Toward a Better Estimation of Working Residential Number (WRN) Rate Among the Undetermined: An Application of Survival Analysis

Hiroaki Minato, Lidan Luo
NORC at the University of Chicago

1. Overview of the methods of working residential number (WRN) rate estimation

1.1 Introduction to WRN

A working residential number (WRN) rate is a percentage of working residential telephone numbers among all telephone numbers whose residential status is known. (Consider a telephone number as a survey sample here.) However, for some telephone numbers we may not be able to resolve their WRN status within a given set of calling rules. For example, we cannot know the WRN status of a number if all the calls for the number result in a combination of ring/no answers, busy signals, quick hang-ups, answering machine pickups, and other technical barriers where the number's residential status cannot be determined. If there are many telephone numbers with unknown WRN status, we may not wish to ignore them from computing a response rate. But, how we deal with these numbers can become an issue, because a higher estimate of the WRN rate among the undetermined deflates the response rate while a lower estimate inflates it. This paper examines the survival analysis estimation of the WRN rate among the unknown WRN status, *e*:

RR = WRN completing interviews / (Known WRN + *e* * Unknown WRN status).

1.2 CASRO method in estimating WRN rate among the undetermined

The Council of American Survey Research Organizations (CASRO) method, also called the "proportional allocation" method, assumes cases of unknown eligibility (WRN status) are eligible (WRNs) in the same proportion as cases for which WRN status can be determined. That is, it simply uses:

WRN rate = Known WRN / (Known WRN + Known nonWRN)

as the estimate of the WRN rate among the undetermined. Note that this method is equivalent to a weighted average of known WRN rates from existing call attempts, where the weights are the numbers of resolved calls from the respective call attempts. The CASRO method is naïve in the sense that the assumption suggests what we do not know systematically is like what we know systematically. The CASRO method would obviously be inappropriate when the WRN rate of the unknown WRN and the WRN rate of the known WRN are significantly different.

1.3 Survival analysis method (SAM) in estimating WRN rate among the undetermined

When there is a variation in the WRN rate over the required numbers of calls, the CASRO method's assumption tends to fail. (The existence of variation does not necessarily make the rate among the known and the rate among the unknown different, because each rate is a summary of multiple rates in an aggregated group.) For such a situation, there is an alternative to the CASRO method in estimating the WRN rate among the undetermined.

The survival analysis approach described by Brick, Montaquila, and Scheuren (2000, 2002) takes advantage of the relationship between the level of difficulty (number of call attempts) in reaching a household and the WRN rate. Provided that phone numbers are censored randomly, undetermined numbers at the end of data collection can be considered as right-censored observations. The survivor function for such data, which describes the probability of a number being resolved at each call attempt, can be partitioned into separate functions, WRN and nonWRN. It is a direct application of the survival analysis method with competing risks, and we call it SAM. Using the similar notation to that in Brick et al. (2000, 2002), the mode-specific survivor functions are written as

$$\hat{S}_{WRN}(t) = \sum_{t' \geq t} \frac{d_{WRN}(t')}{n(t')} \hat{S}(t')$$

and

$$\hat{S}_{nonWRN}(t) = \sum_{t' \geq t} \frac{d_{nonWRN}(t')}{n(t')} \hat{S}(t'),$$

where $d_{WRN}(t')$ and $d_{nonWRN}(t')$ are the numbers of cases resolved to be WRN and nonWRN at the t'-th call attempt, respectively, $n(t')$ is the number of cases available for the t'-th call attempt, and

$$\hat{S}(t) = \prod_{t' < t} \frac{n(t') - (d_{WRN}(t') + d_{nonWRN}(t'))}{n(t')}$$

is the Kaplan-Meier estimate of the marginal survivor function. And, the overall WRN rate is computed as

$$\hat{r}_{\infty} = \frac{\hat{S}_{WRN}(0)}{\hat{S}_{WRN}(0) + \hat{S}_{nonWRN}(0)}.$$

Finally, the WRN rate of the undetermined numbers is estimated as

$$\hat{r}_{UN} = \frac{\hat{r}_{\infty} \cdot n_{TOT} - n_{WRN}}{n_{UN}},$$

where n_{TOT} is the total number of cases, n_{WRN} is the number of cases resolved as WRN, and n_{UN} is the number of undetermined cases.

