STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2806

Measuring and maximizing coverage in the World Trade Center Health Registry

Joe Murphy[1], Robert M. Brackbill[2,*,†], Lisa Thalji[1], Melissa Dolan[1], Paul Pulliam[1] and Deborah J. Walker[3]

[1] *RTI International,*[‡] *230 W. Monroe, Suite 2100, Chicago, IL 60606, U.S.A.*
[2] *Federal Agency for Toxic Substances and Disease Registry, U.S.A.*
[3] *New York City Department of Health and Mental Hygiene, U.S.A.*

SUMMARY

The World Trade Center Health Registry (WTCHR) is a database for following people who were exposed to the disaster of 11 September 2001. Hundreds of thousands of people were exposed to the immense cloud of dust and debris, the indoor dust, the fumes from persistent fires, and the mental trauma of the terrorist attacks on the WTC on 9/11. The purpose of the WTCHR is to evaluate the potential short- and long-term physical and mental health effects of the disaster. The definitions of the exposed groups are broad and defined based on an understanding of which groups had the highest exposures to the WTC disaster and its aftermath. The four exposure groups include rescue and recovery workers, residents, students and school staff, and building occupants and passersby in Lower Manhattan.

While one goal of the WTCHR was to maximize coverage overall and for each exposure group, another was to ensure equal representation within exposure groups. Because of the multiple sample types pursued, several approaches were required to determine eligibility. Estimates of the number of eligible persons in each of the exposed populations were based on the best available information including Census, entity-specific employment figures, and public and private school enrollment data, among other publicly available sources. To address issues of undercoverage and overcoverage a variety of methods were assessed or applied, including a capture–recapture analyses test of overlapping sample building list sources and automated deduplication of sample records.

Estimates of the true eligible population indicate that over 400 000 unique individuals were eligible for the baseline health survey. Interviewer-administered surveys were completed with more than 71 000 persons, resulting in an overall enrollment rate of approximately 17 per cent. Coverage was highest among rescue and recovery workers, followed by residents, students and school staff, and building occupants. Both the accuracy of coverage estimates and the raw number and representativeness of enrollees were maximized by our approach to coverage. In designing a registry which relies on multiple pathways and

[*] Correspondence to: Robert M. Brackbill, Federal Agency for Toxic Substances and Disease Registry, U.S.A.
[†] E-mail: rob1@cdc.gov
[‡] RTI International is a trade name of Research Triangle Institute.

Contract/grant sponsor: Federal Agency for Toxic Substances and Disease Registry (ATSDR)

Received 16 October 2006
Accepted 16 November 2006

sources of data to build the sample, it is important to develop a comprehensive approach that considers all sources of error and minimizes bias that may be introduced through the methodology. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: registry; World Trade Center; coverage; population estimation

INTRODUCTION

The World Trade Center Health Registry (WTCHR) is a database for following people who were exposed to the disaster of 11 September 2001 (9/11). Hundreds of thousands of people were exposed to the immense cloud of dust and debris, the indoor dust, the fumes from persistent fires, and the mental trauma of the terrorist attacks on the WTC on 9/11. The purpose of the WTCHR is to evaluate the potential short- and long-term physical and mental health effects of the disaster. It was conceived as an imperative public health response to document and assess the potential impact on physical and mental health resulting from the WTC disaster on large and diverse populations. It is a collaborative scientific effort by the New York City Department of Health and Mental Hygiene (NYCDOHMH), Agency for Toxic Substances and Disease Registry (ATSDR), and external scientific and community partners. Funding has been provided by the Federal Emergency Management Agency (FEMA). Registry building and baseline data collection activities were conducted by RTI International.

The objectives of the WTCHR are to collect information about physical and mental health effects across a wide range of exposures; to provide data on potential health effects identified by the WTCHR for more in depth follow-up studies; to provide a means for conducting long-term follow-up of a large group of exposed persons; and, to provide data that may assist in the development of screening and intervention programmes. Akin to a longitudinal cohort study, people after baseline recruitment will be followed for up to 20 years. The goals of this methodology manuscript are to describe the steps deployed to establish a population-based cohort; detail the methods used to measure coverage and response rates; and, discuss the strategies used to maximize coverage in the WTCHR for selected groups of registrants.

The definitions of the exposed groups are broad and based on an understanding of which groups had the highest exposures to the WTC disaster and its aftermath. Exposure groups were selected for the Registry based on the following criteria: exposure to the actual event; exposure to the immediate aftermath of the attack; ongoing exposures related to rescue, recovery and clean-up of the WTC site; or living, working, or attending school in the lower Manhattan area. The WTCHR targeted the following populations that met the exposure criteria described above:

1. *Workers*: *workers and volunteers involved in rescue*, *recovery*, *clean-up*, *or other disaster-related activities at the WTC site and/or at the Staten Island recovery operations or on transport barges for at least one shift anytime from 11 September 2001 to 30 June 2002*:

   Workers at the WTC disaster site may have been exposed to potentially toxic contaminants at levels sometimes exceeding regulatory limits and with unknown synergistic effects; faced risks of physical injury from falls, burns and other safety hazards; and have experienced mental trauma from the loss of colleagues and other disturbing experiences. At the Staten Island recovery operations or on the barges transporting materials, workers may have also been exposed to potentially toxic contaminants at levels exceeding regulatory limits in the course of their work of sifting or moving materials, including remains, from the WTC disaster site.

