1. Group 2 includes persons who are have less acute exposures than those in Group 1, such as persons who were on the street south of Chambers on September 11, 2001 but not in one of the 35 damaged or destroyed buildings nearest to the WTC site. People who were in any one of the 35 damaged or destroyed buildings prior to or at the time of the attack were designated Group 1, a subset of people who were south of Chambers Street on September 11, 2001; rescue, recovery, or clean-up workers who can be located by contacting their employer or organization are Group 1 while volunteer workers are Group 2; residents who lived South of Chambers Street or closer to WTC site are Group 1, while residents between Canal and Chambers are Group 2; all school children and staff in schools are Group 1.

Resources were allocated to a list building and active tracing methodology for Group 1 exposure populations. Lists of potentially eligible residents were available for purchase, but the remaining sample of school children, rescue, recovery, and cleanup workers, and occupants of buildings on 9/11 is being built by requesting information directly from appropriate entities. Representatives of these entities are asked to provide a list of all potentially eligible persons, including current contact information. Over 250 lists have been obtained as of July 19, 2004. In addition, any person who thought they may be eligible (that is, Group 1 or Group 2) could self-identify by pre-registering with their contact information on a web site or calling a toll free number to be interviewed. Outreach and media campaigns were mounted during the enrollment phase of the WTC Health Registry to encourage cooperation among those called for their interview and to promote self-identification for enrollment. With hundreds of

1 Data presented in this paper are not final. These data are current as of July 19, 2004 and are subject to change since data collection was still in process at the time this paper was written.

overlapping sample sources, the likelihood of obtaining duplicate sample members is high. This paper outlines the approach taken to minimize the impact of duplicate cases on data quality, costs, and respondent burden.

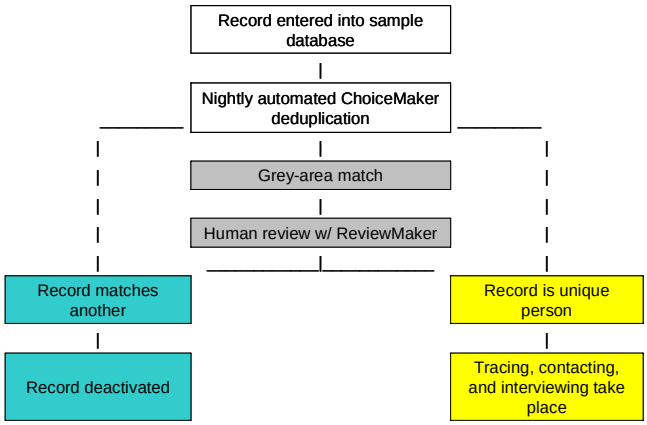
Methods

As of July 18, 2004, over 180,000 unique persons have been identified for the WTC Health Registry and over 48,000 have completed a 30-minute interview. A major concern of working with such a large a database of names from such a small geographic area is sample frame duplication. Kish (1965) describes a sample frame as "perfect if every element appears on the list separately, once, only once, and nothing else appears on the list." Coverage error occurs when some persons are omitted from the list or frame used to identify members of the study population (Groves, 1989). It also occurs when the same person is included in the sample frame more than once or when the person is included in error. Because of the multitude of sample types and list sources, it was expected that many sample members would be included in the frame more than once. For instance, if a person was identified as a resident living south of Chambers Street on the purchased residential list file and was also identified by an employer in one of the 35 damaged or destroyed buildings around the WTC complex, that person's name would appear twice on the sample frame. Similarly, if a person was identified by a rescue/recovery organization as having worked at the WTC site between September 11, 2001 and June 30, 2002 and that person also self-registered via the WTC Health Registry web site, he or she would be listed twice in the sample frame. Deduplication is necessary to avoid overcoverage and to reduce the cost and respondent burden associated with treating the same case as more than one case. (Murphy, Brackbill & Thalji, 2004).

Since potential respondents may be identified on multiple lists, including web self-registration and by inbound calls, accounting for duplicates is necessary to reduce respondent burden that would result from contacting the same person multiple times and to minimize costs associated with unnecessary multiple contacts. In addition, it is necessary to perform deduplication to assure data quality. For persons identified but not interviewed, there would be no feasible way to determine whether records were duplicated without an automated approach. As a result, many more incomplete interviews would be recorded and coverage error associated with duplication would increase.

The deduplication process utilized for the WTC Health Registry compares locator (name, address, phone, Social Security Number) and demographic information between cases in the sample database. ChoiceMaker Technologies developed a batch matching program to evaluate every new record in the sample database relative to existing records in the table. In this process, locator and demographic fields are parsed and standardized. Once the fields are standardized, key matching fields (name, birth date, SSN, etc.) are used to identify a set of cases that are potential matches to the new record. For each of the cases in the set, clues are applied to determine two probabilities – match and differ. New cases with high match probabilities relative to an existing case are determined to be duplicates. New cases with high differ probabilities relative to all existing cases are determined to be unique. The thresholds for match and differ determination are parameters that can be modified as necessary. Figure 1 illustrates the automated process.