Why might SAM work better than the CASRO method? The survival analysis estimation requires an application of random censoring in making calls. Note that without censoring, the CASRO method and SAM give the same estimates. The censoring produces unresolved (surviving) cases during trial in addition to those unresolved cases originally occurring at the end of trial. Since we know the WRN rate during trial from the known WRN cases, we can use this information to estimate the WRN rate of the unknown cases during trial (call them *forced* unknown cases, because they could be resolved if the random censoring were not done) along with the original unknown cases.

If the original unknown cases at the end of trial are large and complex, neither SAM nor the CASRO method would do well, because the unknown cases are dominated by the truly unknown cases that may not be represented by the known or the forced unknown cases. It would be a general lack-of-data problem.

However, if, for example, we know the WRN rate would monotonically decrease beyond the call attempts of the trial, the best estimate would be the WRN rate of the trial's last call attempt. If, in addition, we know the rate of decrease, the estimate may be further adjusted by it.

1.4 A context—Making additional calls under a budget constraint

In some situations, a survey conductor needs to make additional calls to telephone numbers whose WRN status was unknown after the required number of call attempts has been reached. For example, more completed interviews may be needed without ordering or releasing more phone samples.

Perhaps, there are three standard options to pursue. One is to change the entire set of calling rules by increasing the maximum number of call attempts and to call all the undetermined cases a few more times. Two is to randomly select some cases from the undetermined and to call them a few more times. And, three is similar to two but is to target a "promising" subgroup of, say, soft refusal cases and to try to convert them with extra call attempts.

For simplicity, we consider only options 1 and 2 in order to evaluate the CASRO method and SAM.

The first comparison is option 1 with the CASRO estimation vs. option 2 with SAM.

Second, we compare option 1 with the CASRO estimation and option 2 with SAM, when the numbers of total calls and the number of total resolved numbers are the same. Note that the equal budget condition forces option 2 with SAM to have the larger number of maximum call attempts.

Finally, given option 2 with SAM, we vary the maximum number of call attempts, keeping the number of total calls and the number of total resolved numbers same, and compare them.

In this paper, we consider only the most basic models for the above comparisons and state some theoretical results. We also illustrate how one might conduct post-survey censoring, using the Year 2 call history data from the REACH 2010 Risk Factor Survey.

2. Comparisons of the CASRO approach and the survival analysis approach

2.1 Formulation

A basic premise is that when there is a non-constant relationship between the number of call attempts and the WRN rate, the more calls we attempt (at not-yet-resolved cases), the better overall picture of the WRN rate we get. And, obviously, the more known WRN's we have, the less unknown WRN's there are and the less significant the WRN rate among the unknown would become with respect to the response rate computation.

Suppose that a population consists of three types of phone numbers: the first type of phone numbers (t_1) can get resolved by one call, the second type (t_2) requires exactly two calls to be resolved, and the third type (t_3) three calls. Let r_1, r_2, and r_3 be the resolution rates for the first, second, and third call attempts, respectively. r_1 and r_2 must be greater than 0 but less than 1 so that t_1, t_2, and t_3 are all non-empty. r_3 is set to 1 so that the union of t_1, t_2, and t_3 is the entire population. The WRN rates for t_1, t_2, and t_3 are written as w_1, w_2, and w_3, respectively. They can be any value between 0 and 1, inclusively.

2.2 Full extension with CASRO vs. censored extension with SAM

For this population, consider the simplest initial set of calling rules, which attempts to call every phone number once. Assume some numbers are left unresolved after this rule is applied, and we are required to get more eligible phone numbers. The full-extension-with-CASRO method (F2-CAS) extends the rule by calling all the unresolved cases for the second time, while the censored-extension-with-SAM (F1C1-SAM) first censors out some of the unresolved cases at the rate c_1 and continues calling the rest for the second time. c_1 is assumed to be non-zero and non-one. For both F2-CAS and F1C1-SAM, no third call attempt is made and it is represented by $c_2 = 1$, i.e., censoring all unresolved numbers remaining after the second call. (See Table 1 in Appendix for a realization of the F1C1 calling rules in the maximum-three-call-attempt population.)