2. *Residents*: *persons whose primary residence was south of Canal Street on 11 September 2001*:
   Residents, including adults and children, of lower Manhattan on 11 September 2001 whose primary residence was close to the disaster site may have had an increased risk of exposure to potentially toxic contaminants if they were at home at any time between 11 September and 31 December 2001. Many residents were also displaced from their homes, had concerns about toxic exposures, and had potential exposure to physical injury or psychological impact and/or rescue and clean-up efforts.
3. *Students and Staff*: *students who were enrolled in a nursery school/daycare*, *elementary*, *middle*, *or high school south of Canal Street on 11 September 2001 and staff persons employed in a nursery school/daycare*, *elementary*, *middle*, *or high school south of Canal Street on 11 September 2001*:
   Many students and school staff associated with schools (pre-K through 12) south of Canal Street were near a cloud of dust and debris, had to be evacuated, were exposed to the fumes and particulate material of the fire, or were potentially exposed to physical injury or psychological impact.
4. *Occupants and Passersby*: *persons present south of Chambers Street in Manhattan on 11 September 2001 any time between the first plane impact and noon* (*this includes persons who were in collapsed or damaged buildings*, *people in other buildings*, *and people in transit or outdoors*):
   Occupants in buildings and people in transit or pedestrians in and near the WTC on the morning of 11 September 2001 had a high risk of potential or actual injury from burns, falling or exploding debris, glass, and trampling; exposure to a variety of potentially toxic contaminants; and potential exposure to traumatic events.

The four exposure groups were further divided into two groups. Registrants were assigned to one of these two groups based on self-reported responses to the baseline survey. Group 1 is considered higher priority because they are among those considered to be the most highly exposed to the environmental effects of the disaster and generally consist of those individuals in greatest physical proximity to the WTC site either on 11 September 2001 or during the subsequent clean-up. Group 1 consists of all workers and volunteers involved in rescue, recovery, or clean-up at the WTC site, the Staten Island recovery operations, or on the barges for at least one shift any time between 11 September 2001 and 30 June 2002; people whose primary residence was south of Chambers Street on 9/11; students and school staff enrolled or employed in schools or day cares south of Canal Street on 9/11; and persons working in one of the 35 buildings or the three structures damaged or destroyed on 11 September 2001.

Group 2 consists of people whose primary residence was south of Canal Street but on or north of Chambers Street on 9/11; and occupants, employees, visitors, passersby and others who were in buildings or on the subway south of Chambers Street on 9/11 other than the those who were in the 35 damaged or destroyed buildings or the three damaged structures on 9/11. Figure 1 shows a map of lower Manhattan with the Chambers Street and Canal Street boundaries identified in relation to the WTC site.

The WTCHR is a non-probability based sample. It was compiled from various sources, including purchased or acquired lists of likely eligible registrants (‘pre-registrants’), self-identification through a toll-free number, and self-identification through the project web site. It is a consecutive sample, that is, a strict version of convenience sampling where every available subject is selected [1]. Statistical principles can be applied to validate coverage and representativeness. The WTCHR

