Sampling Racially Matched Population Controls for Case-Control Studies: Using DMV Lists and Oversampling Minorities

Ralph DiGaetano, Westat; Barry Graubard, National Cancer Institute; Sowmya Rao, National Cancer Institute; Jacqueline Severynse, Westat; Sholom Wacholder, National Cancer Institute

Overview

The Kidney Cancer Study (KCCS) is currently being carried out for the National Cancer Institute, focusing on renal cell cancer. The incidence rate of renal cell cancer is growing at a higher rate among Blacks than other race/ethnic groups. However, there are many more White cases than Black, so both race groups are of interest in the study. The study is being carried out at two sites with large Black populations in both number and as a percentage of the total population: Chicago (specifically, Cook county) and Detroit (specifically, Wayne, Oakland, and Macomb counties). Controls are frequency matched to cases on the characteristics of age, sex, and race with the age range of interest being 20-79. Controls aged 65 and older are being selected from lists of Medicare Beneficiaries, where age, sex, and race are all available.

Initially, the plan was to select samples of controls aged 20-64 using a list-assisted Random Digit Dialing (RDD) approach for identifying pools of people eligible to serve as controls in this age range (see DiGaetano and Waksberg, 2002). However, due to low response rates in the initial wave of data collection, it was decided to use listings from the Department of Motor Vehicles (DMV) of Illinois and Michigan to form sample frames. Neither of these DMV lists provides race.

Waksberg, Judkins, and Massey (1997) showed that oversampling Census block groups with high percentages of Blacks can be a relatively efficient approach to oversampling Blacks in an area sample design. Identifying the block groups associated with addresses found on a DMV file thus has the potential of accomplishing a similar sort of efficiency conditional on a suitable means for linking addresses to block groups. Geocoding provides such a means

Geocoding is a relatively recently developed methodology. Software programs are used to link addresses to Census areas such as tracts, block groups, and blocks. In urban settings, geocoding assignments of block groups to addresses can be expected to be highly accurate. Thus, geocoding provides a means for oversampling Blacks on list frames where race is not provided but address is.

This paper provides an initial evaluation of using Illinois and Michigan DMV listings for sampling purposes where age and sex are stratification variables, focusing on issues of coverage. In addition, a preliminary assessment of strata based on the use of geocoding for oversampling blacks is provided.

The Sample Design

Elements Specific to Case Control Studies

First, we will briefly describe some of the basic aspects of a case control study sample design. Both cases and controls are regarded as selected through stratified random sampling. The strata are created through the cross-classification of known risk factors of the disease under study (this might be, for example, age and sex or age, sex, and race). Cases are viewed as sampled from the "diseased" population while controls are sampled from members of the general population who do not have the disease. The control sample allocation across strata may be identical to that of cases, double that of cases, etc. This allocation is described as matching at a rate of 1 to 1, 2 to 1, etc. Further discussion about the sample design of case control studies can be found in Korn and Graubard (1999). Specific case control related elements of the KCCS sample design discussed below include the frequency of sampling, the criteria for matching, and the sampling of cases.

Case accrual takes place over roughly a four year period. The Chicago site cases are obtained through Cook County hospitals. The Detroit site cases are obtained through the Detroit Cancer registry that is part of the Surveillance, Epidemiology, and End

Results Program of the National Cancer Institute (the Detroit SEER registry). Controls are selected during the case accrual period at approximately six month intervals, each six month period being characterized as a wave.

The controls are population based and the sample distribution of controls is matched proportionately for each race separately to the expected sample distribution of cases across the age-sex strata established for sampling cases. For controls aged 65-79, selected from lists of Medicare Beneficiaries, the matching is based on expected actual sample sizes. For controls aged 20-64, selected from the DMV lists, this matching is based on expected effective sample sizes, taking into account the design effects associated with the oversampling of controls found in “high density Black” areas as described below.

The frequency matching of controls to cases is two to one for Blacks and one to one for Whites. Note that the ratios of the actual number of participating controls aged 20-64 to corresponding participating cases for Blacks and Whites will be somewhat higher than the originally specified "two to one" and "one to one", respectively, since the matching of this set of controls is based on effective sample sizes accounting for design effects that are greater than 1.

