## Title

Using Marketing Automation to Modernize Data Collection in the California Teachers Study Cohort

**Permalink**
https://escholarship.org/uc/item/35q3n1fb

**Journal**
Cancer Epidemiology Biomarkers & Prevention, 29(4)

**ISSN**
1055-9965

**Authors**
Savage, Kristen E
Benbow, Jennifer L
Duffy, Christine
et al.

**Publication Date**
2020-04-01

**DOI**
10.1158/1055-9965.epi-19-0841

Peer reviewed

Published in final edited form as:

*Cancer Epidemiol Biomarkers Prev.* 2020 April ; 29(4): 714–723. doi:10.1158/1055-9965.EPI-19-0841.

# Using marketing automation to modernize data collection in the California Teachers Study cohort

**Kristen E. Savage**[1,*], **Jennifer L. Benbow**[1], **Christine N. Duffy**[2], **Emma S. Spielfogel**[1], **Nadia T. Chung**[1], **Sophia S. Wang**[1], **Maria Elena Martinez**[3], **James V. Lacey Jr.**[1]

[1]Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA

[3]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA

## Abstract

**Background:** Like other cancer epidemiological cohorts, the California Teachers Study (CTS) has experienced declining participation to follow-up questionnaires; neither the reasons for these declines nor the steps that could be taken to mitigate these trends are fully understood.

**Methods:** The CTS offered their 6th study questionnaire (Q6) in the fall of 2017 using an integrated, online system. The team delivered a web and mobile-adaptive questionnaire to 45,239 participants via email using marketing automation technology. The study's integrated platform captured data on recruitment activities that may influence overall response, including the date and time invitations and reminders were emailed and the date and time questionnaires were started and submitted.

**Results:** The overall response rate was 43%. Participants ages 65 – 69 were 25% more likely to participate than their younger counterparts (OR=1.25, 95% CI, 1.18-1.32) and non-white participants were 28% less likely to participate than non-Hispanic white cohort members (OR=0.72, 95% CI, 0.68-0.76). Previous questionnaire participation was strongly associated with response (OR=6.07, 95% CI, 5.50-6.70). Invitations sent after 2 PM had the highest response (OR=1.75, 95% CI, 1.65-1.84), as did invitations sent on Saturdays (OR=1.48, 95% CI, 1.36-1.60).

**Conclusions:** An integrated system that captures paradata about questionnaire recruitment and response can enable studies to quantify the engagement patterns and communication desires of cohort members.

**Impact:** As cohorts continue to collect scientific data, it is imperative to collect and analyze information on how participants engage with the study.

[*]**Correspondence:** Kristen E. Savage; Department of Computational and Quantitative Medicine; City of Hope, 1500 E Duarte Road, Duarte, CA 91010-3000. Telephone: 626-218-3731; Fax: 626-471-7308; ksavage@coh.org.

**Conflict of Interest:** The authors declare no potential conflicts of interest.

**Keywords**

Marketing automation; web questionnaire; response rates; customer relationship management (CRM); survey development

## Introduction

Cancer epidemiological cohorts (CECs) are a cornerstone of epidemiological research and uniquely positioned to drive future discovery. These cohorts have traditionally collected self-reported data on health status, lifestyle factors, and attitudes from study participants via paper-based, mailed questionnaires, the response rate to which has steadily declined over time (1). Systematic studies of how best to administer these questionnaires to improve response and overall study retention within cohorts are limited (2, 3). How to maintain response to mailed paper surveys is also a challenge in other fields and industries (4, 5).

For this and other reasons, many population-based studies are adopting web-based questionnaires (6, 7). Web-based surveys for epidemiological research enable cohorts to streamline data collection, receive questionnaire responses more quickly, swiftly identify and respond to problems within the questionnaire, and automate recruitment (8). However, the same issues surrounding paper questionnaire logistics apply to web-based administration: what is the best way to distribute questionnaires, and how can researchers use them to ensure high-quality, representative data from study participants?

The move to web-based questionnaires makes it easier for cohort studies to begin to address these questions. Intrinsic in web-based questionnaire administration is the ability to collect paradata: data generated during—and about—the data collection process (9). Web-based surveys can capture key information about participant preferences and the effect of these preferences on response. Recorded at the respondent level, this information—who responds to the questionnaire, how they respond, when they respond, the device(s) on which they respond, etc.—lay the foundation to evaluate survey distribution methods and their effect on the collection of self-reported data.

The California Teachers Study (CTS), an on-going CEC established in the mid-1990s (10), has distributed five paper-based questionnaires and experienced the same declining response seen across the field. In transitioning the study's most recent survey to web, the team designed the questionnaire to collect paradata at each stage of questionnaire recruitment to identify which elements were associated with response.

## Materials and Methods

### Study Population

The CTS began in 1995-1996, when 133,479 female public-school professionals (primarily teachers or administrators, ages 22-104) completed a mailed self-administered questionnaire. All participants have been invited to complete four paper-based follow-up questionnaires, in 1997-1999 (Q2); 2000-2002 (Q3); 2005-2007 (Q4); and 2012-2015 (Q5). Grant award

U01-CA199277, from 2015 to 2020, funded the CTS to administer a 6th follow-up survey (Q6) that gave participants the option to complete the survey electronically or on paper.

All CTS participants who were alive and had not withdrawn or opted out of future contact would be invited to complete Q6 (see Figure 1). Participants who had provided an email were initially recruited for the web-based version of Q6, i.e., Q6web, as described in this paper. Participants who had not provided an email address would receive the paper version of Q6, i.e., Q6paper.

Of the 133,479 original CTS participants, 2 had withdrawn from the CTS; 29,970 were deceased; and 2,060 had opted out of further participation at the time of the questionnaire. Four additional participants were excluded due to discrepancies with their baseline questionnaires. Another 12,144 participants completed the CTS baseline survey in 1995-1996 but had not completed any of the four CTS follow-up questionnaires; we considered this subgroup to be inactive. The remaining 89,299 participants were eligible for Q6 in fall of 2017. Of these participants, 45,239 (50.7%) had provided the CTS with an email address and were considered eligible for the web-based questionnaire. Seven eligible participants died shortly after Q6web recruitment began and are therefore excluded from this analysis.

## Questionnaire Content

The CTS Steering Committee (https://www.calteachersstudy.org/team) determined content with consideration of current and new scientific hypotheses, the need to update key participant lifestyle and health factors, and emerging topics of interest. Question wording was based on previous questionnaires and validated questions from other sources. Q6 topics included physical activity, health conditions, medications, family health history, reproductive history, body size, financial stress, social support, sexual orientation and gender identity, and medicinal cannabis.

## Questionnaire Platform

Since 2014, the CTS has consolidated study management activities within a single study customer relationship management (CRM) platform hosted on Salesforce.com (https://www.salesforce.com/). Initially developed to manage customer acquisition and sales cycles, CRM platforms integrate all customer “touchpoints”—and paradata about those touchpoints —within a single central system with the ultimate goal of improving customer relationships (11). The CTS CRM is used to manage participant contact information and recruitment activities for projects that collect data or biospecimens.

We evaluated multiple survey platforms for their ability to integrate directly with the study CRM, their ease of use, compatibility with mobile devices, and security. The CTS purchased an annual subscription to Qualtrics.com (https://www.qualtrics.com/), an experience management company that provides customer, brand, product, and employee experience platforms. For CTS research, we used the Qualtrics survey tools, which removed the need to custom-code a CTS-specific platform. All Qualtrics surveys are web and mobile enabled, use Transport Layer Security (TLS) encryption, and meet the technical requirements of the Health Information Technology for Economic and Clinical Health Act (HITECH).

## Questionnaire Design

Qualtrics' point-and-click questionnaire development software offers over 150 question templates. After the CTS agreed on topic and question order, one CTS team member added the content to Qualtrics, specified question format, determined topic and question order, governed display logic, created test questionnaires, and directly integrated Qualtrics and the study CRM without support from software developers or institutional IT staff.