For F2-CAS, the unresolved phone numbers after the second call attempt consist of the entire and only t_3, $n_1(1-r_1)(1-r_2)$, whose true WRN rate is w_3. For F1C1-SAM, the undetermined cases include t_3 and some of t_2, $n_1(1-r_1)c_1 + n_1(1-r_1)(1-c_1)(1-r_2)$. The true WRN rate for these undetermined cases is summarized by

$$\frac{(1-r_1)c_1r_2w_2 + (1-r_1)c_1(1-r_2)w_3 + (1-r_1)(1-c_1)(1-r_2)w_3}{(1-r_1)c_1r_2 + (1-r_1)c_1(1-r_2) + (1-r_1)(1-c_1)(1-r_2)}.$$

Using the true parameters, F2-CAS computes the WRN rate for its undetermined cases by

$$\frac{n_1r_1w_1 + n_1(1-r_1)r_2w_2}{n_1r_1 + n_1(1-r_1)r_2}.$$

Thus, the error is

$$E_{F2-CAS} = \frac{r_1w_1 + (1-r_1)r_2w_2}{r_1 + (1-r_1)r_2} - w_3.$$

(This actually is a bias, but we call it "error" in a generic way.)

F1C1-SAM computation is more complicated, but the estimate (in terms of the true parameters) and its error from the truth, $E_{F1C1-SAM}$, can be formulated explicitly. Our first theorem is:

Theorem 1. $|E_{F2-CAS}| \ge |E_{F1C1-SAM}|$ for all w_1, w_2, and w_3 in [0, 1], whenever $n_1 > 0$, $0 < r_1 < 1$, $0 < r_2 < 1$, $r_3 = 1$, $0 < c_1 < 1$, and $c_2 = 1$.

Proof. We can prove this directly by algebra. First, compute and write that $E_{F1C1-SAM} = a(bw_1 + (1-b)w_2 - w_3)$ where $a = (1-r_2)/(c_1r_2 + 1 - r_2)$ and $b = r_1/(r_1 + r_2 - r_1r_2)$. Next, note that $E_{F2-CAS} = bw_1 + (1-b)w_2 - w_3$ and then that $E_{F1C1-SAM} = aE_{F2-CAS}$. Since $0 < a < 1$, we have $|E_{F1C1-SAM}| = |aE_{F2-CAS}| = a|E_{F2-CAS}| \le |E_{F2-CAS}|$, where the equality in the last $\le$ holds only when $E_{F2-CAS} = E_{F1C1-SAM} = 0$. »«

It basically means that in the current context the SAM calculation of the WRN rate among the undetermined with censored cases would give a smaller bias than the CASRO calculation of the WRN rate among the undetermined without censored cases. (Note that the above calculations are hypothetical because the true parameters are assumed to be known.) When the w_1, w_2, w_3, r_1, and r_2 parameters are estimated from data without bias, we can consider the expected values of E_{F2-CAS} and $E_{F1C1-SAM}$ and the similar result would hold. However, the standard errors of those errors would have to be discussed in the future.

This theorem can be generalized in various ways. For example, we can set the initial set of calling rules to be up to five calls. Or, we can lengthen the unobserved tail. These generalizations should carry over the truth-ness of the current theorem. However, generalizing the censoring part can be tricky. We consider one such generalization in the next section.

2.3 Full extension with CASRO vs. censored extension with SAM—With the same numbers of total additional calls and the same numbers of total additionally resolved calls

First, we note that, in the previous population setting, if we set $0<c_1<1$, $0<c_2<1$, and $c_3=1$ for the censored-extension-with-SAM (write F1C2-SAM), i.e., if we censor after the second call attempt as well as the first call attempt, then we get $E_{F1C2-SAM}=0$.

This might indicate how and why the censored-extension-with-SAM works. But, the scenario is unrealistic, because we are allowing calls to the part of population that we assume we cannot normally reach. In fact, since this subpopulation is characterized by setting the resolution rate to 1, $r_3=1$, if we apply the full-extension to them, all cases will be resolved and no undetermined cases will be left for WRN rate estimation. The number of total WRN's in the population would equal the number of known WRN's in F1C2 plus the number of SAM-estimated WRN's from the undetermined, because the SAM estimation is perfect (unbiased and no sampling error) in this case.