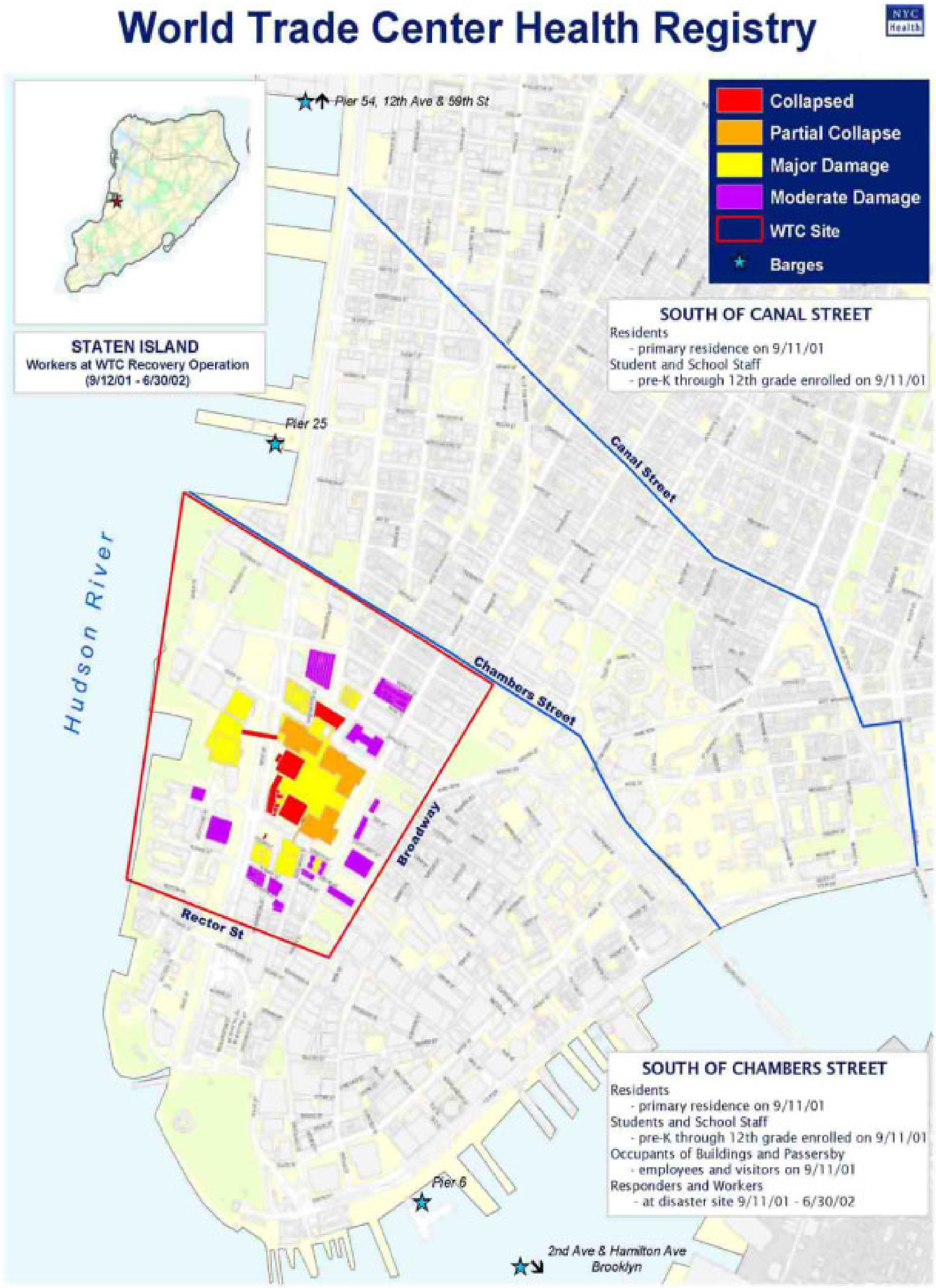
World Trade Center Health Registry
NYC Health
Collapsed
Partial Collapse
Major Damage
Moderate Damage
WTC Site
Barges
Pier 54, 12th Ave & 59th St
STATEN ISLAND
Workers at WTC Recovery Operation
(9/12/01 - 6/30/02)
SOUTH OF CANAL STREET
Residents
- primary residence on 9/11/01
Student and School Staff
- pre-K through 12th grade enrolled on 9/11/01
Pier 25
Hudson River
Canal Street
Chambers Street
Broadway
Rector St
SOUTH OF CHAMBERS STREET
Residents
- primary residence on 9/11/01
Students and School Staff
- pre-K through 12th grade enrolled on 9/11/01
Occupants of Buildings and Passersby
- employees and visitors on 9/11/01
Responders and Workers
- at disaster site 9/11/01 - 6/30/02
Pier 6
2nd Ave & Hamilton Ave
Brooklyn

Figure 1. Map of target area for WTCHR.

DOI: 10.1002/sim

represents a sample frame of individuals who may be contacted for future studies over the next 20 years.

For the WTCHR, maximizing population coverage was a key challenge. Coverage is defined as the extent to which the sample frame matches the true eligible population. A sample frame is ‘perfect if every element appears on the list separately, once, and only once, and nothing else appears on the list.’ [2]. Coverage estimation is used to assess the recruitment of the population at risk, and is a component in response rate calculations. Coverage error occurs when some persons are omitted from the list or frame used to identify members of the study population [3]. Overcoverage results from duplication of records in the sample frame. While one goal of the WTCHR was to maximize coverage overall and for each exposure group, another was to ensure equal representation within exposure groups (i.e. comparable proportions of registrants per group by demographic categories). Non-coverage occurs when portions of the true eligible population are not included in the study. Non-coverage can lead to biased results if those included in the sample frame are systematically different than those who are not included on some key measure.

ESTIMATING THE TRUE ELIGIBLE POPULATION

In epidemiological studies, the true eligible population is often referred to as the *denominator*, since it is the base number of persons from which coverage calculations can be made. An accurate denominator is a critical component of the WTCHR, since it will come to represent the number at risk of developing adverse health conditions due to exposure to dust and debris from the WTC disaster. For the WTCHR, we have used the term true eligible population to be consistent with the eligibility criteria developed for the different populations at risk. The true eligible population is also the denominator used to calculate the WTCHR-specific coverage rates, whereas the registry sample is the base for response rate calculations. The overall denominator has been estimated using strict geographic and temporal boundaries for the four target populations at risk: workers, residents, occupants and passersby, and students and staff.

Estimating the WTCHR denominator represents a unique challenge, due to several difficult temporal and spatial characteristics of the population. The WTCHR aims to include people who were living, working, or going to school in the study area at a point in time that has passed. Thus, we are dealing with a retrospective cohort. A subset of people living and working around the WTC on 11 September 2001 now live or work elsewhere, or may no longer be living. Others have been born into or moved into the area. The current population of lower Manhattan includes many who were not present on 11 September 2001 and excludes many who were.