**Display logic and embedded data.—**Display logic and embedded data are standard components in self-reported questionnaires used by cancer epidemiology cohorts. Display logic, also known as skip logic, determines which questions are presented to participants based on their previous responses within the questionnaire. Display logic helps achieve efficiency and accuracy by hiding questions from participants for whom they would be redundant and asking additional questions of participants for whom follow-up is applicable. Embedded data is additional information stored directly within a questionnaire. For example, previous CTS questionnaires included questions such as "Since 2005, have you…"; the dates embedded in these questions frame the period of interest. In the previous CTS paper surveys, the skip logic and embedded data were incorporated into the design of the pages and therefore held constant across participants' questionnaires.

For Q6web, we used display logic and embedded data to individualize participants' questionnaires. We identified the self-reported data from previous questionnaires that was relevant to Q6 and embedded it in the questionnaire. Each participant's questionnaire contained her name, birthdate, mailing address, email address(es), and phone number(s) for verification; participants were asked to verify or correct their contact information. Participants' menopausal status and the month and year of their most recent questionnaire were used as display logic: women who had previously reported they were postmenopausal were not presented with the menopausal status section, and participants were asked if they had used hormone therapy since the date of their last completed questionnaire.

Applying this individual-level data for all eligible participants generated a personalized, unique questionnaire link for each Q6web recipient.

## Questionnaire Dissemination

The CTS agreed upon the questionnaire recruitment methodology as outlined in Figure 2a, whereby a participant would receive follow-up emails depending on her interaction with the first questionnaire invitation (Invite 1). Study participants could receive up to three invitations and three reminder emails for a maximum of six emails within a 30-day period.

The CTS utilized Pardot.com (https://www.pardot.com/), a marketing automation software supplied by Salesforce.com, to implement this protocol. Marketing automation software enables users to automate activities, such as sending emails, based on pre-defined rules (12, 13). These rules use "if, then" statements to determine action based on time intervals and/or participant behavior.

CTS staff designed email templates in Pardot for each Q6 invite and reminder as shown in Figure 2a. Templates were created once; Pardot automatically personalized the first name and questionnaire URL embedded in the email when it was sent.

Staff added these email templates to a Q6 nurturing campaign. Nurturing campaigns employ behavior-based email automation: the “if, then” rules governing automated tasks depend on the email recipient’s actions. For example, whenever a participant met the CTS’ rule for receiving Invite 2, Pardot automatically sent her the Invite 2 email template with her first name and personal questionnaire link.

## Participant Experience

Participants started their questionnaires by clicking on the unique URL in their email invitation. Each questionnaire link recorded that participant’s progress as she advanced through the questionnaire. This permitted participants to start and stop the questionnaire at will, switch browsers, or change devices while preserving their previous answers. The bottom of each screen contained the CTS’s toll-free number and invited participants to call if they encountered any difficulties or had questions.

## Data Collection & Integration

Within the CTS’s CRM platform, each participant has a “participant record” that includes all the data associated with her relevant CTS activities. Study staff directly integrated Qualtrics and Pardot with the study CRM so activities occurring within these external platforms were automatically documented on each participant’s record (see Figure 3). Questionnaire responses and the paradata associated with those responses—including start time, completion time, start date, completion date, the device used at questionnaire start, operating system used, etc.—were recorded on the participant’s record, as were paradata on all the recruitment activities and their outcomes. These data included the date and time each recruitment email was sent, the total number of invitations and reminders emailed, and email hard bounces, i.e. emails that could not be delivered to the recipient’s inbox.

## Statistical Analysis

This analysis includes data on recruitment and responses collected between October 2017 and June 2019. The primary outcome was completion of Q6web. Secondary outcomes were completion of Q6web after the 1st, 2nd, or 3rd invitations. The primary potential explanatory variables were participant age, race, cancer history, previous questionnaire participation, as well as the paradata on number of invitations and reminders sent; day of week that the invitations were sent; and time of day that the invitations were sent.

We used Chi-square tests to evaluate differences in the distributions of response by categories of exposures. We did not specify the time of day that the email invitations were sent, but we categorized into approximate quartiles (see Table 1). We used logistic regression to calculate odds ratios (ORs) and 95% confidence intervals (95% CIs) associated with participant and recruitment characteristics. All analyses were completed with SAS 9.4 (Cary, NC) using the CTS Data Warehouse (https://www.calteachersstudy.org/for-researchers).

# Results

## Recruitment Population

CTS participants who were recruited for Q6web (Q6web recruits) differed from the group that was not recruited (Table 1a). Q6web recruits were statistically significantly younger than participants who were not recruited. The percentage of participants over age 80 who were recruited for Q6web (18%) was approximately half of the percentage of participants over age 80 who were not recruited for Q6web (33%). Compared with the CTS participants who were not recruited for Q6web, Q6web recruits were less likely to have a personal history of cancer. Within the recruited population, cancer survivors were more likely than participants without a personal history of cancer to be over age 80 (34% vs. 17%).

Q6web recruits were also more likely to be white and approximately twice as likely to have completed the previous CTS questionnaire (Q5, in 2012-2014) or to have completed all the CTS follow-up surveys since the CTS began in 1995-1996.

## Recruitment Invitations

A total of 19,481 of the 45,232 participants, or 43.1%, completed Q6web. Over half (50.8%) of the respondents completed their questionnaire after the first invitation (Invite 1) and without needing any reminder emails (Figure 2b). Another 6% of total responders completed their questionnaire after receiving the first reminder email (Reminder 1); the second and third reminder emails produced fewer responses.

Follow-up email invitations were only half as effective as the previous invitation emails: response after the second (Invite 2) accounted for 23% of the total response, and response after the third (Invite 3) accounted for only 13% of the total response. This pattern of declining return also occurred with the reminder emails: the responses to the second and third reminders were consistently 60% to 75% lower than the responses to Reminder 1 (Figure 2b). This pattern was consistent for the first, second, and third invitation emails. As the cumulative number of emails sent increased after Invite 2, the total cumulative number of completed questionnaires remained relatively flat (Figure 2c).

Table 1b presents the days of the week and the time of day that Invites 1, 2, and 3 were sent. All 45,232 eligible participants were sent Invite 1, but Invite 2 and Invite 3 were only sent to participants who had not yet started their questionnaire. Therefore, the three denominators—45,232 for Invite 1; 31,784 for Invite 2; and 26,079 for Invite 3—are nested subsets that include the same non-responders. Invitation emails were not evenly distributed across time of day and day of week (Table 1b). A higher percentage of invitation emails were sent mid-week than on weekends and in the early morning than later in the day, and the proportion of emails sent before 9AM increased with subsequent invitations.

The ORs in Table 2a model the odds of having completed Q6web: ORs above 1.00 indicate a positive association between that characteristic and completing the questionnaire. Compared with participants under age 65, participants ages 65 to 79 were significantly more likely to complete Q6web (ORs ranged from 1.11 to 1.25, see Table 2a). In contrast, participants over age 80 years were 50% less likely to complete Q6web compared with those under age 65 (OR=0.50, 95% CI, 0.47-0.54). Non-white participants were less likely to complete Q6web compared with non-Hispanic white participants (OR=0.72, 95% CI, 0.68-0.76). Cancer survivorship was not significantly associated with completion of Q6web. Among cancer survivors, type of cancer diagnosis was not associated with Q6web response (Chi-square p-value=0.08).

Participants who completed the previous CTS survey (Q5, 2012-2015) were six times more likely to complete Q6web than participants who did not complete Q5 (OR=6.07, 95% CI, 5.50-6.70). Having completed all the previous CTS follow-up surveys was significantly associated with completing Q6web (OR=3.06, 95% CI, 2.93-3.20). The multivariate ORs were not markedly different from the univariate ORs, which were adjusted for categorical age only.