For a scenario we consider in this section, we expand the population by adding a subpopulation that would require four call attempts to be resolved. We assume this subpopulation has a resolution rate of 1, $r_4=1$. And, it is now assumed that $0<r_3<1$.

Further, we let the number of calls represent cost[1]. Then, F1C1 requires less extra cost than F2. However, the number of additionally resolved cases from F1C1 would be smaller than the number of additionally resolved cases from F2. In order for the censored extension to increase the number of extra resolved cases to match that from the full extension, it would have to call some of those unresolved cases from the second call attempt, within the given cost constraint. For this third call attempt, we may or may not censor. But, in this comparison we force a censoring before the third call attempt, too, and set $0<c_1<1$, $0<c_2<1$, and $c_3=1$. Under this scenario, we compare F1C2-SAM and F2-CAS.

We start with the same initial set of calling rules described in 2.1. And, assume again some numbers could not be resolved by the rule and that we are going to extend it so that we can get more eligible phone numbers. The full extension (F2) is unchanged as a benchmark and sets the number of additionally resolved cases and the extra budget (the number of extra calls):

The number of extra calls = $n_1(1-r_1)$

The number of resolved cases from the extra calls = $n_1(1-r_1)r_2$.

For F1C2, the number of additionally resolved cases[2] is written as:

$$n_1(1-r_1)(1-c_1)r_2+n_1(1-r_1)(1-c_1)(1-r_2)(1-c_2)r_3,$$

which has to be equal to[3] $n_1(1-r_1)r_2$ from F2. The number of extra calls[4] from F1C2 is:

$$n_1(1-r_1)(1-c_1)+n_1(1-r_1)(1-c_1)(1-r_2)(1-c_2),$$

which is restricted to be equal to[5] $n_1(1-r_1)$ from F2.

Suppose we use the CASRO method to compute the WRN rate among the undetermined from F2 and use SAM to compute the WRN rate among the undetermined from F1C2. We *would* hope that the error in F1C2-SAM is smaller than or equal to the error in F2-CAS for all values of r_1, r_2, and r_3 in (0, 1) and for all values of w_1, w_2, w_3, and w_4 in [0, 1].

Conjecture 2. $|E_{F2-CAS}|\geq|E_{F1C2-SAM}|$ for all w_1, w_2, w_3, and w_4 in [0, 1], whenever $n_1>0$, $0<r_1<1$, $0<r_2<1$, $0<r_3<1$, $r_4=1$, $0<c_1<1$, $0<c_2<1$, and $c_3=1$ such that $n_1(1-r_1)(1-c_1)+n_1(1-r_1)(1-c_1)(1-r_2)(1-c_2)=n_1(1-r_1)$ and $n_1(1-r_1)(1-c_1)r_2+n_1(1-r_1)(1-c_1)(1-r_2)(1-c_2)r_3=n_1(1-r_1)r_2$.

Discussion. This conjecture is actually false. The relation between $|E_{F2-CAS}|$ and $|E_{F1C2-SAM}|$ depends on the parameters w_1, w_2, w_3, w_4, r_1, r_2, r_3, c_1 and c_2. We examine $D=|E_{F2-CAS}|-|E_{F1C2-SAM}|$. (An alternative would be to study the more concise but complicated formula, the ratio of the errors $R=|E_{F2-CAS}/E_{F1C2-SAM}|$, assuming $R\geq 0$ if $E_{F1C2-SAM}=0$ with the equality if and only if $E_{F2-CAS}=0$ and $E_{F1C2-SAM}=0$.)

First, we note that the two conditions among n_1, r_1, r_2, r_3, c_1 and c_2 in the conjecture may be simplified to $r_3=r_2$ and $r_2=1-c_1/((1-c_1)(1-c_2))$. The first equation rather restricts the class of applicable populations. But, for the current scenario to be true, we must impose it here. The second equation can be solved also for c_1 or c_2 and, depending on what we substitute, *D* can be represented by a different set of parameters. We choose the above substitutions so that *D* contains only w_1, w_2, w_3, w_4, r_1, c_1, and c_2. We modify Conjecture 2 as follows.