All Black cases are selected for study participation. The final strategy for sampling White cases has yet to be established. The number of White cases is substantially larger than those of Blacks. The current design is to establish sampling rates for White cases so that the expected number of White cases for a given sex-age group combination (e.g., males, aged 45-49) is to be double the number of Black cases expected for that combination.

The samples of controls aged 65-79 are subsamples of samples provided by the Centers for Medicare and Medicaid Beneficiaries (CMS) from their listings of Medicare Beneficiaries. These listings include data on age, sex, race, and address.

Elements Related to the Oversampling of Controls Aged 20-64 Living in High Density Black Areas

The sample design for the sample selection of controls aged 20-64 was developed with the following features: a stratification of block groups into high and low density Black designations; selection of a sample of DMV records to be geocoded so that each sampled address is assigned to a block group and thus a "Black density" stratum; and selection of a sample of persons from a set of strata defined from the cross-classification of the three variables "Black density", sex, and age-group. The following steps were taken to implement the design.

1. Determine, using Census data, a stratification of block groups that can be expected to be relatively efficient in terms of oversampling. For this study two general strata were formed, “high” and “low” Black density. The formation of these strata took into account the concentration of Blacks and non-Blacks in each stratum as well as the coverage of these two subpopulations. For both Detroit and Chicago the “high density Black” stratum covers roughly 85 percent of the Black population and about 85 percent of the people in the stratum are Black, based on Census 2000 figures. The “low density Black” stratum covers roughly 95 percent of the non-Black population and 95 percent of persons are expected to be non-Black. The vast majority of persons non-Black in these two sites is White.

2. Obtain all records from a state's DMV associated with the targeted counties. (Obtaining all records means that one can directly learn about and deal with issues of duplication on the file.) Establish a file of those persons considered to have currently active driver’s licenses and IDs within the targeted age range, based on information such as date of birth, license issue date, and license expiration date. Unduplicate the file using name, address, and date of birth. License number is not available for unduplication. The version of the file containing the currently active records remaining after unduplication represents a state’s sample frame for a given wave.

3. Select a systematic sample of records from the sample frame to undergo geocoding, sorting on sex and age prior to sample selection.

4. Geocode the sampled records, linking addresses to block groups via software algorithms. The use of ZIP Code to assign records with problem addresses to a block group permits a mechanical assignment which does not result in bias from misassignments to incorrect block groups but has the potential of reducing the effectiveness of the stratification or increasing the variance of study estimates. There have been relatively few such records, so any impact of misassignments on study estimates due to the use of ZIP Codes when address information does not suffice will be small.

5. After geocoding, assign the sampled records to sample strata based on the cross-classification of three variables: Black density (high, low), sex, and age group (five-year intervals from 20-24 through 60-64).

6. Evaluate the coverage of the DMV listings by comparing the DMV sample distribution across the sample strata to that expected based on Census 2000 figures.

7. Evaluate the potential impact of geocoding on the effectiveness of the stratification. This involves looking at the percentage of records assigned to the high density Black stratum based on geocoding compared to the percentage expected based on 2000 Census data for each age-sex cross-classification.

8. Determine sampling rates within strata taking into account the expected number of Blacks and Whites per stratum and the degree of oversampling necessary to obtain sample yields with desired power after accounting for design effects.

DMV Listings: IDs as Well as Driver's Licenses

Some people do not have a driver's license but need the type of identification that a driver's license provides. Thus, state Departments of Motor Vehicles generally will provide an ID to such a person who is a resident of that state. Including DMV records representing IDs on a sample frame can thus enhance coverage of the target population eligible to serve as controls. This can be particularly helpful if a state provides all DMV records associated with the area of interest for the study. However, some people may have both an ID and a license. Thus, if a state provides only a sample of DMV records, the extent of duplication between IDs and driver's licenses cannot be determined.

IDs represent a considerable proportion of the Chicago DMV records (20-45%, depending on age-sex cross-classification) but a negligible proportion for Detroit (roughly .5% overall). This may be attributable to the larger role played by mass transit in the Chicago area compared to Detroit.