Table 2b presents the associations between the time of day and the day of week that Invite 1 was emailed and the odds of those participants completing Q6web. The univariate ORs are adjusted for categorical age only; the multivariate ORs are adjusted for race, cancer survivorship, and having completed Q5. Compared with participants to whom Invite 1 was emailed before 9AM, participants to whom Invite 1 was emailed later in the day were statistically significantly more likely to have completed Q6web. Multivariate adjustment attenuated these associations, but the pattern of associations remained consistent. Participants to whom Invite 1 was sent after 2PM were 50% more likely to have completed Q6web than participants to whom Invite 1 was emailed before 9AM (OR=1.51, 95% CI, 1.43-1.60).

Response to Q6web was also associated with the day of week that Invite 1 was sent. Compared with participants to whom Invite 1 was emailed on a Tuesday, participants whose first invitation was emailed on any other day except Friday (OR=1.04, 95% CI, 0.97-1.11) were significantly more likely to complete Q6web (ORs ranged from 1.16 to 1.26, see Table 2b). Multivariate adjustment for participant characteristics reduced the magnitude of the associations with Saturday and Sunday (OR=1.34, 95% CI, 1.24-1.46 and OR=1.22, 95% CI, 1.13-1.31, respectively) but slightly increased the magnitude of the associations with Wednesday and Thursday (OR=1.26, 95% CI, 1.17-1.35 and OR=1.16, 95% CI, 1.08-1.24, respectively).

Table 2c shows the associations with Q6web response in models that a) included Invite 1 time of day and day of week and b) were adjusted for the participant characteristics of age, race, cancer survivorship, and Q5 completion. The reference groups for these ORs are the participants who were sent Invite 1 before 9AM and on a Tuesday. Among all participants who were sent Invite 1, those who were sent that invitation email after 2PM (OR=1.80, 95% CI, 1.69-1.92) or on a Saturday (OR=1.79, 95% CI, 1.64-1.96) were almost twice as likely to complete Q6web. Invite 1 emails sent any time after 9AM, or on any day other than a Tuesday, were significantly associated with completion of Q6web. Additional adjustment for retirement age (<65 vs. >=65) or self-reported retirement at Q5 did not materially change these associations.

Because the majority of the Q6web responders had completed Q5, these associations were nearly identical in models restricted to participants who had completed Q5. However, among the subgroup of Q6web respondents who had not completed Q5, day of week was not as strongly or clearly associated with response. Compared with participants who were sent Invite 1 on a Tuesday, only the participants who were sent Invite 1 on a Wednesday were more likely to complete Q6web (OR=2.22, 95% CI, 1.55-3.18) (see Figure 4a). Time of day remained associated with Q6web response, but the associations were less precise than among the entire population that was invited to complete Q6web (see Figure 4b). Among this subgroup, invitations sent between 10 and 11 AM were almost two and a half times more likely to be completed (OR=2.44, 95% CI, 1.78-3.32), but the association was less exact than for all Q6web recruits (OR=1.20, 95% CI, 1.14-1.27). Invitation emails sent to Q5 non-responders bounced twice as often (12% bounce rate) as invitation emails to participants who had completed Q5 (6% bounce rate).

A total of 640 Q6web recruits (1.4%) called the 1-800 number. Participants age 75 and older called more frequently than their younger counterparts; these age groups comprised 34% of the recruited population but constituted 52% of the caller population (Figure 5).

## Discussion

To distribute our 6th questionnaire, the CTS used an integrated combination of a commercial survey platform, marketing automation software, and the existing CRM. This approach provided a professional-grade and user-friendly questionnaire experience to study participants. It also generated new data about how and when participants responded. The days of the week and times of day on which the invitations were emailed were significantly associated with whether participants completed the questionnaire. Even among the subgroup of CTS participants who have responded most consistently to all follow-up surveys, timing mattered. Participants to whom we sent invitation emails on Saturdays or in the afternoons were more likely to complete the survey than participants to whom we sent invitation emails in the middle of the week or in the mornings, respectively. These associations raise interesting questions about how CEC questionnaire recruitment protocols ultimately influence participant response.

This integrated approach had multiple strengths. The commercially available survey platform was user-friendly, customizable, and could be easily configured for this survey. The software was easily managed by study staff; the entire implementation—from content brainstorm to questionnaire delivery—was completed in 10 months. The range of question and answer templates accommodated the CTS' scientific needs, and the survey platform allowed us to extensively personalize the questionnaire at scale, for over 43,000 participants. This streamlined the experience for participants, increased the efficiency of our data collection process, and improved data consistency. The Qualtrics platform intrinsically collected new paradata and included multiple options for pilot-testing and reviewing results immediately after participants submitted their survey. Internal data suggest our survey costs were approximately 50% lower for Q6web than for previous CTS surveys.

This approach also has some limitations. Our team had to develop proficiency in the marketing automation software used to disseminate the survey. Integrating the survey platform, marketing automation software, and our CRM required that someone on staff manage each platform. This strategy built on our existing experience using CRM for a 2013-2016 CTS biobanking project. Studies that have not previously used CRM or integrated their data in this way would have a steeper learning curve, but all three of the platforms we used are user-friendly and designed to be configured by users for their specific experience.

Another limitation is the absence of paradata from earlier CTS questionnaires. Previous surveys were mailed, scannable questionnaires. Data on response to those surveys based on day or date of mailing were not collected. In the absence of those data, it is impossible to determine whether the associations we observed here existed during previous surveys; if they did, it is impossible to know whether these associations have changed over time.

We did not directly ask participants why they responded when they did. Future surveys should try to determine how much those patterns reflect participants' preferences vs. potentially random behavior. It would also be informative to know why participants chose not to respond, although that information is challenging to ascertain.

This is a hypothesis-generating analysis of how we collected data, rather than an analysis of existing scientific data. When we designed the email invitation schedule, we did not actively choose which days of the week or times of day to send the survey. Although we observed associations between these "exposures"—the paradata on how we delivered Q6—and the outcome of response vs. non-response, these associations could be confounded by other data that we did not measure, such as what participants were experiencing when they received the invitations. We assume those choices were non-random, but more data would be needed to understand those associations.

This analysis of CTS paradata demonstrates that every recruitment activity represents informative data that should be collected, analyzed, and utilized to improve CEC research. There is a need for this type of infrastructure science and paradata, especially for real-world data (14). These CTS paradata are a positive side effect of our decision to use commercially available survey and marketing automation software; paradata are essential components of marketing and communications in other industries and sectors. There are enough similarities between those industries and how CECs need to collect patient-reported data to consider making detailed paradata a standard component of a CEC's data collection strategy, regardless of whether surveys are electronic or on paper. The associations between Q6web response and time of day, day of week, and number of recruitment emails sent are based on one CTS survey. Whether these patterns are generalizable to future CTS surveys or other CECs will be unknown until additional data from other surveys is available for comparison. Future CTS data collection protocols should be designed based on results of extensive pilot testing, using this type of paradata. For example, if pilot testing confirmed that emails sent later in the day and later during the week generated higher response, then the subsequent full-scale recruitment emails could be scheduled to only be sent during high-yield time windows. Other large-scale CECs, which require significant investments of time, personnel, and funding to collect data, might also benefit from a similar approach to understand their participants' behavior and align their recruitment strategies to those behavior patterns.

Overall, these data indicate that response to CEC follow-up surveys is associated with more than just the demographic characteristics of the participants. As CECs consider how to efficiently collect additional self-reported data from participants and patients (14), further exploration of these types of metadata and paradata has the potential to improve data collection protocols and the resulting research.

