Implementation of an Efficient Sampling and Case Processing Method

Shelton Jones[1], Tenbroeck Smith[2], Katherine Treiman[1], Mai Nguyen[1], Connie Hobbs[1], Alyssa Troeschel[2]
[1]RTI International, Research Triangle Park, NC 27709
[2]American Cancer Society, Atlanta, GA 30303

Abstract
We describe a method for sampling and processing cases that enables researchers to efficiently and collaboratively build sampling frames in a multisite study. This method builds the sampling frames by using data from multiple data sources, thus introducing additional variables for stratification, clustering, or both. It applies the No Personal Identification Disclosed (NOPID) approach (Jones et al., 2011) to ensure respondent confidentiality, while attempting to maximize response rates and reduce measurement error. This method was used for the Patient Reported Outcomes Symptoms and Side-Effects Study (PROSSES) funded by the American Cancer Society to study cancer patients' symptom experiences. To maximize accurate recall of symptom experiences during curative treatment, every month we sampled cross-sections of new cancer patients and mailed them recruitment materials. PROSSES, which was conducted with 17 Cancer Centers dispersed across the United States, achieved an overall response rate of approximately 60%. We describe the key steps to successfully fielding PROSSES on the basis of the NOPID method as well as the challenges faced and the lessons learned.

Key Words: Confidential sampling, NOPID sampling, multisite frame construction

1. Introduction

Research in the United States must function within the guidelines of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The act ensures the protection of human subjects as determined by the U.S. Department of Health and Human Services, which develops regulations protecting the privacy and security of certain health information. The two rules covered by HIPAA are

- the HIPAA Security Rule, which establishes security standards for the protection of electronic health information; and
- the HIPAA Privacy Rule, which establishes standards for privacy of individually identifiable health information.

Researchers are required to show strict compliance with the Privacy Rule by doing one or more of the following:

- The researcher must receive written permission from individuals in the form of an authorization before conducting the research, or

- the researcher must receive a waiver of the authorization for a data-use agreement for research using information on deceased persons, or

- the researcher can obtain a de-identified file containing health information, because de-identified health information is not considered protected health information (PHI) and is not protected by the Privacy Rule.

This research is based on making full use of the de-identified file for research purposes. We will note that the following 18 identifiers must be removed for the file to be considered de-identified and in compliance with the HIPAA Privacy Rule:

- name of the patient
- geography smaller than state
- all dates except year
- telephone numbers
- fax numbers
- e-mail addresses
- Social Security Numbers
- medical record numbers
- health plan beneficiary numbers
- account numbers
- certificate numbers
- vehicle identifiers
- device identifiers
- Web URLs
- IP address numbers
- biometric identifiers
- full-face photo images
- other unapproved IDs

Constructing sampling frames, designing surveys, and selecting probability samples are challenging in the absence of this key design information. However, the NOPID (No Personal Identification Disclosed) approach (Jones et al., 2011) introduced a sampling, data processing, and analysis method that enables the researcher to complete each of these tasks in the absence of the 18 key identifiers. The NOPID approach was first described in both rounds of the National Cancer Institute Community Cancer Centers Pilot Program (NCCCP). The basic requirements for the method are given below:

1. Each entity obtains approval from its institutional review board (IRB).

2. The HIPAA-covered entity maintains personally identifiable information (PII) and PHI.

3. The HIPAA-covered entity removes PII to de-identify records and forwards to the non-covered HIPAA entity.

4. The non-covered entity stores the de-identified PHI for stratification and sampling.

5. The non-covered entity selects and forwards the sample to the covered entity to contact the patient sample.

6. The sampled patients respond only to the non-covered entity.

For the Patient Reported Outcomes Symptoms & Side–Effects Study (PROSSES), we applied the NOPID method to conserve study resources and to ensure confidentiality of study participants. The objectives of PROSSES were to obtain accurate recall of cancer patients' symptom experiences during curative treatment. To bolster the sample size, we collected data for 13 monthly cross-sectional data collection periods. The target population consisted of cancer patients who received treatment for either breast cancer or colon cancer from one or more of the 17 participating Cancer Centers. Only participants who were age 21 or older and treated in Stages I–III were eligible to participate.