Because of the multiple sample types pursued (residents, building occupants and passersby, students/school staff, and rescue/recovery workers), several approaches were required to determine eligibility. For some sample types, such as residents, lists of households or individuals were much easier to acquire than for others (for instance, passersby). Details on the methods used to estimate denominators for each sample type follow.

Rescue, recovery, and clean-up workers

As compared with other sample types, data on the rescue, recovery, and clean-up workers came from a wider variety of sources. This included employer rosters, government agencies, unions, and other research projects involving this population. Estimating a denominator for this group

DOI: 10.1002/sim

Table I. Estimated true eligible population of Group 1 rescue/recovery and related workers, WTCHR.

Organization type	Number eligible	Number of organizations
Total	*91 469*	*449*
City agencies	26 659	36
State agencies	8897	46
Federal agencies	5122	31
FEMA	3499	22
Volunteer	26 480	13
Rescue/recovery	20 397	238
Other	415	63

Table II. Estimated true eligible population of Group 1 and Group 2 residents, WTCHR.

Group	Estimated eligible population
Total	*57 511*
Group 1 (south of Canal)	21 926
Group 2 (on or North of Chambers and south of Canal)	35 585

was especially difficult, given the fact that the time period for eligibility was over eight months. Project staff collected the number and names of potentially eligible registrants from pre-specified agencies known to have employed WTC rescue and recovery workers. These contacts provided lists of employee names and contact information that make up the portion of these workers to be actively traced. Table I presents the denominator estimates for this sample type by organization type.

Residents south of Canal Street

The most recent U.S. Census of population and housing was conducted on 1 April 2000—a little over a year prior to the 11 September 2001 terrorist attacks on the WTC. The decennial census collects a limited number of data elements on every person and housing unit in the United States, including age, sex, race/ethnicity, tenure (whether the home is owned or rented) and vacancy characteristics [4]. Summary data are available down to the block level, but are also available at the block group, tract, and ZIP Code Tabulation Area (ZCTA) level. While census data are subject to a small amount of under or overcoverage (estimated between 0.12 per cent undercoverage and 0.50 per cent overcoverage nationwide [5, 6] they provide the most timely and accurate estimates of the residential population south of Canal Street in lower Manhattan on 11 September 2001.

Because of their high level of coverage and the short amount of time between the Census and WTC disaster, Census data are used as the primary source for estimating the resident portion of the denominator. Census data are publicly and freely available from the U.S. Census Bureau. Table II presents the Census population counts by Group 1 and Group 2 for the catchment areas defined for the WTCHR.

DOI: 10.1002/sim

Table III. Estimated true eligible population of Group 1 student/school staff, WTCHR.

School type	Number eligible	Number of schools
Total	*15 197*	*37*
Public schools	12 623	14
Private schools	847	5
Preschools/daycares	1727	18

Students and school staff

The denominator for the students enrolled and staff employed in schools as of 9/11 was estimated from school-level data gathered within the catchment area. Public school denominator data were obtained from the 2000–2001 Common Core of Data (CCD) school year data set that is publicly available from the National Center for Education Statistics (NCES). Denominator information for private schools, preschools, and day cares was collected during telephone contacts and/or in-person visits with the school representatives. Table III presents the estimated number of students, children and staff in each of the area schools and day cares.

Occupants of collapsed and damaged buildings

The denominator for employees and other non-residential building occupants is made up of all such persons present south of Chambers on 11 September 2001. Employees from businesses in 35 buildings and three structures identified by the New York City Department of Buildings as buildings that could not be occupied after the 11 September 2001 attacks and collapse of the WTC towers comprise the Group 1 building occupants and were identified for active tracing through the sample building process. Figure 2 shows the location of 35 of these buildings and two of the three structures [7] in and around the WTC complex (concourse not shown in the exhibit).

As compared to the number of residents, the number of building occupants who were physically present was much more difficult to estimate. There is no one data source like the Census that can provide an accurate enumeration for this group. However, through direct contact with businesses, estimates of the number eligible were obtained. As part of the WTCHR sample building process, project staff collected the number and names of potentially eligible registrants from businesses in the 35 buildings. Businesses provided lists of employee names and contact information and these data were used to actively trace potential registrants. Direct counts were obtained for 239 of the 1212 total businesses in these buildings. Using data on business size and office space from secondary sources, estimates of the number eligible were imputed for the remaining businesses. Using this method, the total estimated number of eligible Group 1 building occupants is 62 092. The estimated total number of individuals who were in WTC Towers 1 and 2 is 24 015. The WTCHR estimate is higher than what has been reported in other research,[§] but this current estimate is based on methodology that we believe to be defensible.

[§] For instance, an average turnstile count estimates 14 154 people typically arrive between midnight and 8:47 AM in the towers [8]. Based on eligibility of a sample of survivors interviewed, the National Institute of Standards and Technology estimates the tower population to have been 17 400 [7].