## Acknowledgements

The California Teachers Study and the research reported in this publication were supported by the National Cancer Institute of the National Institutes of Health under award number U01-CA199277; P30-CA033572; P30-CA023100; UM1-CA164917; and R01-CA077398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

The collection of cancer incidence data used in the California Teachers Study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The opinions, findings, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the official views of the State of California, Department of Public Health, the National Cancer Institute, the National Institutes of Health, the Centers for Disease Control and Prevention or their Contractors and Subcontractors, or the Regents of the University of California, or any of its programs.

The authors would like to thank the California Teachers Study Steering Committee that is responsible for the formation and maintenance of the Study within which this research was conducted. A full list of California Teachers Study team members is available at https://www.calteachersstudy.org/team.

Great thanks are also due to the California Teachers Study participants, who have given their time and commitment to the study since 1995. The data and information they provided have been an invaluable contribution to women's health and cancer research.

**Financial Support:** U01-CA199277; P30-CA033572; P30-CA023100; UM1-CA164917; and R01-CA077398

## References

1. Galea S, Tracy M. Participation rates in epidemiologic studies. Annals of epidemiology. 2007 9;17(9):643–53. PubMed PMID: 17553702. Epub 2007/06/08. eng. [PubMed: 17553702]
2. Teague S, Youssef GJ, Macdonald JA, Sciberras E, Shatte A, Fuller-Tyszkiewicz M, et al. Retention strategies in longitudinal cohort studies: a systematic review and meta-analysis. BMC medical research methodology. 2018 11 26;18(1):151. PubMed PMID: 30477443. Pubmed Central PMCID: 6258319. [PubMed: 30477443]
3. Booker CL, Harding S, Benzeval M. A systematic review of the effect of retention methods in population-based cohort studies. BMC public health. 2011 4 19;11:249. PubMed PMID: 21504610. Pubmed Central PMCID: 3103452. [PubMed: 21504610]
4. Lacey JV Jr., Savage KE. 50 % Response rates: half-empty, or half-full? Cancer causes & control : CCC. 2016 6;27(6):805–8. PubMed PMID: 27100357. [PubMed: 27100357]
5. Czajka JL, Beyler A. Background Paper Declining Response Rates in Federal Surveys: Trends and Implications. Washington, DC: Mathematica Policy Research. 2016.
6. Russell CW, Boggs DA, Palmer JR, Rosenberg L. Use of a web-based questionnaire in the Black Women's Health Study. American journal of epidemiology. 2010 12 1;172(11):1286–91. PubMed PMID: 20937635. Pubmed Central PMCID: PMC3025633. Epub 2010/10/13. eng. [PubMed: 20937635]

7. Smith B, Smith TC, Gray GC, Ryan MA. When epidemiology meets the Internet: Web-based surveys in the Millennium Cohort Study. American journal of epidemiology. 2007 12 1;166(11):1345–54. PubMed PMID: 17728269. Epub 2007/08/31. eng. [PubMed: 17728269]

8. van Gelder MM, Bretveld RW, Roeleveld N. Web-based questionnaires: the future in epidemiology? American journal of epidemiology. 2010 12 1;172(11):1292–8. PubMed PMID: 20880962. Epub 2010/10/01. eng. [PubMed: 20880962]

9. Callegaro M Paradata in web surveys. Improving surveys with paradata: Analytic uses of process information. 2013:261–79.

10. Bernstein L, Allen M, Anton-Culver H, Deapen D, Horn-Ross PL, Peel D, et al. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). Cancer causes & control : CCC. 2002 9;13(7):625–35. PubMed PMID: 12296510. Epub 2002/09/26. eng. [PubMed: 12296510]

11. Chen IJ, Popovich K. Understanding customer relationship management (CRM): People, process and technology. Business Process Management Journal. 2003;9(5):672–88.

12. Heimbach I, Kostyra DS, Hinz O. Marketing Automation. Business & Information Systems Engineering. 2015 4 01;57(2):129–33.

13. Järvinen J, Taiminen H. Harnessing marketing automation for B2B content marketing. Industrial Marketing Management. 2016;54:164–75.

14. Data Science Opportunities for the National Cancer Institute: Interim Report of the National Cancer Advisory Board Working Group on Data Science. National Institutes of Health, 2018 8/14/2018.

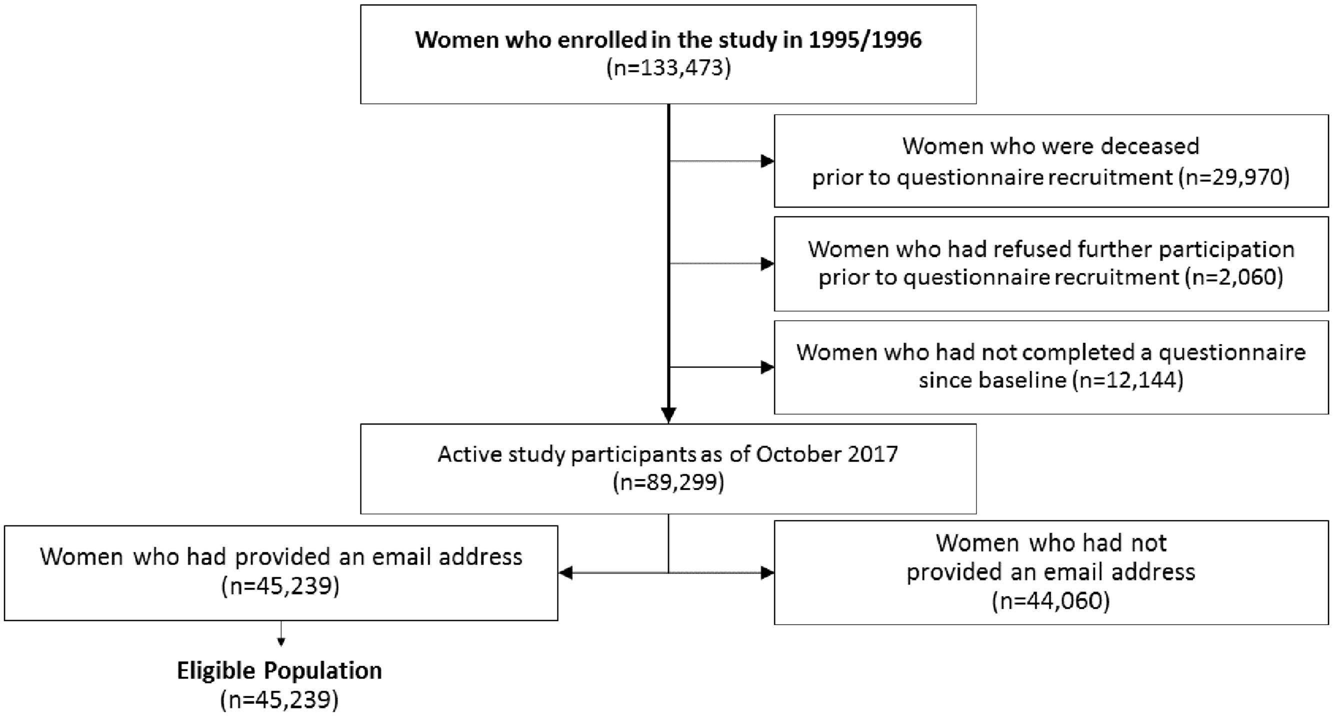

Women who enrolled in the study in 1995/1996
(n=133,473)
Women who were deceased
prior to questionnaire recruitment (n=29,970)
Women who had refused further participation
prior to questionnaire recruitment (n=2,060)
Women who had not completed a questionnaire
since baseline (n=12,144)
Active study participants as of October 2017
(n=89,299)
Women who had provided an email address
(n=45,239)
Women who had not
provided an email address
(n=44,060)
Eligible Population
(n=45,239)

**Figure 1.**

Eligibility for recruitment in Questionnaire 6 Web (Q6web). Of the 133,479 original study participants, two have asked to be removed from the study completely. An additional four had discrepancies in their baseline questionnaires. The exclusions applied to the remaining 133,473 California Teachers Study participants are displayed in the flowchart above. Seven participants died shortly after Q6web recruitment began. The response patterns of the remaining 45,232 eligible participants are analyzed in this paper.