Figure 1. Deduplication Process



Records that have match probabilities between the match and differ thresholds ("grey-area matches") are marked for human review. The ReviewMaker tool is used by the reviewers to work the cases marked for human review. It allows the reviewer to see all cases in a potential match compound and allows them to make hard match/differ decisions for the new case relative to each existing case in the compound.

It is important to demonstrate the benefits of this approach to deduplication for reference in other large-scale surveys involving a sample frame constructed from many overlapping list sources. In the next section, we consider questions related to data quality, cost, and respondent burden associated with deduplication.

Analysis

Without deduplication, the WTCHR sample database would include more than 207,000 records. Through the deduplication process, 17,751 cases were determined to have duplicate values in the sample

database. Eliminating the 20,449 records representing the 17,751 duplicate cases reduced the number of cases in the database by more 9.5%.[2] The majority of the 17,751 duplicate cases appeared twice in the sample database prior to deduplication, but another 1.3% appeared three or more times. Table 1 shows the distribution of instances cases appeared in the sample database prior to deduplication.

Table 1. Number of Cases by Number of Records in the WTCHR Sample Database

Records per Case	Cases	Percent of Cases	Records	Percent of Records
1	169,460	90.5%	169,460	81.6%
2	15,316	8.2%	30,632	14.8%
3	2,003	1.1%	6,009	2.9%
4+	432	0.2%	1,559	0.8%
Total	*187,211*	*100.0%*	*207,660*	*100.0%*
Cases Included More than Once	17,751	9.5%	--	--
Extra Records Due to Duplicate Cases	--	--	20,449	9.8%

Deduplication occurred at different rates based on sample priority group (Group 1 or 2) and sample type (resident, student/school staff, building occupant, rescue/recovery worker). Table 2 presents the distribution of sample cases by group and type before and after deduplication.[3] In general, deduplication improved data quality by assuring that the sample frame did not include extra cases in error. For instance, without deduplication, the resident sample type would have accounted for almost 21% of the sample frame. The resident share of the sample frame after deduplication is significantly lower at about 19%.

Table 2. Sample Frame Group and Type Distribution Before and After Deduplication

Before Deduplication				
Sample Type	**Group 1**	**Group 2**	**Total**	**Percent**
Residents	17,892	24,651	42,543	20.5%[a]
Students / School Staff	2,415	0	2,415	1.2%
Building Occupants	94,280	16,904	111,184	53.5%[b]
Rescue / Recovery Workers	34,653	4,811	39,464	19.0%[a]
Unknown (Mostly Ineligible)	0	12,054	12,054	5.8%[b]
Total	149,240	58,420	207,660	
Percent	71.9%[b]	28.1%[a]		100.0%
After Deduplication				
Sample Type	**Group 1**	**Group 2**	**Total**	**Percent**
Residents	15,028	21,165	36,193	19.3%
Students / School Staff	2,219	0	2,219	1.2%
Building Occupants	86,863	16,305	103,168	55.1%
Rescue / Recovery Workers	29,218	4,794	34,012	18.2%
Unknown (Mostly Ineligible)	0	11,619	11,619	6.2%
Total	133,328	53,883	187,211	
Percent	71.2%	28.8%		100.0%

[a]Significantly higher than after deduplication at $p<.05$
[b]Significantly lower than after deduplication at $p<.05$

Changes in sample composition before and after deduplication can also be tracked by geography and sample demographics to determine whether frame error would have occurred had deduplication not taken place.

Deduplication is also a potential cost-saving method since it can identify and eliminate erroneous sample records before they are actively worked in data collection. For instance, without deduplication, it may not be known that a single person is in the sample database twice. If both records in the database are put

[2] A case here is defined as a record or set of records defining a unique individual. A record is any row in the sample database. Before deduplication, one case may have several records. After deduplication, there is only one record per case.

[3] Deduplication did not occur at one time but was a nightly process.

through locating and contact processes only to find out that they represent a single individual when that person is contacted a second time, money is used less efficiently than it would be otherwise, and response burden is increased.