DOI: 10.1002/sim

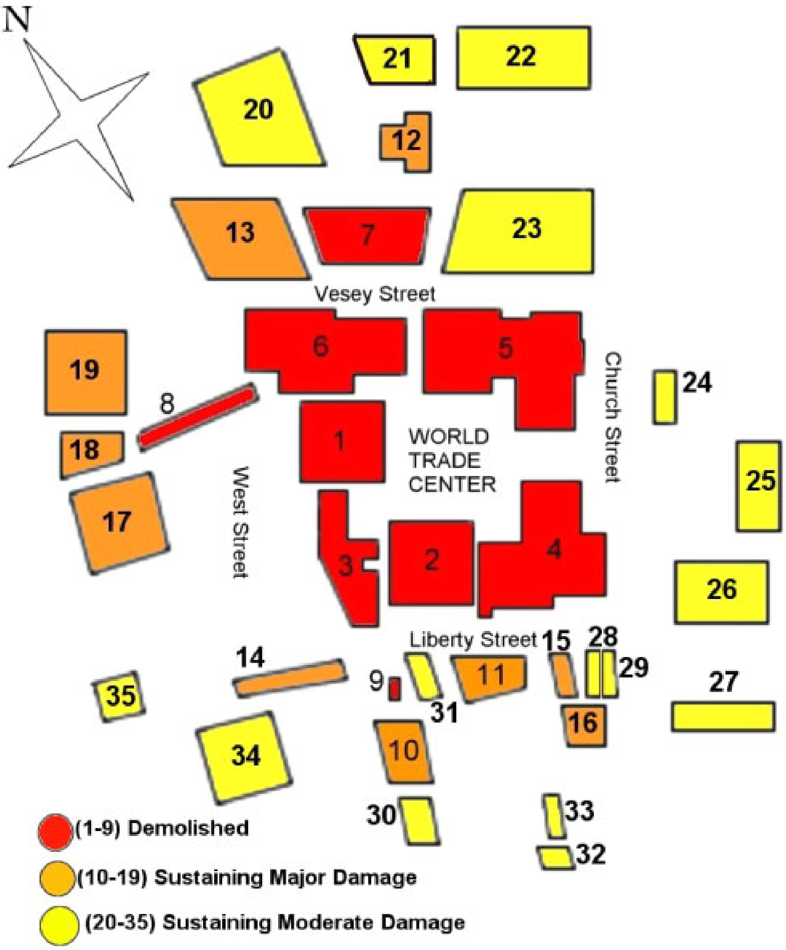

Figure 2. Buildings damaged or destroyed during the 11 September 2001 attack on the World Trade Center. Federal Emergency Management Agency (FEMA) World Trade Center Building Performance study (http://www.fema.gov/library/wtcstudy.shtm).

Other employees, occupants of buildings south of Chambers, and passersby (Group 2) entered the pre-registry database through self-identification rather than being actively traced. Because of the wide variation in composition of this group, data for denominator estimation for this group were difficult to obtain. An overall estimate for this group of 300 000 people was based on 2000 Census Journey to Work data which included number of commuters to Lower Manhattan (Area 1) (NYC Department of City Planning, 2000 Census to Work, special tabulation). Table IV presents the estimated denominator for building occupants and passersby by group.

Estimating the overall true eligible population for the WTCHR

The four registrant sample types (rescue, recovery, and clean-up workers; residents; students and school staff; and building occupants and passersby) do not represent mutually exclusive categories.

Table IV. Estimated true eligible population of Group 1 and Group 2 building occupants and passersby, WTCHR.

Group	Estimated eligible population
Total	*362 092*
Group 1 (building occupants of 35 damaged/destroyed buildings and 3 structures)	62 092
Group 2 (building occupants in buildings south of Chambers, passersby)	300 000

There is some level of overlap between the types, such that the sum of each subgroup denominator exceeds that for all registrants. The level of this overlap must be assessed in order to compute one number estimating the overall population at risk.

For instance, an individual could have worked at the WTC and lived in lower Manhattan on 9/11. If this person was included in both sample types, he or she would end up being counted twice in the overall denominator. Completed interview data can provide the basis for computing the overall estimate. Each registrant completing an interview was placed in one or more of the four sample types and a priority group based on his or her responses. Completed interviews of each sample group and type were classified by the total number of groups and types under which they were eligible. This distribution was then applied to the marginal estimates of true eligibles for each sample group and type to predict the number that would be classified under multiple sample types in the population. Because multiple types represent duplication in the count of unique persons, counts of cases identified under more than one sample type were divided by the number of types under which one would be classified, thereby converting person-type counts into simply person counts. The formula for this calculation is

$$\text{Overall true eligible population} = \sum\sum[(r_{xy}/r_{x}. * a_{x}.)/y]$$

where r_{xy} is the completed interviews of type x eligible under y (1–4) number of types, $r_x.$ the marginal number of interviews of type x and $a_x.$ the marginal number of true eligibles of type x.