Sampling and data collection of PROSSES involved considerable collaboration between the American Cancer Society (ACS), the American College of Surgeons (ACoS), the 17 Cancer Centers, and RTI International. ACS screened and filtered the files from the Rapid Quality Reporting System and delivered the files to RTI to construct the sampling frames, populate the study's patient control system, prepare patient packets for the centers, and collect the survey data from study participants. **Section 2** highlights the importance of collaboration among all entities in order for the study to be completed successfully. **Sections 3** and **4** describe how confidentiality can be maintained even as patient materials are prepared and delivered to the patients. **Section 5** describes how the monthly sampling frames are built and processed. **Sections 6** and **7** list the variables recorded for processing and sampling purposes as well as the tracking of nonresponding patients. In **Section 8**, we discuss our data collection eligibility and response rates for the Cancer Centers, and **Section 9** shows the modal choices of the patients and how they varied from center to center

2. Collaboration Among All Entities

RTI worked with the ACS and the ACoS, key partners in the study, and were responsible for coordinating all study activities with the participating Cancer Centers. Key factors contributing to the successful implementation of the study were establishing clear roles and responsibilities for each entity, communicating frequently, and identifying and addressing issues on a timely basis.

Some operational issues occurred, particularly early in the data collection period with the first few samples. For example, there were some issues with the study IDs and inclusion of some ineligible cases in the samples. RTI worked closely with the participating Cancer Centers to review each sample and identify those types of issues. RTI immediately reported issues to the ACS and ACoS, reviewed the issues, and worked with these partners to reach resolution. Refinements were made to the sampling files as needed.
Data collection was closely monitored and weekly data collection reports were generated. The reports provided the numbers of completed surveys and the response rate. In weekly meetings, RTI and the ACS reviewed the data collection activities and determined whether any adjustments were needed to the survey schedule (e.g., release of samples).

Over the course of the study there were some changes to the design, including the addition of two Cancer Centers. RTI was responsive to these changes, working with the additional Cancer Centers to get them started with data collection. Our goal was to meet ACS and study needs as they arose, for example by providing an interim data set.

Some of the lessons learned are given below:

1. With multiple players involved in a complex study, working as a team and ensuring open and frequent communication was critical to success.
2. It was essential for RTI to play a hands-on role with the participating Cancer Centers. Study team members established communication with liaisons at each Cancer Center from the start and provided both formal training and one-on-one support as needed throughout the study.
3. The study required careful budget monitoring and several modifications to the scope of work and budget based on changes to the study (e.g., addition of two Cancer Centers, an increase in the sample size).

3. Maintaining Patient Confidentiality

Protection of the privacy rights of the patients was the most important feature of this study. Data collection procedures were designed to allow only the Cancer Centers access to identifying patient information, whereas RTI had access only to the de-identified patient data. Data de-identification is the core principle of the NOPID sampling and analysis approach. RTI did not have access to PII, and the Centers did not have access to patient survey responses. All study procedures were reviewed and approved by the National Cancer Institute IRB, RTI IRB, and the local IRBs before any involvement with the patients.

Challenges existed, but solutions were found. For example, to ensure patient confidentiality, we established transparent communication between entities. The close communication between RTI and the Centers and everyone's willingness and dedication to implement and adhere to specified procedures contributed to the overall success in maintaining patient confidentiality. Only a study ID number was used to identify each patient. This ID number was used on the paper survey and in correspondence between RTI and the Centers.

Although signed consent was waived, a consent sheet was included in all survey packets mailed to the patients by RTI. This form provided study information and provided IRB contact numbers as well as a toll-free study number for participants. RTI established this dedicated toll-free study telephone line for study participants to call if they had questions or wished to opt out of the survey. The voicemail message was carefully crafted to include instructions for the sample member to leave a message without the need for a call-back and instructions for a requested call-back. The messages asked that the caller use the study ID number found on their survey, not their name, in the message.

Occasionally, notes written on a survey mailed by a family member, or a message left on the toll-free number voicemail, informed the researchers of the death of the selected patient. (In a few instances the patient name was used instead of the study ID.) The Cancer Centers and RTI followed a protocol in reporting the death of a selected patient. At no time was the patient name included in an e-mail. Notification was done via personal telephone contact between the Centers and RTI. It was important that RTI learn of any deaths or opt-outs that were received by Centers so no further follow-up surveys would be sent to those patients.

RTI staff manually reviewed all open-text fields of keyed surveys to ensure that no names of patients, family members, or friends were found. Data entry staff were instructed to omit such names from their keying operations. The manual review was performed as another layer of quality assurance. Any discovered name was removed from the survey data file.