To determine whether the WTC Health Registry deduplication process successfully reduced the costs associated with erroneously contacting the same person multiple times, we compare the cost per case of deduplication to the cost per case of tracing and locating. Because cost structures differ based on when and by whom data are collected, we present cost here in relative terms. Let the cost of deduplication per case equal D. D is calculated by dividing the total cost of deduplication by the number of records successfully deduplicated. Let the cost of tracing and contacting a case equal C. If C is greater than D, then cost savings are realized by using deduplication, since tracing and contacting are not needed on the deduplicated cases. For the WTC Health Registry, C is 1.16 times greater than D. This means that the cost per case for deduplicated cases under the method we employed is 16% less than it would have been had we actively attempted to trace and contact the cases. Applying these costs to the overall number of valid and duplicate cases, we can see what our costs would have been with and without deduplication. Here D=1 and C=1.16:

Without deduplication: 207,660 cases * C = 240,886.

With deduplication: (187,211 cases * C) + (20,449 duplicates * D) = 231,132.

Dividing the total with deduplication by the total without deduplication, we find that actual costs for tracing and locating were 96% what they would have been without deduplication. In other words, by using deduplication, 4% of the tracing and contacting budget was saved.

Finally, we consider the effect of deduplication on respondent burden. It is important for any survey to minimize the burden placed on respondents by avoiding unnecessary contact and engagement wherever possible. Deduplication makes it possible to avoid unnecessary contact by determining in advance which records should not be pursued because they do not represent unique cases. Without deduplication, a single individual may be contacted two or more times to complete an interview, even though he or she has already completed an interview. When the respondent is reached to complete the interview again, he or she must explain that the interview has already been completed. This puts undue burden on the respondent in terms of time, and possible annoyance. Recontacting a respondent in error may also reduce the perceived integrity of the survey organization.

While we have no data to measure the amount of annoyance or other qualitative effects of unnecessary contacts on the WTC Health Registry, we can calculate the approximate reduction in burden time due to deduplication. As mentioned previously, over 48,000 interviews have been completed to date. Each interview takes 28.6 minutes on average. This equates to more than 22,000 hours of contact with respondents. In addition to the completed interviews, contact has been made with ineligible or noncompliant sample members. To date, more than 10,000 sample members have fallen into this category. The amount of time spent on the phone with these respondents is only about 3 minutes, on average. This equates to about 500 hours of contact with respondents.

To determine the proportion of time saved by not contacting duplicate records, we apply the approximate amount of time that would have been spent on the phone with these individuals, had deduplication not taken place. We estimate that these calls would average 30 seconds in length. This includes the time for the respondent to answer the phone, the interviewer to read the introduction script, and the respondent to explain that he or she has already completed an interview.

Assuming the proportion of individuals contacted to date is equal to the proportion of duplicates that would have been contacted without deduplication, we estimate that approximately 6,339 records would have been contacted in error. At 30 seconds a call, this comes to about 53 hours of unnecessary contact with duplicate cases. While this equals less than one percent of the total amount of respondent burden time, it is likely that the unnecessary contact would cause undue annoyance for thousands of people. This could also translate into a decreased in perceived integrity for the Registry sponsors and data collectors.

Conclusion

Use of deduplication matching algorithms across multiple list sources was successful for the World Trade Center Health Registry across three dimensions: data quality, cost, and respondent burden. Deduplication improved data quality by assuring that the sample frame did not include extra cases in error, as shown by the changes in the Registry's sample composition before and after deduplication. Deduplication was cost effective in that with use of matching algorithms the cost per case for de-duplicated cases was 16% less than it would have been without use of these methods; a total of 4% of the tracing and contacting budget was saved. Deduplication minimized respondent burden in the time that was avoided on the telephone with individuals who

would have been recontacted without use of this method.

Deduplication methods are applicable to other registries as well as probability-based samples. Environmental exposure registries often face the challenge of compiling a cohort exposed to agents that were present in various media across years or decades. Because these cohorts disperse across time, multiple data sources are often used to identify potential registrants. As with the World Trade Center Health Registry, use of various list sources can result in overcoverage error that one can attempt to minimize and measure through deduplication techniques. Probability-based samples that use multiple, potentially overlapping sample frames can benefit from deduplication methods to measure their potential overcoverage as well, and larger samples are most likely to find this cost effective. Finally, researchers responding to imperatives to maximize the analytic utility of existing databases can benefit from deduplication methods. The New York City Department of Mental Health and Hygiene will use these techniques in matching the World Trade Center Health Registry with existing immunization and tuberculosis registries. Environmental health researchers attempting to integrate databases of hazards, exposures, and health effects face the same methodological challenge of measuring error while minimizing costs. Formal deduplication methods can help to improve quality and minimize respondent burden and costs for all of these efforts.

References

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley and Sons, Inc.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons, Inc.

Murphy, J., Brackbill, R. and Thalji, L. (2004). Coverage Issues in the World Trade Center Health Registry. Paper presented at the City Futures Conference: http://www.uic.edu/cuppa/cityfutures/papers/webpapers/cityfuturespapers/session6_3/6_3coverageissues.pdf