As an example, the solution for the Group 1 (G1) occupant component of this formula follows:

$$\begin{aligned}
&\text{True eligible population component for Group 1 occupants}\\
&\quad = [(7802 \text{ G1 occupant completes of one sample type}/10\,393 \text{ G1 occupant completes}\\
&\qquad * 62\,092 \text{ G1 occupant true eligibles})/1 \text{ sample type}]\\
&\qquad +[(2457 \text{ G1 occupant completes of two sample types}/10\,393 \text{ G1 occupant completes}\\
&\qquad * 62\,092 \text{ G1 occupant true eligibles})/2 \text{ sample types}]\\
&\qquad +[(133 \text{ G1 occupant completes of three sample types}/10\,393 \text{ G1 occupant completes}\\
&\qquad * 62\,092 \text{ G1 occupant true eligibles})/3 \text{ sample types}]\\
&\qquad +[(1 \text{ G1 occupant completes of four sample types}/10\,393 \text{ G1 occupant completes}\\
&\qquad * 62\,092 \text{ G1 occupant true eligibles})/4 \text{ sample types}]\\
&\quad = 54\,218
\end{aligned}$$

DOI: 10.1002/sim

Summing over all sample types and groups, an estimated overall denominator of 409 492 persons is obtained. While this number does not figure into any of the final WTCHR outcome rates, it is important as a stand-alone figure as it represents the best estimate the project can provide of the true number of persons exposed to the events of 11 September 2001, as exposure is defined by the WTCHR.

MAXIMIZING COVERAGE

Building the sample frame

During the design phase of the WTCHR, it was determined that the ideal strategy for Group 1 recruitment and enrollment involved requesting and obtaining lists of names from the various entities that had payroll lists or registries. Such entities included business operations that had employees who worked in the 35 buildings; contractors, government agencies, unions that hired and managed rescue, recovery, and clean-up workers at the WTC site or separate WTC locations such as Staten Island or barges; or the New York City Department of Education and individual public or private schools. List building and tracing individuals to verify contact information was considered the best methodology for identifying and interviewing the largest proportion of the higher priority groups. This approach to developing the sample for Group 1 of the WTCHR was first described in the WTCHR protocol.[¶] It was then subsequently enhanced and implemented by the WTCHR baseline data collection contractor RTI International.

List building and active tracing

The list building and tracing process consisted of four basic stages. These include enumeration of list source entities, contacting the entities and requesting lists, obtaining the lists, and building a pre-registration data base with contact information of persons on lists. The fourth stage included correcting contact data on the lists that did not contain a telephone number. This was done by using standard tracing and locating methods to obtain verifiable telephone numbers or address information. These four stages were applied to the four sample types with variations in the approaches for each.

Lists of residents to be traced were obtained primarily from sample files purchased from Genesys Sampling Systems. Additional sample elements were generated by asking respondents from the resident group about other members of the household; this includes other adults who lived at the same address and children. Additional potential contacts were also generated by collecting the names of persons who were offered by a respondent during a face-to-face interview or during a locating and tracing contact. This method is similar to network sampling, which is used in surveys of special populations to increase coverage [9]. Residents who self-identified were added to the pre-registry database as well.

The primary source of businesses contacted for the WTCHR was a sample list purchased from Genesys Sampling Systems. This list contained every business in the Genesys business database with an address at one of the damaged or destroyed buildings as of 10 September 2001. In all, Genesys provided contact information for 906 businesses.

[¶] Brackbill and Thomas (2003). Protocol for the World Trade Center Health Registry. New York City Department of Health and Mental Hygiene in collaboration with Agency for Toxic Substances and Disease Registry. New York, NY (available upon request).

DOI: 10.1002/sim

To maximize coverage for businesses in the 35 buildings, two additional lists of organizations were obtained at no cost. The first was a list of 587 businesses that the Downtown Alliance derived from Dun & Bradstreet listings as of August 2001. Although there was significant overlap between the Dun & Bradstreet list and the Genesys list, it did provide some unique records and improved the coverage of businesses. Similarly, the New York Metropolitan Transportation Council (NYMTC) published a report titled 'post September 11th impacts: inventory of affected businesses.' [10]. The businesses listed in this report lost space in the World Trade Center and adjacent properties destroyed and damaged on 11 September 2001. Some, but not all, of the 35 buildings identified for the WTCHR are included in the NYMTC report. The NYMTC report identifies 879 businesses in the 35 buildings, but there is significant overlap with the Genesys and Downtown Alliance lists. Since many businesses are included on two or three of the above list sources, deduplication is conducted to eliminate duplicate business-level records. As mentioned previously in this report, businesses were contacted to supply the names and contact information of employees who were likely to be eligible for the WTCHR.

Self-identification

Anyone who learned about the WTCHR by receiving information from their employer, a media campaign, or community outreach could self-identify by pre-registering on the WTCHR web site or calling a toll free number. Eligibility for the WTCHR as a member of one of the defined exposure groups was determined during the initial portion of the interview.