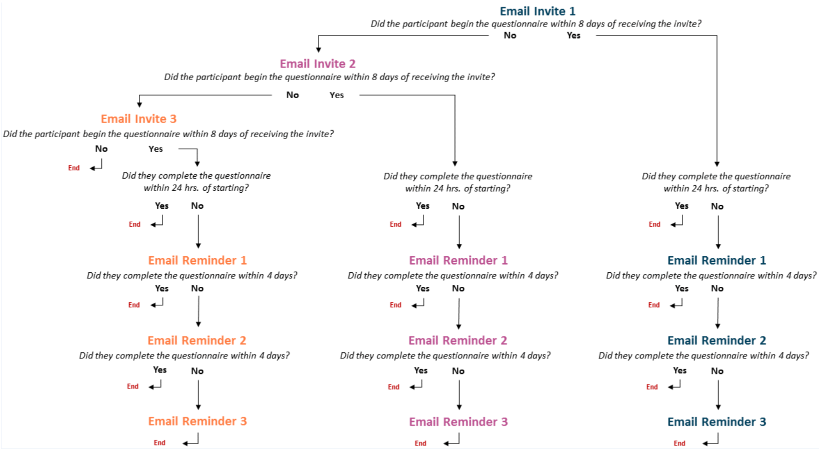

Email Invite 1
Did the participant begin the questionnaire within 8 days of receiving the invite?
No
Yes
Email Invite 2
Did the participant begin the questionnaire within 8 days of receiving the invite?
No
Yes
Email Invite 3
Did the participant begin the questionnaire within 8 days of receiving the invite?
No
Yes
End
Did they complete the questionnaire within 24 hrs. of starting?
Yes
No
End
Email Reminder 1
Did they complete the questionnaire within 4 days?
Email Reminder 2
Did they complete the questionnaire within 4 days?
Email Reminder 3

**Figure 2a.**

Q6web recruitment nurturing campaign. All eligible participants received Invite 1. A participant’s actions determined the follow-up emails she received as noted in the figure. Invites 1, 2, and 3 all contained different content and images. Reminder emails differed from one another but were identical across follow-up paths, i.e. a participant who received Reminder 1 after opening Invite 1 received the same email template as a participant who received Reminder 1 after opening Invite 3.

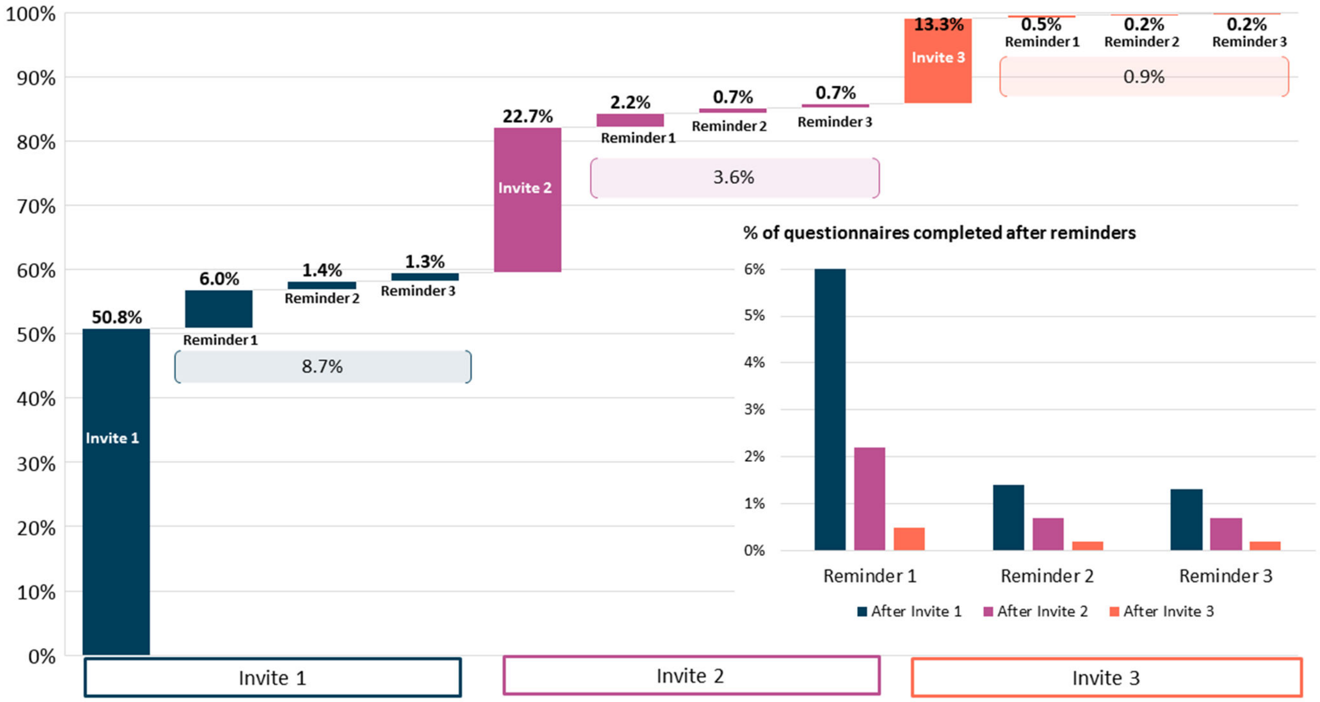


**Figure 2b.**

Percent of questionnaires completed after each recruitment touchpoint. The call-out box highlights the % of completed questionnaires following each reminder email. Reminders sent to participants who opened Invite 1 had more than twice the return as reminders sent to participants who opened Invite 2.

