LINKING WEBSITE EXPOSURE DATA TO SURVEY DATA: A SINGLE-SOURCE SOLUTION

Eric Melton and Jayne Krahn, Kantar Media
Jeff Landi, Safecount

Summary

Kantar Media in partnership with Safecount produced a test of a passive, direct, and involuntary approach for the collection of visitation behavior focused primarily on health-related websites. The approach involved a single read of available cookie data. The results were matched back to survey data that include website visitation behavior collected via respondent recall. The results did not produce strong agreement between data collected passively versus recalled. There exists attrition of cookie-based usage history as that history moves farther into the past.

Background

The MARS (Multimedia Audience Research Systems) OTC/DTC Pharmaceutical Study is a syndicated national multimedia and product usage survey conducted annually in the United States since 2001. As a print service, MARS has been accredited by the Media Rating Council since 2006. MARS includes audience estimates for 102 magazines and 4 national newspapers as well as media consumption information on television broadcast, radio, Internet, and various out-of-home opportunities for advertising exposure. The study also includes detailed information on over 70 ailments plus treatment options, drug brand usage for prescription and over-the-counter remedies as well as extensive psychographic, attitudinal, and behavioral information of interest to the pharmaceutical, healthcare, and medical insurance industries. The value of MARS to users is the ability to estimate the size of ailment target groups while providing an understanding of who the group members are and what media they use. Sampling is address-based supplemented with households with known ailment sufferers drawn from convenience lists and covers the adult general population, age 18 or older, in fifty states plus the District of Columbia; the methodology is a mail/paper-based self-administered interview. The latest study was in the field January to March 2011 and released April 2011. The total sample size is 20,539 respondents.

The MARS Online Behavior Study (OBS) is a re-contact of the subset of MARS respondents identified through study responses as Internet Users (accessed in the last 30 days). The OBS was first introduced to the market in 2010. The stated objectives of the study are twofold: One, to provide insight into how the Internet, in particular, influences consumers' health and wellness decisions. Two, to provide information about how Internet Users use various information channels, online and offline, for their research during various stages of an illness or condition. Survey content emphasizes different channels within the online space such as general health-related websites, drug manufacturer websites, search, and online communities/social networking. Respondents identified in MARS as Internet Users are invited by mail and email to participate in a web-based self-administered interview. Completed interview data are appended to the MARS data and reweighted to the set of U.S. adult Internet Users to create a new research database that is an extension of MARS. In the 2011 MARS release, 14,880 respondents were identified as Internet Users and invited to participate in OBS. In the latest OBS study, data was collected from February to May 2011 and released June 2011. The total sample size is 5,703 respondents.

In and of itself, paper/mail-based interviewing is not conducive to passive measurement, but conditions in the U.S. are not yet supportive of a wholesale migration of the project to online self-administered interviewing. MARS is a general population study, and in 2011 Internet penetration is approaching 80%. But penetration in a prime target group, prescription drug users aged 50 years or older, remains less than 60%.

The introduction of OBS conducted via an online instrument against Internet Users allows for the use of passive digital media measurement technologies. Safecount is a sister company of Kantar Media within the Kantar family that provides online passive measurement solutions. Safecount employs experts in live web data collection, and their mission is to improve transparency in digital data collection for consumers, publishers, and researchers. The company has existing relationships with hundreds of website publishers/media owners who carry Safecount's research tags for the purposes of intercept sampling and surveying visitors. Based on the breadth of tracking and the number of websites and networks across which Safecount operates, an estimated 80% of the U.S. online population is covered.

Test Approach

Safecount has a number of different solutions for the collection, warehousing, and analysis of online behavior data. However, the internal goal for the research was modest. Kantar Media wished to know how long the OBS respondent pool would need to be tracked in order to report reliable website visitation data for a short list of websites of interest to users of MARS and OBS, primarily health-related websites. Safecount and Kantar Media determined that a quick and cost effective approach to answering this research question would be to insert a survey tag or web beacon at the end of the OBS instrument. The survey tag instructed the respondent's browser software to read each Safecount-owned cookie on the system and pass all available website exposure history for the short list of websites to a database within Safecount's systems. In essence, the approach "piggybacks" on Safecount's existing online ad tracking activities. The technology solution is platform neutral, working in all known browser software. Kantar Media would develop visitation accumulation curves by analyzing the exposure data backward in time from the point the respondent encountered the survey tag. This information taken from the recent past would give some idea of visitation accumulation if Kantar Media were to use a real-time tracking solution.

Although it was not an original research objective, a happy consequence of Kantar Media's exploration with this passive approach is that the resulting online behavior data may be matched with data collected in both MARS and OBS where website usage is recorded on a respondent recall basis. Kantar Media can participate in a discussion and comparison of passive and recall-based website behavior collection outcomes.