4. Preparing and Delivering Patient Materials

RTI prepared patient packets for each USID indicated in each of the 13 de-identified sample files received from ACS. Boxes containing these initial patient packets were shipped via FedEx to all applicable Centers within about 5 weekdays when the sample was released. Center staff responsible for matching each packet to the sampled patient via the USID number, generating the patient notification letter, inserting it into the packet, generating/affixing address labels, and mailing the packets to the patients.

The study protocol also stipulated that two follow-up mailings be sent by Centers to nonrespondents. RTI determined the nonrespondents for each sample on the basis of a 21-business day window after the previous mailout, where feasible. Care was taken to avoid an overlap of deliveries of patient packets to Centers in the same week that represented different samples and types of mailings. For example, if a second follow-up appeared to be due for a particular sample and we also had an initial mailing scheduled, we prioritized the initial mailing and delayed the shipment of packets for the follow-up mailout. Allowing this flexibility in follow-up mailouts eliminated possible confusion at the Centers that could have resulted in errors.

5. Processing Monthly Sampling Frames

A total of 13 monthly files were received from ACS over the data collection period. The first file was received in March of 2012 and contained patients diagnosed from January 2011 through February 2012 for a total of 14 months. This first file contained 1,063 total eligible cancer patients, or 24.4% of the 4,359 patients were received on the first month of data collection. **Figure 1** shows the variability in patient counts over the monthly periods. Although a distant second, in October 2012, we received 561 (12.9%) patients. The smallest number of patients (104, or 2.4%) was received in the June 2012 subsample, which also had the lowest response rate (49%).

Upon receiving the monthly data files from ACS, we confirmed that the files only contained the two data fields (NCCCPsiteID and USID) necessary for frame construction. The NCCCPsiteID was the identifier of the Cancer Center and the USID was the unique patient identifier within the Cancer Center. We also confirmed that all patient USIDs were unique for each Cancer Center. The files were loaded into the study's patient control system for continuous monitoring and tracking. An independent backup Master Frame containing all incoming patient and center IDs was also created to ensure quality. There was no subsampling of eligible patients for this study. That is, all eligible patients were invited to participate in the study with certainty.