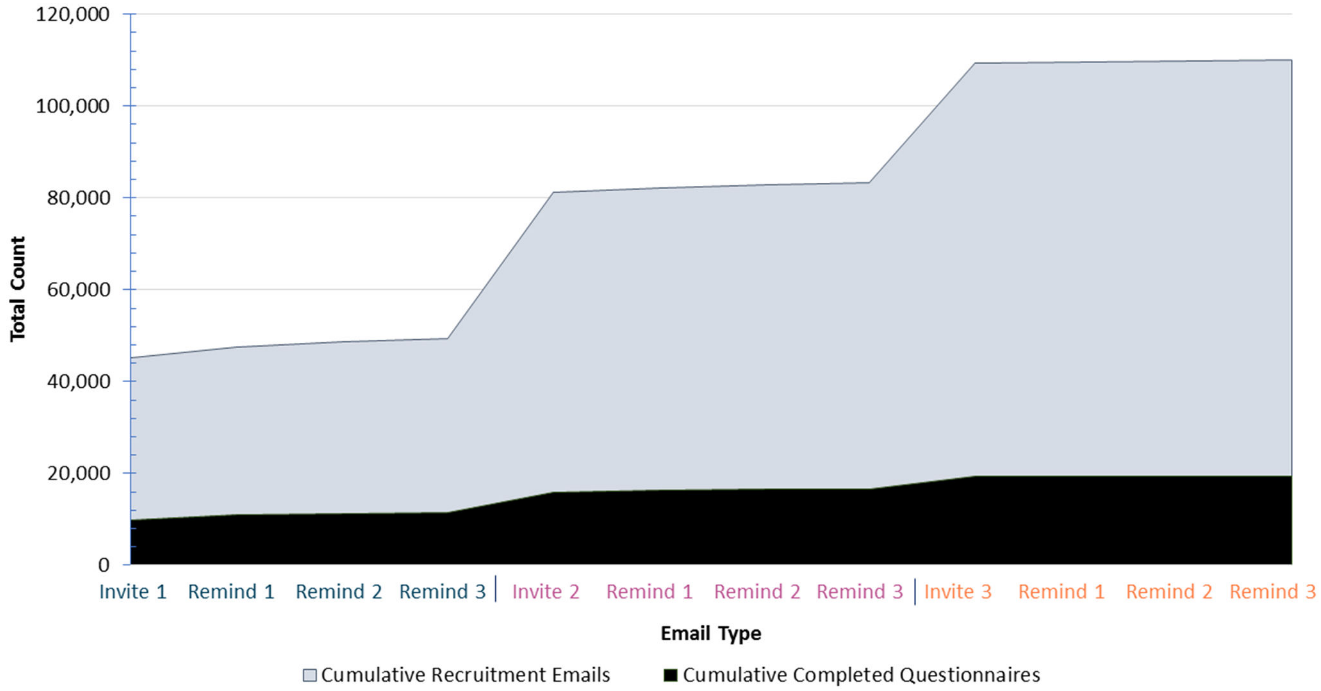


**Figure 2c.**

Cumulative recruitment emails vs. completed questionnaires. This figure illustrates the cumulative recruitment emails alongside the cumulative completed questionnaires at each email stage. Most completed questionnaires were received after Invite 1 was emailed. As additional recruitment emails were sent, the total number of completed questionnaires increased incrementally.

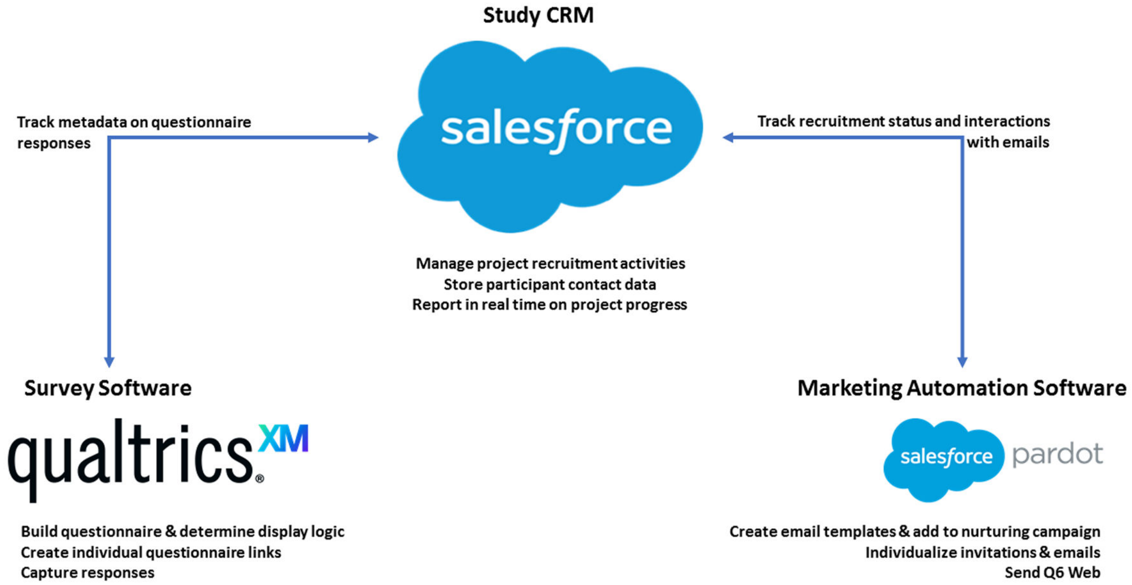

Study CRM
salesforce
Track metadata on questionnaire responses
Track recruitment status and interactions with emails
Manage project recruitment activities
Store participant contact data
Report in real time on project progress
Survey Software
qualtrics XM
Build questionnaire & determine display logic
Create individual questionnaire links
Capture responses
Marketing Automation Software
salesforce pardot
Create email templates & add to nurturing campaign
Individualize invitations & emails
Send Q6 Web

**Figure 3.**

Integrated customer relationship management (CRM) schema used to disseminate Q6web. The figure demonstrates the integrated approach leveraged to disseminate Q6web and collect questionnaire responses. Staff built the questionnaire within Qualtrics, which directly integrated with the study CRM hosted on Salesforce.com. Questionnaire responses were immediately mapped to the participant's record in Salesforce upon questionnaire completion. Pardot, a marketing automation software supplied by Salesforce.com, was used to create Invite and Reminder templates and disseminate these recruitment emails according to the rules created in the Q6web nurturing campaign. Recruitment metadata and questionnaire paradata were automatically recorded on each eligible participant's record in Salesforce.

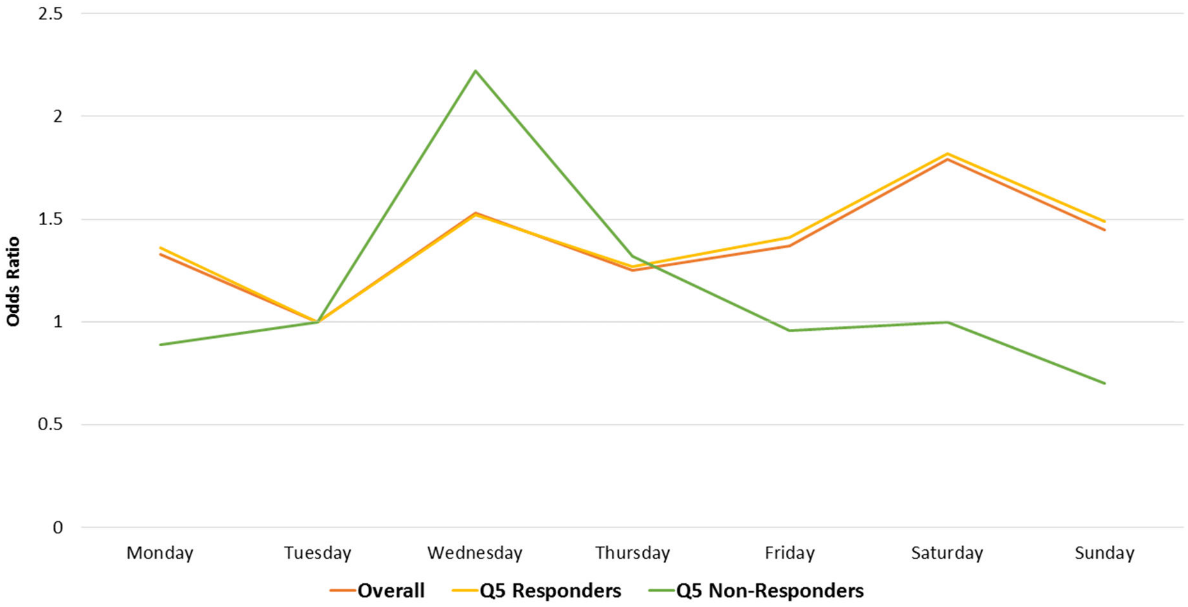


**Figure 4a.**

Likelihood of completing Q6web by Invite 1 day of week. This figure displays the odds ratios (ORs) of Q6web completion by the day of week Invite 1 was sent, stratified by participants who responded to Questionnaire 5 (Q5 responders) and participants who did not respond (Q5 non-responders). Q5 responders comprise the majority of Q6web recruits; therefore, there is little variability between the overall ORs and ORs for Q5 responders.