Deduplication

To avoid *overcoverage*, or error due to duplication of records from different sources, all sample elements were first standardized to the extent possible. For instance, address formats were specified to avoid multiple abbreviations for the same information (St., Str., Street, etc.). Similar standard approaches were taken with other data elements, such as name prefixes and suffixes. Software was utilized for the pre-registry database to reduce the presence of duplicate records. This software package developed by Choicemaker Technologies, Inc. utilized an algorithm for identifying likely occurrences of duplication in lists based on name, address, and social security number information, among other data elements [11]. This process was run on the pre-registry database on a nightly basis. For more information on deduplication techniques used in the WTCHR, see [12].

REPORTING OF COVERAGE

To evaluate the recruitment of eligible registrants from the estimated true eligible population we calculated the *enrollment rate*. This is equivalent to the total number of registrants completing an interview relative to the estimated total number who were eligible to participate.

Table V presents the WTC enrollment rate by sample type and group, where applicable.

An overall enrollment rate of 17.4 per cent was achieved for the WTCHR. Among residents, the Group 1 enrollment rate of 37.3 per cent was much higher than the Group 2 rate of 15.0 per cent. Similarly, the enrollment rate for Group 1 building occupants was 16.7 per cent while the Group 2 rate was lower at 11.0 per cent. Coverage was relatively high for rescue/recovery workers (33.5 per cent) and relatively low for students/school staff (17.4 per cent).

DOI: 10.1002/sim

Table V. Enrollment rates by sample type and group.

Sample type/group	Completed interviews	Estimate eligible population	Enrollment rate (per cent)
Overall (*without ranking*)*	*91 463*	*526 269*	17.4
Rescue/recovery workers	30 665	91 469	33.5
Residents	14 665	57 511	25.5
Group 1 (higher exposure)	8170	21 926	37.3
Group 2 (lower exposure)	5319	35 585	15.0
Students/school staff	2646	15 197	17.4
Building occupants/passersby	43 487	362 092	12.0
Group 1 (higher exposure)	10 393	62 092	16.7
Group 2 (lower exposure)	33 094	300 000	11.0

*The numerator for the overall coverage rate does not equal the sum of numerators by sample type. Because the denominator is the sum of all estimated eligible person-types (without ranking), the numerator must equal the total number of person-type combinations among completed interviews (91 463) rather than simply the total number of completed interviews at the person level (71 463).

DISCUSSION

In designing a registry like the WTCHR that relies on many sources of data to build the sample, it is important to develop a comprehensive approach that considers all possible sources of error and minimizes bias that may be introduced through the methodology. For the WTCHR, we developed a tracking and reporting process that allowed us to monitor and control for non-coverage to the greatest extent possible, across all key exposure groups. This process was a well-crafted feedback loop that allowed sample building processes to inform subject recruitment and enrollment, and *vice versa*.

One of the most difficult analytic issues we faced in this process was estimating the denominator for the different populations at risk included in the WTCHR. An important lesson to be learned is that all possible avenues of obtaining information about the eligible populations should be pursued in the early stages of registry development. For instance, some major sources of information about where people were and when they were there may only be available for a short period of time in the aftermath of an event. In addition, this information may originate from unexpected places. It is also important to include denominator estimation early in the registry design process so that the appropriate data for making calculations will be available. All potential sources of error and methods for measuring error should be investigated prior to data collection so the information necessary for evaluating coverage is present.

Another important lesson learned is that list acquisition and tracing are effective approaches to building a sample for a registry. Prior experiences with this approach only involved obtaining a list of names from one or two specific sources (e.g. post office survey; NIOSH cohort studies; other WTC disaster related studies [13]). It had not been applied to trying to gain the goodwill and cooperation of hundreds of businesses, agencies, and unions as was the case for the WTCHR baseline recruitment programme. Overall, more than 130 000 of the total 197 000 potential registrants were identified from list building activities. The remainder self-identified *via* phone or web.

We attempted to obtain the estimated denominator for building occupants by testing a method in which the proportion of listed sample members who self-identified was applied to those who were not listed. This is an application of the capture–recapture method of population estimation, which is often employed in epidemiological studies [14]. This methodology was originally developed by ecologists who were interested in estimating the size of animal populations. One caveat of applying this method is that the chances of being included on a list and self-identifying are likely not independent. Outreach and the response of businesses will vary and influence this proportion. For instance, some entities sent messages to their employees encouraging them to register or provided them with self-identification kits, while others were much more passive. Some companies were within the media reach of the WTCHR while others were outside. Several critiques of the application of capture–recapture to epidemiological studies have been made, with suggestions for refinement [15]. These critiques emphasize the fact that cases must be intensively scrutinized in an attempt to identify those characteristics that might act to exclude a case from one list or another. Estimates of building occupants were produced using this capture–recapture method but the assumptions underlying this method were not considered robust enough to warrant inclusion of the results or further discussion in this paper. As a result, we only present the final methodology used to estimate building occupants in this paper.

Another difficult issue in this type of registry is the sheer magnitude of duplicate cases, a source of error simply too great to ignore. For the WTCHR, we identified over 20 000 cases as duplicates. The implementation of sophisticated software algorithms to eliminate duplicate cases and identify potential duplicates for human review is critical. Without such a process, the bias introduced by duplicates would be significant. It is vital that the most accurate methods of deduplication be applied to all the different sources of pre-registration or sample so that an accurate estimate of coverage can be achieved.