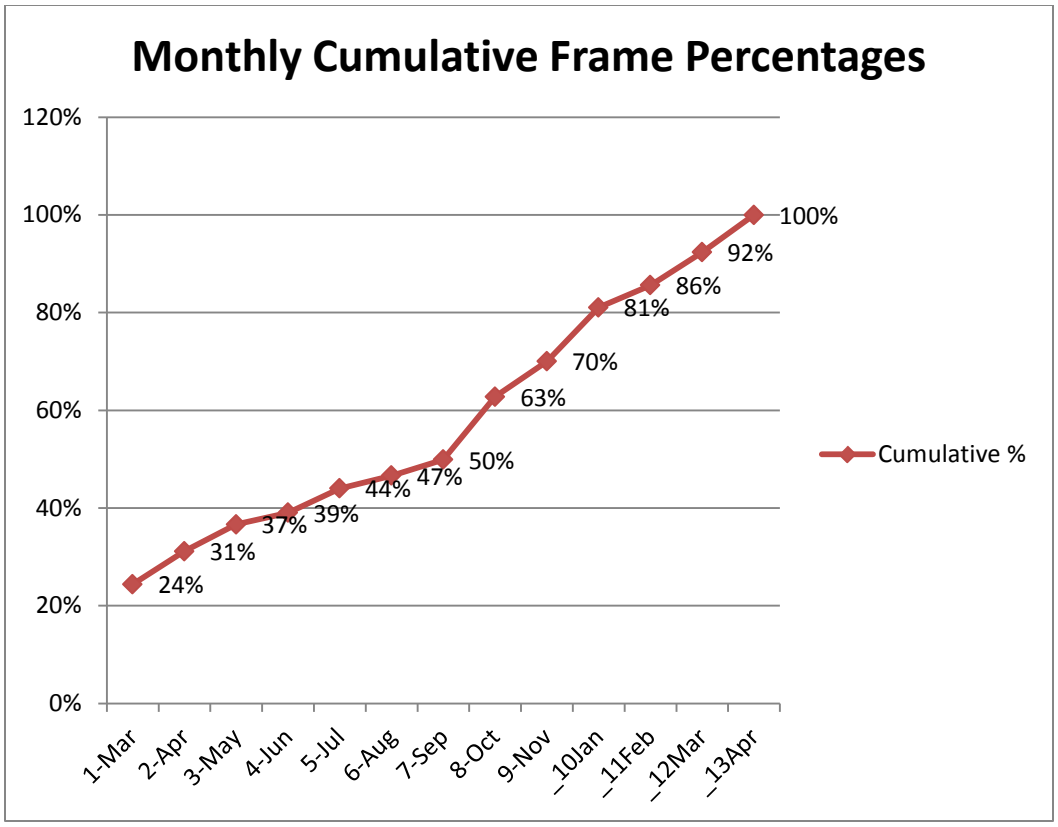

Figure 1: Sampling Frame Patient Counts by Data Collection Month

Included in the patient packet were a study questionnaire and the web address to the study. Patients were invited to complete the hardcopy questionnaire or to complete the survey online in the study's website. Patients who did not participate after approximately 21 business days were pre-contacted with a priority second mailing. Those that still refused after 21 additional business days were randomized and systematically assigned to a methodological experiment being – either a third mailing or a telephone follow-up call conducted by the sites. This mode assignment was a methodological experiment to ensure that virtually half of the final follow-up attempts would be conducted by mail and the remaining half by telephone. The inclusion of the experiment was to determine whether participation gains could be achieved by following up the nonrespondents via telephone. Consequently, the third mailing proved to be more productive than the telephone follow-up by yielding 218 interviews compared to 144 from the telephone follow-up.

6. Processing Cases Monthly

Sample cases or data were loaded into the RTI Control System (CS), an internally developed system used for sample management and survey operation management. As sample data were loaded, several new variables were created to ensure successful operations in mailing survey packages, following up with non-respondents, tracking survey status, and managing paper survey receipt to include the following variables:

(1) Internal case ID (Our internal ID included data critical in mailing processes such as Center ID and sample number)
(2) Unique password to access the online survey

(3) Sample release number
(4) Sample released date
(5) Initial mailing and non-respondent follow-up mailing status and dates of occurrence
(6) Survey status and survey mode
(7) Receipt status of hard-copy survey

These variables were updated when their associated events took place by the nightly data processing system. The nightly process would also generate near "real-time" reports to provide staff with an accurate view of all operations and response rates by center as well as overall Centers.

7. Conducting Monthly Nonresponse Follow-Up

Nonrespondent follow-ups were crucial operations to ensure that the study achieved the maximum response rate possible. The follow-up mailings to nonrespondents took place 3-4 weeks from the previous mailing. Patient nonresponse was identified from the following variables:

(1) Status of the previous mailing
(2) Decease or refusal status of the patient
(3) Survey status
(4) Receipt status
(5) Center ID associated with each non-responding patient

Mail-merge files were created for each center that only comprised patients who did not respond for the first two mailings. These files were used for preparing survey packages at RTI. However, since RTI did not have respondent's contact info, RTI e-mailed the Centers the corresponding list of USIDs so they could prepare mailing labels for nonrespondents from their databases. This was a unique process that required a combination of automation and close collaboration between both RTI and staff at each center.

8. Patient Eligibility and Response Rates by Cancer Center

Eligibility and response rate computations were built within the daily data tracking and monitoring protocol. At any point in time, these quality metrics could be computed for each Cancer Center, monthly as well as overall months. AAPOR (American Association of Public Opinion Research) Rule #4 was used because it assumes ineligibility among the no contacts at a similar rate as those already surveyed. This rule also counts break-offs or partial responses as valid responses (AAPOR, 2011). **Figures 2** and **3** shows the eligibility and response rates for all Cancer Centers and how these rates varied for the Centers. All patients were considered eligible to participate unless decease or unable to complete the survey due to the severity of their illness. The overall eligibility rate was 97.0% among those with known eligibility. The Center with the highest eligibility rate was Center 10 at 99.2%. Center 21 had the lowest eligibility rate (50.0%) due to an error in the de-identified patient assignment that resulted in erroneously producing duplicate patient identifiers. A total of 2,517 patients completed the survey for an overall unweighted response rate was 59.5%. Center 23 had the highest response rate at 71.9%, and Center 15 had the lowest response rate at 39.5%.