Theorem 2. Suppose that w_1, w_2, w_3, and w_4 are in [0, 1] and that $n_1>0$, $0<r_1<1$, $0<r_2<1$, $0<r_3<1$, $r_4=1$, $0<c_1<1$, $0<c_2<1$, and $c_3=1$. Also, suppose that $r_3=r_2$ and $r_2=1-c_1/((1-c_1)(1-c_2))$. Then, for any w_1, w_2, w_3, w_4, and r_1, there exist some c_1 and c_2 such that $|E_{F2-CAS}|=|E_{F1C2-SAM}|$. Further, for some w_1, w_2, w_3, w_4, and r_1, there exist some c_1 and c_2 such that $|E_{F2-CAS}|>|E_{F1C2-SAM}|$.

[1] The number of calls is only an approximation of actual cost, which usually depends on interview-time length rather than how many calls.

[2] The number of total resolved cases is this plus $n_1 r_1$ for either F1C2 or F2.

[3] We can weaken this condition by replacing "equal to" with "greater than or equal to". But, identification becomes even more difficult.

[4] The number of total calls is this plus n_1 for either F1C2 or F2.

[5] We can weaken this condition by replacing "equal to" with "less than or equal to". But, identification becomes even more difficult.

Proof. To show that the first statement is true, we define $D = |E_{F2-CAS}| - |E_{F1C2-SAM}|$ and solve it for arbitrary values of w_1, w_2, w_3, w_4, and r_1. It turns out that D = 0 if[6] $c_2 = 1 - c_1/(1-c_1)$. The function is nonlinear but monotonically decreasing with the domain (c_2) being (0, 1) and the range (c_1) being (0, 0.5).

For the second assertion, consider $c_2 = 1 - c_1/(1-c_1) + k$. For some w_1, w_2, w_3, w_4, and r_1, we can find k such that $|E_{F2-CAS}| > |E_{F1C2-SAM}|$. We simply give an example. Take w_1 = 0.6, w_2 = 0.4, w_3 = 0.5, w_4 = 0.2, and r_1 = 0.7. And, set c_1 = 0.2 and k = -0.1 and thus c_2 = 0.625. Then, $0.2924675325 \approx |E_{F2-CAS}| > |E_{F1C2-SAM}| \approx 0.2660929434$. »«

We note that D is normally not monotonic with respect to k. Monotonicity may occur in special cases when some or all of w_1, w_2, w_3, and w_4 take the boundary values. Also, notice that for some w_1, w_2, w_3, w_4, and r_1, there do not exist any c_1 and c_2 such that $|E_{F2-CAS}| > |E_{F1C2-SAM}|$. For example, if w_1, w_2, w_3, and w_4 are all zeros, then $|E_{F2-CAS}| = |E_{F1C2-SAM}| = 0$ for all r_1, c_1, and c_2. The next step may be to specify w_1, w_2, w_3, w_4, and r_1, for which there *always* exist c_1 and c_2 such that $|E_{F2-CAS}| > |E_{F1C2-SAM}|$. Then, the theorem may become more constructive.

In general and practice, c_1 needs to be determined only some time before all the second call attempts are made, and c_2 before all the third call attempts are completed. And, r_1, r_2, r_3, w_1, w_2, and w_3 are all estimable from the call history data, even during the calling period. For example, r_1 and w_1 may be estimated from the call history data after some first call attempts are made[7]. The only truly unknown parameter is w_4, and D can be studied over $w_4 \in [0,1]$ for given censoring rates c_1 and c_2 and for given estimates of r_1, r_2, r_3, w_1, w_2, and w_3.

2.4 Large and short censored extension with SAM vs. small and long censored extension with SAM—With the same numbers of total additional calls and the same numbers of total additionally resolved calls

Say, in order to increase the resolved and thus eligible cases, you decide to take option 2, i.e., to randomly select some cases from the undetermined and to call them a few more times, and to use SAM to estimate the WRN rate among the undetermined. You also have a fixed budget to extend your set of calling rules. A question is how you distribute extra calls.

We compare two strategies for the same population assumed in 2.3. The first strategy is to call a subset of remaining unresolved numbers only one more time. The second strategy is to allow up to two more calls to resolve a subset of remaining unresolved numbers. The latter subset must necessarily be smaller than the former subset in order to save some budget for calling some still-unresolved numbers for the second time. The budget, the total number of calls, is kept constant. Also, the total numbers of resolved cases must be the same.