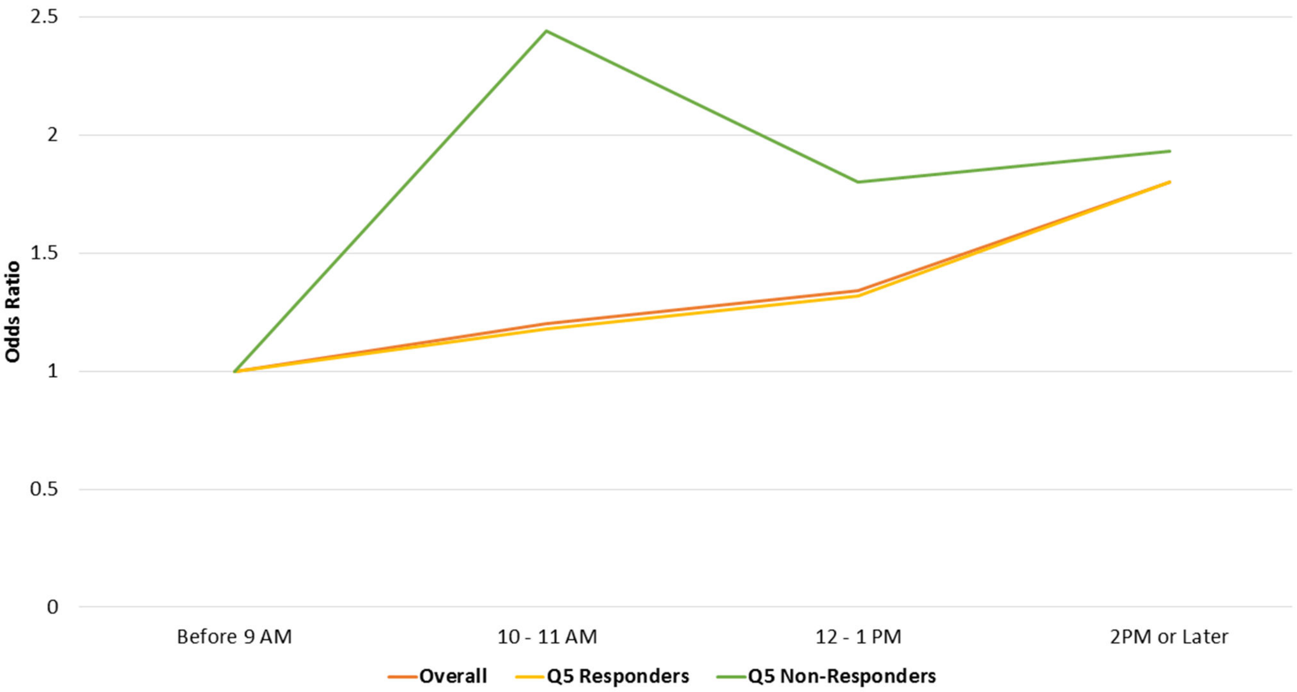
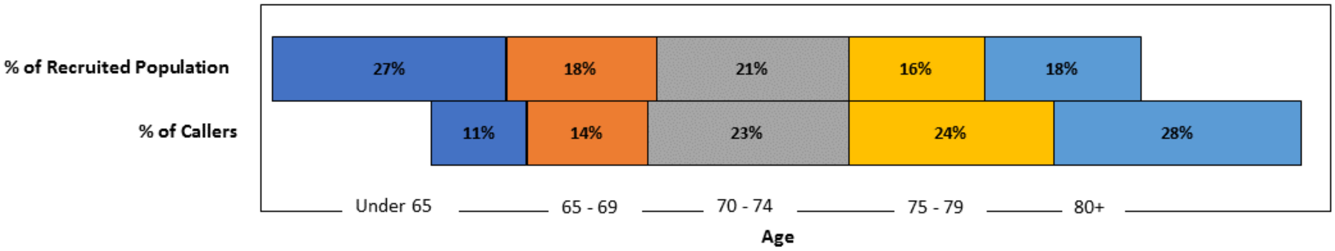


**Figure 4b.**

Likelihood of completing Q6web by Invite 1 time of day. This figure displays the odds ratios (ORs) of Q6web completion by the time of day Invite 1 was sent. Compared with invitations sent before 9AM, invitations sent later in the day were more likely to be completed among Q5 responders and Q5 non-responders.



**Callers age 75 and older comprised 34% of the recruited population but constituted 52% of the caller population.**

**Figure 5.**

Q6web recruited population vs. caller population by age group. This figure illustrates the difference in age distribution between the population recruited for Q6web and the population that called the CTS 1-800 number. Participants age 75 and older called the toll-free number more often than their younger counterparts.

## Table 1a:

Participant characteristics of Q6web recruits and non-recruits

| Participant Characteristics | Recruited for Q6W<br>N = 45,239* | Rest of Population*<br>N = 44,060 | P Value |
|---|---|---|---|
| **Age** [a] | | | |
| *Younger than 65* | 12,295 (27.8%) | 11,434 (26.0%) | <0.0001 |
| *65 – 69* | 8175 (18.1%) | 6050 (13.7%) | |
| *70 – 74* | 9510 (21.0%) | 6640 (15.1%) | |
| *75 – 79* | 7119 (15.7%) | 5385 (12.2%) | |
| *80+* | 8140 (18.0%) | 14,551 (33.0%) | |
| **Race** | | | |
| *Non-Hispanic White / Caucasian* | 39,975 (88.4%) | 37,039 (84.1%) | <0.0001 |
| *Non-White* | 5264 (11.6%) | 7021 (15.9%) | |
| **Personal History of Cancer at Q6** | | | |
| *None* | 42,345 (93.6%) | 35,069 (79.6%) | <0.0001 |
| *Cancer survivor at Q6* | 2894 (6.4%) | 8991 (20.41%) | |
| **Responses to previous CTS questionnaires** | | | |
| *Completed Questionnaire 5 (Q5)* | 41,446 (91.6%) | 21,407 (48.6%) | <0.0001 |
| *Did not complete Q5* | 3793 (8.4%) | 22,653 (51.4%) | |
| | | | |
| *Completed all prior Questionnaires* | 31,138 (68.8%) | 13,275 (30.1%) | <0.0001 |
| *Did not complete all prior Questionnaires* | 14,101 (31.2%) | 30,785 (69.9%) | |

[a] Seven participants died shortly after Q6web recruitment began. Those participants are included in Table 1a but excluded from the other reported results.

## Table 1b:

Distribution of recruitment email time of day and day of week

| **Invite 1** | ***Day of Week*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Sun.** | **Mon.** | **Tue.** | **Wed.** | **Thu.** | **Fri.** | **Sat.** | **Total sent** | **% completed** |
| *Time of Day* | | | | | | | | | |
| Before 9am | 598 | 499 | 1042 | 2606 | 1718 | 4199 | 1949 | 12,611 | **22.7%** |
| 10am-11am | 2650 | 2725 | 2275 | 2061 | 1647 | 348 | 1234 | 12,940 | **24.4%** |
| 12pm-1pm | 1933 | 3462 | 11 | 927 | 994 | 2693 | 0 | 10,020 | **26.9%** |
| After 1pm | 281 | 1617 | 4100 | 998 | 2402 | 0 | 263 | 9661 | **29.8%** |
| *Total sent* | 5462 | 8303 | 7428 | 6592 | 6761 | 7240 | 3446 | 45,232[a] | |
| **% completed** | **13.1%** | **18.0%** | **15.1%** | **15.2%** | **14.1%** | **15.4%** | **9.1%** | | **25.6%** |
| | | | | | | | | | |
| **Invite 2** | ***Day of Week*** | | | | | | | | |
| | **Sun.** | **Mon.** | **Tue.** | **Wed.** | **Thu.** | **Fri.** | **Sat.** | **Total sent** | **% completed** |
| *Time of Day* | | | | | | | | | |
| Before 9am | 1838 | 1931 | 1607 | 3541 | 2794 | 3307 | 2721 | 17,739 | **14.5%** |
| 10am-11am | 20 | 245 | 359 | 1973 | 254 | 31 | 265 | 3147 | **14.9%** |
| 12pm-1pm | 0 | 22 | 993 | 17 | 136 | 413 | 20 | 1601 | **20.1%** |
| After 1pm | 1678 | 1795 | 2107 | 1976 | 473 | 321 | 947 | 9297 | **18.8%** |
| *Total sent* | 3536 | 3993 | 5066 | 7507 | 3657 | 4072 | 3953 | 31,784 | |
| **% completed** | **17.5%** | **15.7%** | **12.4%** | **14.7%** | **18.2%** | **19.2%** | **17.6%** | | **16.1%** |
| | | | | | | | | | |
| **Invite 3** | ***Day of Week*** | | | | | | | | |
| | **Sun.** | **Mon.** | **Tue.** | **Wed.** | **Thu.** | **Fri.** | **Sat.** | **Total sent** | **% completed** |
| *Time of Day* | | | | | | | | | |
| Before 9am | 1504 | 2722 | 1958 | 4747 | 2387 | 2651 | 2349 | 18,318 | **9.1%** |
| 10am-11am | 1205 | 446 | 1375 | 554 | 88 | 48 | 0 | 3716 | **12.1%** |
| 12pm-1pm | 25 | 10 | 855 | 163 | 46 | 292 | 744 | 2135 | **14.7%** |
| After 1pm | 120 | 132 | 170 | 777 | 404 | 243 | 64 | 1910 | **18.2%** |
| *Total sent* | 2854 | 3310 | 4358 | 6241 | 2925 | 3234 | 3157 | 26,079 | |
| **% completed** | **10.3%** | **9.5%** | **8.5%** | **10.4%** | **11.5%** | **15.0%** | **10.5%** | | **10.7%** |