The enrollment rates by sample type ranged from 11.0 to 37.3 per cent with the lowest rates obtained for building occupants/passersby. As noted earlier in this paper, this exposure group was the most difficult for which to estimate the true eligible population. It was also challenging to obtain lists of names of individuals in the 35 damaged buildings to contact for enrollment. In May 2004, nine months into a planned twelve month data collection period, we received a list from the Port Authority of approximately 95 000 names who had been issued a security badge to one or more of the seven WTC buildings at any time during the past five years. Due to budget and time constraints, we subsetted the list according to completeness of contact information such that 30 per cent of the list were contacted by mail and phone follow-up, 51 per cent by mail only, with the remainder not pursued at all due to insufficient contact information. Given the concerns over the low enrollment rates and the possible impact on registry results, we conducted an initial investigation of non-response and self-selection bias. The data indicate that there may have been a modest amount of non-response bias that could result in slight overestimates of prevalence rates on the outcome measures of interest. For more information on this analysis, please see [16].

Finally, it is advisable to design a small post-enumeration survey to validate the various methods for assessing coverage on a registry. With more information available about the nature of and causes for coverage error, the likelihood that errors will be accounted for or reduced in key population estimates will increase significantly.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention/Agency for Toxic Substances and

Disease Registry (ATSDR). The findings have undergone external peer review as required by ATSDR policy and the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA).

REFERENCES

1. Lunsford TK, Lunsford BR. The research sample. Part I: Sampling. *Journal of Prosthetics and Orthotics* 1995; **7**(3):105–112.
2. Kish L. *Survey Sampling*. Wiley: New York, 1993.
3. Groves RM. *Survey Errors and Survey Costs*. Wiley: New York, 1989.
4. U.S. Bureau of the Census. *Introduction to Census 2000 Data Products*. Washington, DC. http://www.census.gov/prod/2001pubs/mso-01icdp.pdf#search=%22introduction%20to%20census%202000%20data%20products%22 (accessed 10/11/06), 2003.
5. Robinson JG. ESCAP II: Demographic Analysis Results, Executive Steering Committee for A.C.E. Policy II. U.S. Bureau of the Census, Report 1, October 13, 2001. http://www.census.gov/dmd/www/pdf/Report1.PDF #search=%22%22ESCAP%20II%3A%20Demographic%20Analysis%20Results%E2%80%9D%22 (accessed 10/12/06), 2001.
6. U.S. Bureau of the Census. *Technical Assessment of A.C.E. Revision II*. Washington, DC. http://www.census.gov/dmd/www/ace2.html (accessed 10/11/06), 2003.
7. Dwyer J, Flynn K. *102 Minutes*: *The Untold Story of the Fight to Survive Inside the Twin Towers*. Times Books: New York, 2005.
8. Averill JD. World Trade Center Investigation Status Project 7: Occupant Behavior, Egress, and Emergency Communications. http://wtc.nist.gov/pubs/June2004OccupantBehaviorEmergencyCommunications.pdf (accessed 10/4/06), 2004.
9. Sudman S, Kalton G. New developments in the sampling of special populations. *Annual Review of Sociology* 1986; **12**:401–429.
10. Urbanomics. *Post September 11th Impacts*: *Technical Memorandum*. Prepared for the New York Metropolitan Transportation Council (NYMTC). http://www.nycp.org/reports/ImpactStudy.pdf#search=%22 urbanomics%20post%20september%2011th%20impacts%22 (accessed 10/11/06), 2003.
11. Choicemaker Technologies Incorporated. http://www.choicemaker.com/ (accessed 10/11/06).
12. Murphy J, Pulliam P, Lucas R. Sample Frame Deduplication in the World Trade Center Health Registry: minimizing overcoverage and cost. *Proceedings of the American Statistical Association*, *Section on Survey Research Methods* [CD-ROM], Alexandria, VA, American Statistical Association, 2004; 4072–4076.
13. National Institute for Standards and Technology. *NIST Interim Report on Analysis of First-Person Accounts from Survivors of the WTC Evacuation on 11 September 2001*: *Appendix O*: *Interim Report on Telephone Interview*. 20 October 2004; Washington, DC. http://wtc.nist.gov/progress_report_june04/appendixn.pdf, 2004.
14. Hay G, McKeganey N, Wiessing L *et al*. *Methodological Guidelines to Estimate the Prevalence of Problem Drug Use on the Local Level*. University of Glasgow Centre for Drug Misuse Research, 1999.
15. Jarvis S, Lowe PJ, Avery A *et al*. Children are not goldfish—mark/recapture techniques and their application to injury data. *Injury Prevention* 2000; **6**:46–50. DOI 10.1136/ip.6.1.46.
16. Murphy J, Brackbill R, Sapp JH *et al*. An analysis of nonresponse bias in the World Trade Center Health Registry. *2005 Proceedings of the American Statistical Association*, *Section on Survey Research Methods* [CD-ROM], Alexandria, VA, American Statistical Association, 2005.