Coverage Evaluation

We have evaluated coverage thus far for both the Chicago and Detroit sites based on the first set of files containing all DMV records (rather than a sample) provided to us by the Illinois and Michigan DMVs. The Chicago file was established by the Illinois DMV in June 2003, while the Detroit file was established by the Michigan DMV in October, 2002.

Direct comparisons between DMV and 2000 Census figures were made based on the cross-classification of sex and age group variables (age groups being five year intervals from 20-24 through 60-64). "Coverage rates" were computed as ratios of DMV counts to Census 2000 counts multiplied by 100. The DMV counts represent the set of persons with active drivers' licenses or IDs at the time the DMV file was produced after unduplication. Note that, generally, one expects some population growth over a two or three year period, so these ratios may overstate coverage somewhat.

It should be noted that, for case control studies, analyses are frequently done after excluding cases that are not found on the sample frame for controls. Thus, if controls are selected through RDD telephone screening operations, cases without a telephone at home are excluded. Requiring that the cases and controls be sampled from the same population (in this case individuals reachable by telephone) permits unbiased analyses. Similarly, for this study it is planned that cases that indicate that they do not have a driver's license or DMV ID will be excluded from analyses where corresponding controls have been selected from DMV listings. Nevertheless, if there is some degree of undercoverage of the target populations associated with a DMV file, two concerns arise. The greater the undercoverage, the greater the degree of uncertainty that the study results pertain to the general population. In addition, for situations of relatively low coverage, one can expect some loss of cases, those that do not have drivers' licenses or IDs, which might otherwise have qualified for inclusion in study analyses. Thus, the strategy of removing from analyses cases that have neither a driver's license nor ID can hamper the ability to generalize from the analyses and can reduce power and efficiency.

Table 1 shows estimated coverage rates for Chicago (Cook County), Detroit (all three counties combined), and Wayne County alone. The DMV figures represent the sum of persons with a driver's license or ID or both (DLs + IDs). Note that Wayne County contains close to 85 percent of all Blacks in the three counties and 91 percent of the Blacks found in the block

groups characterized as "high density Black" based on 2000 Census figures. Considering the Wayne County figures alone was done to identify potential issues related to the coverage of the Black population by DMV records.

Table 1. Estimated coverage ratios for Chicago, Detroit, and Wayne County

	Chicago			Detroit			Wayne		
Sex by Age	Census	DMV (DLs + IDs)	Coverage rate	Census	DMV (DLs + IDs)	Coverage rate	Census	DMV (DLs + IDs)	Coverage rate
M_20_24	192,950	174,894	90.6	115,553	113,039	97.8	62,775	55,345	88.2
M_25_29	222,291	208,859	94.0	140,775	135,659	96.4	72,155	64,122	88.9
M_30_34	215,342	216,585	100.6	153,041	156,949	102.6	75,114	73,055	97.3
M_35_39	210,144	203,232	96.7	162,438	154,656	95.2	77,152	69,942	90.7
M_40_44	200,178	200,176	100.0	163,738	157,397	96.1	78,213	68,948	88.2
M_45_49	173,419	184,476	106.4	145,582	149,168	102.5	69,701	66,162	94.9
M_50_54	148,948	154,361	103.6	125,544	127,408	101.5	59,512	56,787	95.4
M_55_59	112,214	123,284	109.9	93,603	102,817	109.8	43,053	44,837	104.1
M_60_64	90,692	94,442	104.1	67,268	70,063	104.2	31,836	30,402	95.5
Males	1,566,178	1,560,309	99.6	1,167,542	1,167,156	100.0	569,511	529,600	93.0
F_20_24	192,371	177,159	92.1	117,925	107,745	91.4	65,340	52,389	80.2
F_25_29	222,074	201,434	90.7	146,337	131,855	90.1	78,153	63,252	80.9
F_30_34	213,117	201,250	94.4	156,058	152,608	97.8	78,888	72,103	91.4
F_35_39	212,788	188,897	88.8	166,174	151,838	91.4	81,241	68,564	84.4
F_40_44	209,737	196,385	93.6	171,058	158,016	92.4	83,244	70,177	84.3
F_45_49	186,522	189,288	101.5	152,912	153,459	100.4	75,290	68,771	91.3
F_50_54	164,401	165,525	100.7	133,372	132,204	99.1	64,830	60,114	92.7
F_55_59	128,287	136,332	106.3	99,727	105,720	106.0	47,732	47,066	98.6
F_60_64	107,528	107,829	100.3	76,886	71,635	93.2	38,267	31,778	83.0
Females	1,636,825	1,564,099	95.6	1,220,449	1,165,080	95.5	612,985	534,214	87.1