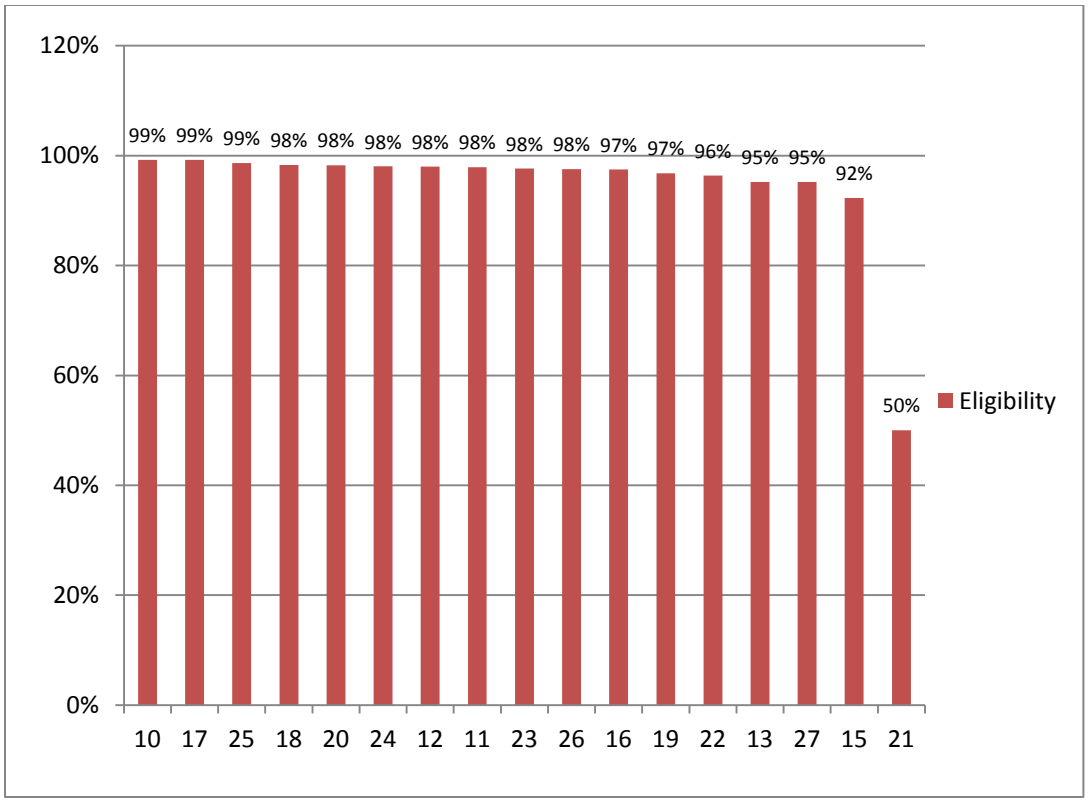

Figure 2: Cancer Center Eligibility Rates

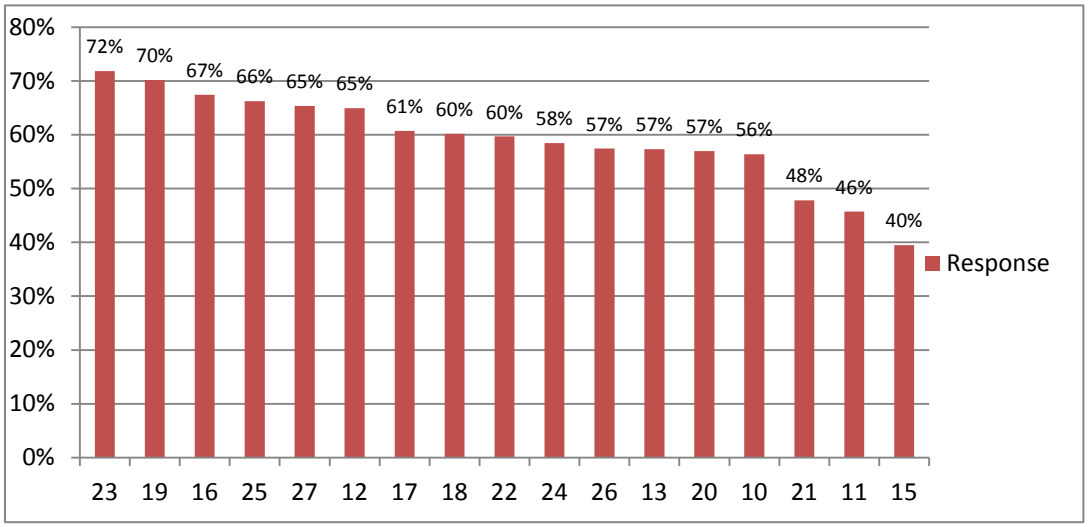

Figure 3: Cancer Center Response Rates

9. Modal Differences in Response Rates by Cancer Center

Consistent with most patient surveys, the majority (90.6%) of the patients chose to complete their surveys on the hardcopy version that was enclosed in the patient packets rather than to log on the study's secure website. The 9.4% who chose to complete the

survey on the website was not uniformly distributed over the 17 Cancer Centers. As shown in **Figure 4**, only 2.9% of participating patients from Cancer Center 15 chose the web version, which had the lowest web percentage. Only one patient out of 35 respondents chose the web version. The largest percent (14.0%) of web participants were from Cancer Center 26. Although it is not clear why such a small percentage of participants from Center 15 chose the web, it may be that a larger percentage of patients from this geographical area do not have web access and therefore unable to do a web version in comparison to many of the other Centers. The variability in modal choice for each Center is very apparent in the stacked bar chart given in **Figure 4**.