The first strategy may be specified by setting $0 < c_{1,1} < 1$ and $c_{1,2} = 1$; the second strategy by $0 < c_{2,1} < 1$, $0 < c_{2,2} < 1$, and $c_{2,3} = 1$. And, the constraints are:

$$n_1(1-r_1)(1-c_{1,1}) = n_1(1-r_1)(1-c_{2,1}) + n_1(1-r_1)(1-c_{2,1})(1-r_2)(1-c_{2,2})$$

and

$$n_1(1-r_1)(1-c_{1,1})r_2 = n_1(1-r_1)(1-c_{2,1})r_2 + n_1(1-r_1)(1-c_{2,1})(1-r_2)(1-c_{2,2})r_3.$$

One might conjecture that the estimation error of the first strategy (F1C1-SAM) is greater than or equal to the error of the second strategy (F1C2-SAM).

Conjecture 3. $|E_{F1C1-SAM}| \ge |E_{F1C2-SAM}|$ for all w_1, w_2, w_3, and w_4 in [0, 1], whenever $n_1 > 0$, $0 < r_1 < 1$, $0 < r_2 < 1$, $0 < r_3 < 1$, $r_4 = 1$, $0 < c_{1,1} < 1$, $c_{1,2} = 1$, $0 < c_{2,1} < 1$, $0 < c_{2,2} < 1$, and $c_{2,3} = 1$ such that

$$n_1(1-r_1)(1-c_{1,1}) = n_1(1-r_1)(1-c_{2,1}) + n_1(1-r_1)(1-c_{2,1})(1-r_2)(1-c_{2,2})$$

and

$$n_1(1-r_1)(1-c_{1,1})r_2 = n_1(1-r_1)(1-c_{2,1})r_2 + n_1(1-r_1)(1-c_{2,1})(1-r_2)(1-c_{2,2})r_3.$$

Discussion. This conjecture basically says that, given the same number of extra calls and the same number of additionally resolved calls, selecting a smaller number of unresolved cases and spreading them over more call attempts would give a less-biased estimate of the WRN rate of the undetermined than selecting a larger number of unresolved and spreading them over fewer call attempts would.

However, we know that we can easily find a counter example to the general conjecture. Take, for example, $w_1 = 0.8$, $w_2 = 0.6$, $w_3 = 0.5$, $w_4 = 0.8$, $r_1 = 0.5$, $r_2 = 0.3$, $r_3 = 0.3$, $r_4 = 1$, $c_{1,1} = 0.2$, $c_{1,2} = 1$, $c_{2,1} = 0.3$, $c_{2,2} = 39/49$, and $c_{2,3} = 1$. Then, $E_{F1C1-SAM} \approx 0.0404$ and $E_{F1C2-SAM} \approx -0.0525$. The conjecture could be true if we constrain some of the parameters. At this moment, we believe that this conjecture would be true, at least if the w_1, w_2, w_3, and w_4 are constrained as $1 \ge w_1 \ge w_2 \ge w_3 \ge w_4 \ge 0$ or $0 \le w_1 \le w_2 \le w_3 \le w_4 \le 1$.

3. A Post-survey experiment on the call history data from the Year 2 REACH 2010 Risk Factor Survey

3.1 Censoring after the survey is completed

If disposition codes of individual calls are recorded in call history data, the censoring and survival analysis method can be applied to the call history data retrospectively, i.e., after the survey data have been collected.

[6] There are other conditions that would give D = 0, e.g., $w_1 = w_2 = w_3 = 0$.

[7] The second call attempt might be started before all cases are called once. But, if enough first calls have been made, the r_1 and w_1 could be estimated reasonably well.

The REACH 2010 Risk Factor Survey has two-stage sample design. The first stage conducts screening on the telephone samples for WRN numbers as well as some other survey-specific eligibility (race/ethnicity and geography). The result of each call is classified into one of 56 disposition codes: e.g., "Business/Gov. Dorm", "Call-Ring No Answer", "Call-Privacy Blocker", "Hostile Refusal (English)", "Appointment", "Complete", etc. Only telephone numbers that are identified as WRN and are racially/ethnically and geographically eligible become the sample for the second stage, the household member interview.