For both Chicago and Detroit the coverage rates appearing in Table 1 are generally around 100 or higher, particularly for the age groups covering the age range 45-64 (coverage of the older age groups is of greater import for this study since they are sampled more heavily than the younger ones, as the case population is concentrated more heavily in the older age groups). However, the rates for Wayne County alone are generally 5 to 10 percentage points lower than the overall Detroit rates.

The apparent undercoverage in Wayne County raises some issues. As discussed earlier, cases without driver's licenses or IDs can be excluded from analyses, and possible bias related to this undercoverage is not a concern in analyses when such cases are excluded. However, some concern could arise about the applicability of study results to persons without driver's licenses or IDs and the coverage rates estimated here are expected to overstate coverage to some degree, as mentioned above. Since cases are associated with older age groups, an examination of coverage among older ages is appropriate. The coverage rate for males aged 45-64 is about 97.1 in Wayne County while for females it is 91.9. The ratios for both males and females aged 45-64 is about four percent higher than the corresponding ratios for males and females over all age groups in Wayne, providing a somewhat more encouraging picture.

Preliminary Evaluation of the Effectiveness of the Stratification Based on the Geocoding of DMV Records

Once a sample frame was established for a site, a sample of the records was selected for geocoding. The initial sample of records selected for geocoding (wave 1 from the Detroit area) was kept small, about 3,000 records. This was due to concern about the possible need to manually code a nontrivial number of records if there were problems with the address fields. However, this did not prove to be the case. Few records had problem addresses and most of those could be assigned via software to a block group based on their ZIP Code. The sample from the sample frame established for Chicago sent for geocoding was increased to about 20,000 records. The plan is to use samples of at least 20,000 in the future.

When the sampled records were provided with a block group through geocoding, each record could be assigned as either in residence in a high or low density Black area. This permitted an evaluation of the estimated population percentage assigned to the "high density Black" strata for each sex-age group cross-classification by comparing the "DMV" percentage based on geocoding to corresponding Census figures. These ratios have sampling error associated with them as well as possible misclassification error attributable to the geocoding process.

Table 2 provides a comparison of the estimated "high density Black" percentages for Chicago, and Table 3 provides a similar comparison for Detroit. The geocoding of the 20,000 sampled records for Chicago shows the DMV percentages generally close to those obtained from Census data. The largest difference for males was in the 60-64 age group where a ratio of .8 was obtained. No other age group had a ratio below .9. For females there were four age groups with ratios in the vicinity of .9, the lowest being the age group of 60-64 with a ratio of .88. The Chicago ratios do not suggest that the effectiveness of the planned stratification will be reduced much due to issues of misclassification associated with geocoding or coverage of the Black community.

Table 2. Evaluating percentage high density across age-sex cross-classification: Chicago (n = 20,000)

Sex by Age-Group		Percentage High Density		Ratio
		DMV	Census	
Male:	20-24	25.4%	24.4%	1.04
	25-29	20.0	20.0	1.00
	30-34	19.4	19.7	0.98
	35-39	22.4	21.6	1.04
	40-44	22.8	22.8	1.00
	45-49	22.9	22.4	1.02
	50-54	21.1	22.4	0.94
	55-59	22.0	24.0	0.92
	60-64	20.4	25.5	0.80
Female:	20-24	30.0	27.0	1.11
	25-29	24.2	24.3	0.99
	30-34	23.3	24.3	0.96
	35-39	24.0	26.2	0.92
	40-44	24.6	27.1	0.91
	45-49	24.4	27.0	0.91
	50-54	27.0	26.6	1.02
	55-59	26.7	27.3	0.98
	60-64	25.1	28.6	0.88