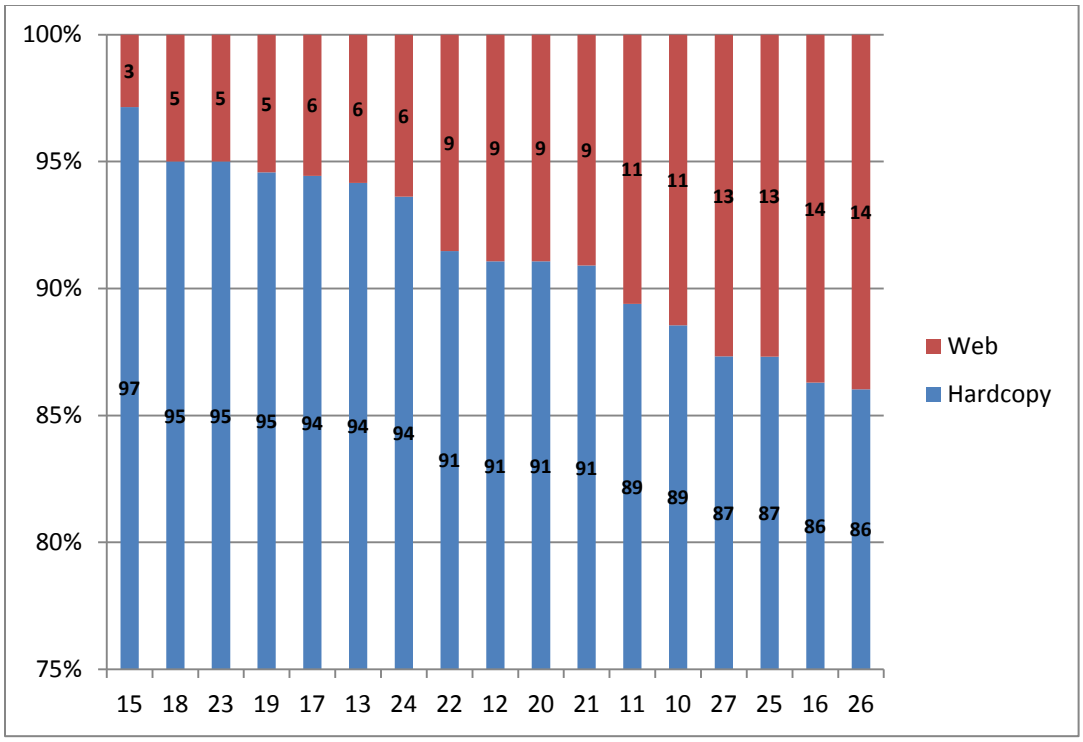

Figure 4: Percentage of Interviews Completed By Data Collection Mode

10. Discussion

This study had a series of entities involved, yet the study was implemented successfully while following HIPAA's Privacy Rule using de-identified data based on the NOPID (Jones et al., 2011) sampling, case processing, and analysis approach. We accomplished this through open collaboration between the 20 entities involved; the American Cancer Society, the American College of Surgeons, 17 Cancer Centers, and RTI International (the contractor). The primary requirements were that each entity clearly understood its purpose, adhered to collaboration, and followed pre-determined study guidelines. The NOPID approach is recommended for similar studies where respondent confidentiality is an important issue.

11. References

American Association of Public Opinion Research. (2011). Standard definitions: Final disposition of case codes and outcome rates for surveys. Deerfield, IL: Author.

Jones, S. M., McCormack, L. A., Johnson, T., Hobbs, C. L., McMichael, J. P., & Clauser, S. (2010, August). Sample design methodology for studying patients using registry data. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 6040–6051. Retrieved from https://www.amstat.org/membersonly/proceedings/2011/papers/400197.pdf

U.S. Department of Health and Human Services. (1992). The privacy rule. Available from http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html