To apply the survival analysis method, the first step is to classify these call-level disposition codes into WRN status known and WRN status unknown. WRN status known can be further classified into WRN and nonWRN. Only a phone number's call history until its WRN status is resolved is necessary for estimating the WRN rate. Thus, the second step is to trim the call history data so that each number's call history ends with the first WRN-status-known disposition code, if any. Some numbers may display all WRN-status-unknown disposition codes in their histories; then, their WRN statuses are undetermined. (For the screening, the REACH survey used a up-to-seven-call rule. However, for some special cases, more than seven call attempts were made, but those cases are dropped from the current analysis.)

The next step is to censor. Censoring can be either forward censoring and backward censoring. The forward censoring imitates an actual set of calling rules that implements censoring. To begin with, note that we normally want every phone number to be called at least once and thus that we do not censor before the first call attempt. To censor after the first call attempt, collect all those cases that were not resolved at the first call attempt and randomly censor $c_1 \times 100$ % of them. To censor after the second call attempt, collect all those cases that were called for the second time but were not resolved and randomly censor $c_2 \times 100$ % of them. Repeat the procedure until the second call attempt from the maximum call attempt; all unresolved numbers after the last call are automatically censored. We must note that the choice of c_i completely determines the composition of the undetermined and thus the estimate of the WRN rate among the undetermined.

The backward censoring is a purely retrospective method and it is possible only after the survey. First, group the cases by the number of calls that required to resolve the WRN status. The cases that were never resolved are excluded from the censoring. In a given group, each case's call history is truncated after a random call-attempt point. For example, in a group of cases that were resolved at the third call, each case may be truncated after the first or second call attempt with some prescribed probabilities. (The sum of the probabilities should be less than 1; otherwise, we lose all resolved-at-the-third cases.) This option can control the composition of the censored cases in a straightforward manner. Note, however, that the original undetermined cases (i.e., the cases that were never resolved at or before the maximum call attempt) remain unchanged and unknown. From the censored call history data, when properly aggregated, the Kaplan-Meier estimates of the survivor functions can be computed by hand (see Brick et al. (2002)) or by SPSS or SAS. And, finally, the survival analysis estimate of the WRN rate of the undetermined with censoring can be computed.

3.2 Applying the survival analysis method to estimate the WRN rate among the undetermined

The REACH 2010 Risk Factor Survey targets racial and ethnic groups in certain communities with specific geographies in the U.S. Samples of this survey come both from listed and unlisted RDD telephone numbers. Known business and disconnected numbers are removed a priori and are not actually called for screening. We implemented a forward uniform censoring method with constant censoring rates of 3 % and 10 % for the call history data from nine of the REACH communities. (The SPSS syntax program is available upon request.)

Table 2 in Appendix shows the survival analysis estimates of the WRN rates among the undetermined for two sample types and nine communities. (The sampling weights are not used here.) The CASRO estimates of the WRN rates among the determined/undetermined are also computed. Comparative scenarios are: (1) an initial set of calling rules is to call every number once, but the rule is extended to call all unresolved numbers up to six more times and (2) an initial set of calling rules is to call every number once, but the rule is extended to call, up to six more times, a subset of unresolved numbers randomly selected at a constant rate after each call attempt. (The first scenario is equivalent to have an initial set of calling rules of calling every unresolved number up to seven times and to stick with the rule.) Theorem 1 should imply that the second scenario gives at most as big an estimation bias as the first scenario. (The varying sampling errors are ignored at this moment.)

4. Conclusion

What is left for future research is plenty. Theorem 1 should be formally generalized. Theorem 2 must be further refined. Conjecture 3 must be conditioned and proved.

In addition to the estimation bias, we need to understand the estimation error due to sampling, including the random censoring. Namely, the standard error of the WRN rate among the undetermined could be quite complicated.

Once both variance and bias are better understood, the evaluation of the survival analysis method in this area would become more conclusive.