Table 3. Evaluating percentage high density across age-sex cross-classification: Detroit (n = 3,000)

Sex		Percentage high density		Ratio
		DMV	Census	
Males	20-44	16.5%	22.2%	.743
	45-64	17.2	21.0	.819
Females	20-44	21.1	25.6	.826
	45-64	22.6	24.2	.936

The data from Detroit suggest that the stratification may not be as effective as planned for this site. Only about 3,000 records were geocoded for Detroit, so to use relatively stable DMV estimates for comparison, just two age groups were considered: ages 20-44 and 45-64. The ratios of DMV to Census percentages are generally much lower than those found in Chicago, with three groups ranging between .743 and .826 while females aged 45-64 had a ratio of .936. This finding is consistent with the suspicion that the relatively low coverage ratios associated with Wayne County are due to low coverage of the Black community by the Michigan DMV listings.

The immediate impact of this result has been on sample size. In order to achieve the targeted power, sample sizes for Detroit had to be increased more so than in Chicago to compensate for the estimated design effects resulting from the oversampling of the high density strata. Specifically, the estimated design effects associated with oversampling the high density strata in Detroit ranged from 1.35 to 1.45 across age-sex combinations. The corresponding range for Chicago was 1.25 to 1.35.

Currently, there is insufficient data to learn about the numbers and percentages of responding Blacks found in the high density vs. low density strata for the two sites. These race data will be collected from the interviews as the study progresses.

Implications for Future Studies

A common approach for the last two decades for case control studies when selecting population-based controls under the age of 65 has been to use an RDD methodology to obtain a pool of persons eligible to serve as controls (DiGaetano and Waksberg, 2002). This approach has become more problematic over the last few years as response rates for RDD surveys have steadily declined. The use of DMV listings provides an alternative that seems viable but has its own set of problems. We will discuss them here as well as make a few observations related to RDD surveys and the use of Medicare Beneficiary Listings.

In addition to the decline in response rates, there are some other potential concerns associated with RDD studies in the future. Phone numbers may become portable, allowing people to use the same phone number regardless of where they happen to move to within the U.S. If this becomes prevalent, the ability to carry out surveys other than at the national level may become problematic as coverage of the target population will become an issue. This is of particular concern to case control studies, which are generally conducted in one or several localities. Another uncertainty is the extent to which households will only use cell phones as time progresses. There are legal issues surrounding the calling of cell phone numbers for survey purposes, so coverage issues may arise if households increasingly rely only on cell phones.

Nevertheless, there are some potential bright spots for RDD methodologies as well. With the recent advent of the national "don't call" listings, the number of telemarketing calls is likely to be substantially reduced to many, perhaps most, households. Those involved in scientific research projects using RDD methodologies to collect data will have more credibility when contacting the household, as contacting households for such studies is still permitted under the law. In addition, the ability to link telephone numbers to addresses has been increasing. This permits advance letters to go out prior to telephone contact, adding to the credibility of the data collection effort and thus potentially increasing response rates and permitting the identification of addresses that are outside an area of interest. Moreover, changes in payment structure associated with cellular phones (e.g., where the caller to a cellular phone, rather than the receiver of the call, pays for the call) may make calling cellular phones a viable approach.

The viability of DMV files serving as sample frames for case control studies depends on a number of issues. First, some states may decline to provide such listings. For one health-related study (not case control) Georgia and Mississippi have