[a] Excludes the 7 participants who died shortly after Q6web recruitment began.

## Table 2a.

Participant characteristics associated with Q6web response

| | Univariate | | Multivariate* | |
|---|---|---|---|---|
| | **OR** | **95% CI** | **OR** | **95% CI** |
| **Age** [a] | | | | |
| *<65* | 1.00 | Ref. | 1.00 | Ref. |
| *65 – 69* | 1.25 | 1.18-1.32 | 1.19 | 1.12-1.26 |
| *70 – 74* | 1.24 | 1.18-1.31 | 1.16 | 1.09-1.22 |
| *75 – 79* | 1.11 | 1.04-1.17 | 1.01 | 0.95-1.07 |
| *80+* | 0.50 | 0.47-0.54 | 0.46 | 0.43-0.49 |
| | | | | |
| **Race** | | | | |
| Non-Hispanic White | 1.00 | Ref. | 1.00 | Ref. |
| Non-white | 0.72 | 0.68-0.76 | 0.74 | 0.70-0.79 |
| | | | | |
| **Cancer History** | | | | |
| No history of cancer | 1.00 | Ref | 1.00 | Ref. |
| Cancer survivor at Q6 | 1.06 | 0.98-1.14 | 0.94 | 0.87-1.02 |
| | | | | |
| **Previous Response** | | | | |
| *Completed Q5* | | | | |
| No | 1.00 | Ref. | 1.00 | Ref. |
| Yes | 6.07 | 5.50-6.70 | 6.01 | 5.45-6.63 |
| *Completed all previous Qs* | | | | |
| No | 1.00 | Ref. | 1.00 | Ref. |
| Yes | 3.06 | 2.93-3.20 | 3.04 | 2.90-3.17 |
| | | | | |

**Table 2b.**

Recruitment characteristics associated with Q6web response

| | | | | |
|---|---|---|---|---|
| **Invite 1 Time** | | | | |
| *Before 9AM* | 1.00 | Ref. | 1.00 | Ref. |
| *Hour 1 (10 – 11 AM)* | 1.27 | 1.20-1.33 | 1.14 | 1.09-1.21 |
| *Hour 2 (12 – 1 PM)* | 1.38 | 1.31-1.46 | 1.30 | 1.23-1.37 |
| *Hour 3 (2 PM +)* | 1.75 | 1.65-1.84 | 1.51 | 1.43-1.60 |
| **Invite 1 Day** | | | | |
| *Monday* | 1.27 | 1.19-1.35 | 1.20 | 1.13-1.29 |
| *Tuesday* | 1.00 | Ref. | 1.00 | Ref. |
| *Wednesday* | 1.21 | 1.13-1.29 | 1.26 | 1.17-1.35 |
| *Thursday* | 1.10 | 1.03-1.18 | 1.16 | 1.08-1.24 |
| *Friday* | 1.04 | 0.97-1.11 | 1.04 | 0.97-1.11 |
| *Saturday* | 1.48 | 1.36-1.60 | 1.34 | 1.24-1.46 |
| *Sunday* | 1.32 | 1.23-1.42 | 1.22 | 1.13-1.31 |

### Table 2c.

Recruitment characteristics associated with Q6web response adjusted for both time of day and day of week and stratified by Q5 response

| | | | | |
|---|---|---|---|---|
| | ***Multivariate model including both time of day & day of week*** | | | |
| | **Invite 1 Time** | | | |
| | *Before 9AM* | 1.00 | Ref. | |
| | *10 – 11 AM* | 1.20 | 1.14-1.27 | |
| | *12 – 1 PM* | 1.34 | 1.26-1.42 | |
| | *2 PM +* | 1.80 | 1.69-1.92 | |
| | **Invite 1 Day** | | | |
| | *Monday* | 1.33 | 1.24-1.43 | |
| | *Tuesday* | 1.00 | Ref. | |
| | *Wednesday* | 1.53 | 1.42-1.65 | |
| | *Thursday* | 1.25 | 1.17-1.35 | |
| | *Friday* | 1.37 | 1.26-1.48 | |
| | *Saturday* | 1.79 | 1.64-1.96 | |
| | *Sunday* | 1.45 | 1.34-1.57 | |
| | | | | |
| **Among Q5 Responders** | | | | |
| | ***Multivariate model including both time of day & day of week*** | | | |
| | **Invite 1 Time** | | | |
| | *Before 9AM* | 1.00 | Ref. | |
| | *10 – 11 AM* | 1.18 | 1.11-1.25 | |
| | *12 – 1 PM* | 1.32 | 1.24-1.40 | |
| | *2 PM +* | 1.80 | 1.69-1.93 | |
| | **Invite 1 Day** | | | |
| | *Monday* | 1.36 | 1.27-1.47 | |
| | *Tuesday* | 1.00 | Ref. | |
| | *Wednesday* | 1.52 | 1.41-1.64 | |
| | *Thursday* | 1.27 | 1.18-1.36 | |
| | *Friday* | 1.41 | 1.30-1.53 | |
| | *Saturday* | 1.82 | 1.66-1.99 | |
| | *Sunday* | 1.49 | 1.37-1.61 | |
| | | | | |
| **Among Q5 Non-Responders** | | | | |
| | ***Multivariate model including both time of day & day of week*** | | | |
| | **Invite 1 Time** | | | |
| | *Before 9AM* | 1.00 | Ref. | |
| | *10 – 11 AM* | 2.44 | 1.78-3.32 | |
| | *12 – 1 PM* | 1.80 | 1.37-2.36 | |

| | | | | |
|---|---|---|---|---|
| | ***Multivariate model including both time of day & day of week*** | | | |
| | *2 PM +* | 1.93 | 1.27-2.93 | |
| | **Invite 1 Day** | | | |
| | *Monday* | 0.89 | 0.58-1.37 | |
| | *Tuesday* | 1.00 | Ref. | |
| | *Wednesday* | 2.22 | 1.55-3.18 | |
| | *Thursday* | 1.32 | 0.93-1.86 | |
| | *Friday* | 0.96 | 0.63-1.45 | |
| | *Saturday* | 1.00 | 0.50-2.00 | |
| | *Sunday* | 0.70 | 0.37-1.33 | |

[a] Adjusted for age, race, cancer status, and having completed Q5.