Appendix

Table 1: A realization of a F1C1 calling rules in the maximum-three-call-attempt population

Call Attempt	Call Size	Resolution Rate	Resolved	Working Rate	Working	Non-working	Unresolved	Censoring/ Subsample Rate	Censored/ Undetermined
1	n_1	r_1	$n_1 r_1$	w_1	$n_1 r_1 w_1$	$n_1 r_1 (1-w_1)$	$n_1(1-r_1)$	c_1	$n_1(1-r_1)c_1$
2	n_2 $= n_1(1-r_1)(1-c_1)$	r_2	$n_2 r_2$	w_2	$n_2 r_2 w_2$	$n_2 r_2 (1-w_2)$	$n_2(1-r_2)$	$c_2 = 1$	$n_2(1-r_2)$
3	n_3 $= n_2(1-r_2)(1-c_2) = 0$	r_3	$n_3 r_3$	w_3	$n_3 r_3 w_3$	$n_3 r_3 (1-w_3)$	$n_3(1-r_3)$		
Total	$n_1 + n_2 + n_3$		$n_1 r_1 + n_2 r_2 + n_3 r_3$		$n_1 r_1 w_1 + n_2 r_2 w_2 + n_3 r_3 w_3$	$n_1 r_1 (1-w_1) +$ $n_2 r_2 (1-w_2) +$ $n_3 r_3 (1-w_3)$	$n_1(1-r_1) +$ $n_2(1-r_2) +$ $n_3(1-r_3)$		$n_1(1-r_1)c_1 +$ $n_2(1-r_2)$

Table 2: Survival analysis estimates of the WRN rates of the undetermined listed and unlisted RDD telephone numbers from 9 REACH communities

The City to Which the REACH Community Belongs	Sample Type	Sample Size	CASRO	Survival Analysis Method	
			Uncensored	3% Censored	10% Censored
Boston, MA	Listed	3577	81.76%	82.43%	84.38%
	Unlisted RDD	1965	81.87%	81.98%	82.94%
Chicago, IL	Listed	700	76.80%	76.36%	77.62%
	Unlisted RDD	1831	60.01%	60.50%	61.16%
Bronx, NY	Listed	1170	69.27%	68.82%	70.11%
	Unlisted RDD	2099	78.24%	78.40%	78.33%
Nashville, TN	Listed	1828	65.71%	67.72%	69.40%
	Unlisted RDD	8380	27.47%	28.72%	31.36%
Charleston, SC	Listed	5974	80.47%	81.68%	84.13%
	Unlisted RDD	5312	45.90%	47.37%	51.00%
New Orleans, LA	Listed	2733	67.46%	69.61%	72.85%
	Unlisted RDD	3915	42.04%	43.20%	45.72%
Portland, OR	Listed	7984	84.96%	86.19%	87.76%
	Unlisted RDD	12944	34.23%	36.83%	41.12%
Kansas City, MO	Listed	911	75.09%	77.24%	80.76%
	Unlisted RDD	7630	36.16%	37.59%	41.50%
Las Vegas, NV	Listed	3567	81.56%	81.80%	82.23%
	Unlisted RDD	16498	49.94%	50.52%	51.63%

Acknowledgement
We would like to thank the session's discussants, Michael Cohen and Victoria Albright, for their helpful comments.

References

American Association for Public Opinion Research (2004), *Standard Definitions, Final Dispositions of Case Codes and Outcome Rates for Surveys: RDD Telephone Surveys, In-Person Household Surveys, Mail Surveys of Specifically Named Persons*, Lenexa, KS: Author.

Brick, J. M., Montaquila, J., and Scheuren, F. (2000), "Estimating Residency Rates for Undetermined Telephone Numbers," in *American Statistical Association Proceedings of the Survey Research Methods Section*, pp. 1045-1050.

--(2002), "Estimating Residency Rates for Undetermined Telephone Numbers," *Public Opinion Quarterly*, 66, 18-39.

Council of American Survey Research Organizations (1982), *On the Definition of Response Rates: A Special Report of the CASRO Task Force on Completion Rates*, Port Jefferson, NY: Author.

Durand, C., Chevalier, S., and Vachon, S. (1998), "Using Survival Regression to Predict Occurrence of First Contact, Outcome at First Contact and Completion of Interview in Telephone Surveys," unpublished paper presented at the 14th World Congress of Sociology, Montreal, August, 1998.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data* (2nd ed.), Hoboken: John Wiley.

Smith, T. W. (2003), "A Review of Methods to Estimate the Status of Cases with Unknown Eligibility: Report Prepared for the AAPOR Standard Definitions Committee," unpublished paper to be presented at AAPOR, Phoenix, May, 2004.