chosen not to do so. Second, states do not necessarily have race as a variable on their DMV file. Two that have reported having race are Tennessee and Florida. When race is not available, one must either screen for race or use an approach such as the one discussed here, oversampling strata expected to have disproportionately high numbers and/or percentages of a targeted racial group. Depending on the relative density of a targeted racial group, screening without an oversampling strategy could have an impact on survey resources. Third, some states permit people with drivers' licenses to exclude their names from distribution on DMV files provided to outside sources. This reduces the coverage of the general population to some extent. To the extent that persons who self-exclude differ from those in the general population, a potential for introducing bias into survey estimates exists unless one drops from analyses cases who indicate they have drivers licenses but have also "self-excluded". Fourth, some states will only provide a sample of records (e.g., 1% of the records are systematically selected) or the cost of obtaining the full file is prohibitive. If only a sample of records is obtained, the extent to which duplicate records on the full file is a concern will be uncertain. If only a sample of records is to be provided, it is important to work with DMV staff to ensure to the extent possible that the file from which the sample is selected has been unduplicated prior to sample selection. If one is requesting that DMV staff restrict the sample to selected counties, one should check that omissions have not resulted in the process. Finally, the broader coverage issues discussed earlier apply. If the coverage of the population, particularly the portion that experiences higher incidence rates for cases, is relatively low, using DMV files may not be as appealing as they might otherwise appear. For example, relatively low coverage of older age groups in a DMV file may be problematic if cases tend to be found in the older segment of the general population.

However, it should be pointed out that the use of files of Medicare Beneficiaries obtained from the CMS, often taken to be a "gold standard" for covering ages 65-79, may result in undercoverage of people in that age range. We have examined ratios of estimates based on samples from these CMS files for the county areas associated with Chicago and Detroit to corresponding 2000 Census figures. Data were obtained from five four-percent systematic samples selected at three-month intervals during 2002 and 2003. The average number of persons in the three age groups 65-69, 70-74, and 75-79 were computed and compared to 2000 Census figures. The results appear in Table 4 below.

Table 4. Ratios of population estimates, averaging across five quarterly CMS samples, to corresponding Census 2000 figures

	Ratios	
Age	Chicago	Detroit
65-69	.82	.88
70-74	.88	.91
75-79	.94	.98

The ratios in Chicago range from .82 to .94 and in Detroit from .88 to .98. Coverage does not appear as complete for the younger age ranges.

Of course, there are important advantages to using the CMS data. Age, sex, race, and address information are readily available without screening or oversampling designated areas. However, for many studies asking cases if they are a Medicare Beneficiary and excluding them if they are not is not undertaken, unlike the approach used for RDD or DMV methodologies (where cases that live in households without a standard telephone or without driver's licenses, respectively, are determined and eliminated from analyses). Since the 65 and older cases often make up the bulk of the cases of interest, consideration might be given to doing such screening if concerns exist about a lack of comparability.

Moreover, currently, CMS files are not being made available to studies, and this practice is likely to continue into the indefinite future. Thus, RDD or DMV files may be the only ways to obtain population-based controls for the 65 and older age group. This would entail much more screening for persons aged 65 and older for controls obtained via RDD. DMV files should be evaluated to determine if there are potential coverage issues for persons aged 65 or older.

There are some additional advantages to working with DMV files in the under 65 age range. If only age and sex are needed for matching cases to controls, oversampling is not an issue as both date of birth and sex are standard items on a DMV file. This simplifies the sample design considerably. Moreover, if it is suspected that location within a targeted area is correlated with the status of being a case, one can sort controls by a geographic indicator such as ZIP Code within the strata formed by

the matching criteria. This implicit stratification can potentially help increase the power to detect geography as a risk factor since controls will be distributed across the geographic area while cases may not be.

An issue related to the analysis of controls where oversampling of geographic areas is employed is worth mentioning. If sample weighting is employed, as is generally done in survey research settings other than case control studies, the design effects associated with oversampling can be reflected in the standard errors with an appropriate variance estimation structure developed for complex surveys (e.g., replication or Taylor series methodologies; see Korn and Graubard, 1999). If not, variables should be added to the analytic model to reflect this component of the sample design.

In conclusion, we find that for urban areas like Detroit and Chicago, where minority populations are highly clustered geographically, geocoding addresses from the DMV listing is a potential way to construct geographically-based sampling strata for oversampling population Black controls, even though race is not indicated on the DMV listing. Further empirical work is needed to determine the viability of establishing strata for oversampling minorities from DMV listings in other areas of the United States.

References

DiGaetano R. and Waksberg J. (2002). Commentary: Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology,* 155 (8), pp.771-775.

Korn, E.L. and Graubard, B.I., (1999). *Analysis of Health Surveys.* New York, NY:Wiley.

Waksberg, J., Judkins, D. and Massey, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology,* 23, pp. 61-71.