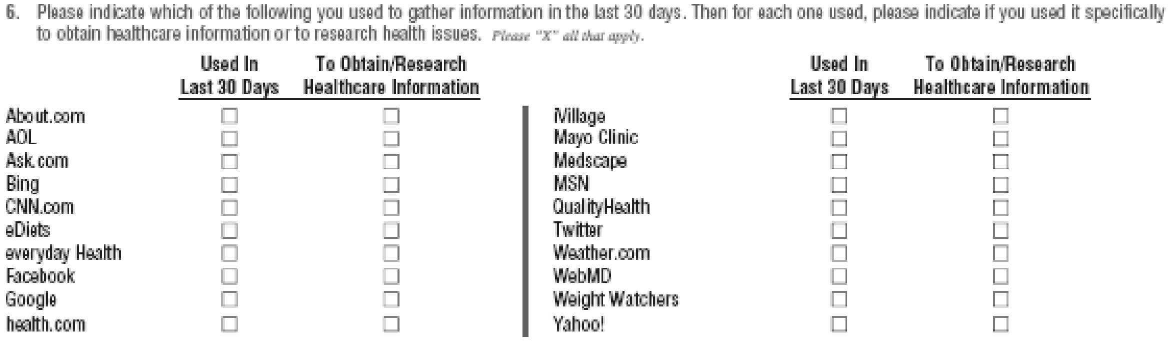
The approach is passive in that no respondent recall through survey questions is required and that there is no respondent interaction needed to do the tasks of collecting and recording online behavior data. The approach is direct in that the Safecount survey tag was configured so that the OBS respondent's unique ID would also be passed to Safecount's systems. The research results were directly matched back to the OBS data using the unique ID as a key, creating a single-source database. The cookies read by the survey tag are those associated with the same browser software and operating system user account used to complete the OBS survey. Available cookie information is passed to Safecount immediately upon the respondent's exposure to the survey tag. Depending on the specific browser software used and the features that are supported and enabled, the survey tag has the ability to cross read cookies placed with other browser software installations on the same computer and user account. Regular simultaneous use of multiple browser software among the general population of Internet Users is not common. A validation step during OBS data processing compares personally identifying demographic data collected in OBS to the same information collected in MARS. OBS respondents that do not have matching demographic data with MARS are not tabulated in the OBS report. Identity issues can exist with the approach but the validation step is the way to best ensure that the passive data are associated with the correct respondent. The approach is involuntary in that no respondent opt-in was presented in the OBS instrument. All respondents exposed to the survey tag have available Safecount-owned cookies read. Note that Safecount does operate a separate program at http://www.safecount.net that allows individuals to opt-out of Safecount tracking and surveying.

Fielding with the Safecount survey tag took place from March to May 2011. As OBS had commenced fielding in February, some respondents completed the OBS interview but were not exposed to the Safecount survey tag. Kantar Media fielded a mini (1 screen) online survey with two questions with the main intent to recontact OBS respondents and expose those to the Safecount survey tag who had not been exposed in the OBS interview. Had the Safecount survey tag been incorporated in the OBS instrument prior to fielding, all 5,703 OBS respondents would have been exposed to the Safecount survey tag. Including respondents exposed to the survey tag in the mini online survey, a total of 5,007 respondents were exposed to the Safecount survey tag.

Safecount collected all available visitation information on the following 22 website domains or subdomains. At the time of data collection, Safecount already had existing agreements with all of the properties for them to carry Safecount's research tags:

AARP.org
About.com: Health
AOL Health
Everyday Health
FitnessMagazine.com
Health.com
HealthGrades
Healthline
iVillage Health
Mayo Clinic
Medscape
MensHealth.com
MSN Health
The New York Times
Prevention.com
USA Today
The Wall Street Journal (wsj.com)
Weather.com
WebMD
WeightWatchers.com
WomensHealthMag.com
Yahoo! Health

MARS collects website usage information based on a “last 30 days” timeframe within a paper-based questionnaire:

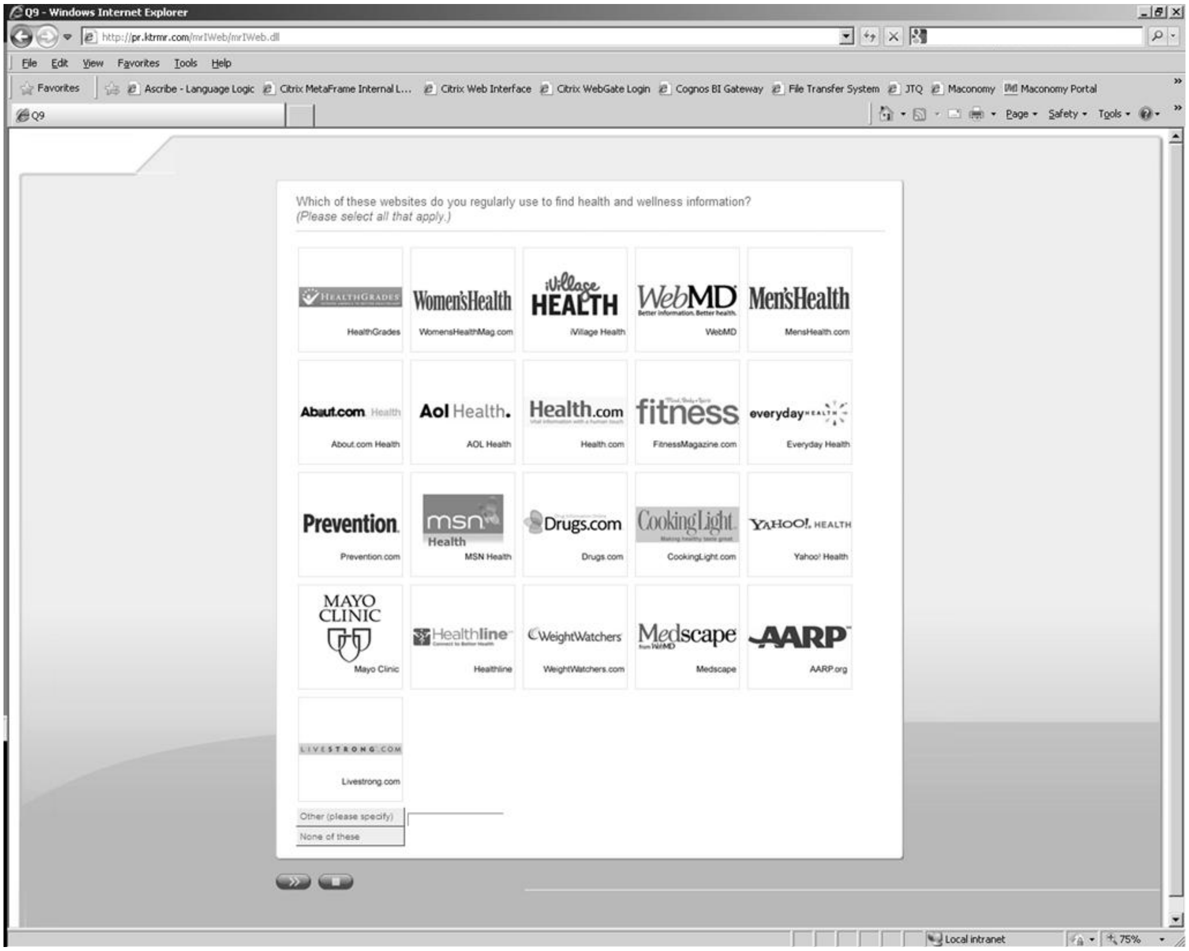


6. Please indicate which of the following you used to gather information in the last 30 days. Then for each one used, please indicate if you used it specifically to obtain healthcare information or to research health issues. *Please "X" all that apply.*

	Used In Last 30 Days	To Obtain/Research Healthcare Information		Used In Last 30 Days	To Obtain/Research Healthcare Information
About.com	□	□	iVillage	□	□
AOL	□	□	Mayo Clinic	□	□
Ask.com	□	□	Medscape	□	□
Bing	□	□	MSN	□	□
CNN.com	□	□	QualityHealth	□	□
eDiets	□	□	Twitter	□	□
everyday Health	□	□	Weather.com	□	□
Facebook	□	□	WebMD	□	□
Google	□	□	Weight Watchers	□	□
health.com	□	□	Yahoo!	□	□

Question 6 is presented on page 9 of a 24 page questionnaire. Editing rules apply based on whether the respondent checked “No” for a question appearing above question 6 on the same page that asks if the respondent had accessed the Internet in the last 30 days. Any data were cleared. But by definition all OBS respondents are Internet Users, having checked “Yes” for accessing the Internet in the last 30 days. In the 2011 MARS release, 1,259 out of the 14,880 respondents (1,259 ÷ 14,880 = 8%) identified as Internet Users left question 6 with no boxes checked even though the response list includes high traffic websites such as Google, Bing, Facebook, MSN, and Yahoo!. Note that About.com, AOL, iVillage, MSN, and Yahoo! are asked at the domain level. It may be inferred that the obtaining of healthcare information or researching of health issues occurs at the health-related subdomain level.

OBS collects website usage in the context of finding health and wellness information:



Which of these websites do you regularly use to find health and wellness information?
(Please select all that apply.)

HealthGrades	WomensHealthMag.com	iVillage Health	WebMD	MensHealth.com
About.com Health	AOL Health	Health.com	FitnessMagazine.com	Everyday Health
Prevention.com	MSN Health	Drugs.com	CookingLight.com	Yahoo! Health
Mayo Clinic	Healthline	WeightWatchers.com	Medscape	AARP.org
Livestrong.com				

Other (please specify)

None of these

Question Q9 is presented near the end of an online interview with a median length of 20 minutes. The presentation order of websites is randomized. Logos are presented in full color. No particular timeframe is invoked in asking this question although the first question of the OBS interview asks if the respondent used the Internet for health and wellness research in the past year. Respondents who selected "I never use the Internet for health and wellness research" in response to a frequency question, "How often do you typically use the Internet to research health and wellness topics?" are not presented Q9. In the 2011 OBS release, 351 respondents out of 5,703 tabulated selected "I never use the Internet for health and wellness research" and so were not presented Q9. Out of the remaining 5,352 respondents who indicated that they use the Internet to research health and wellness topics 691 respondents (691 ÷ 5,352 = 13%) did not select any of the 21 named websites.

Results

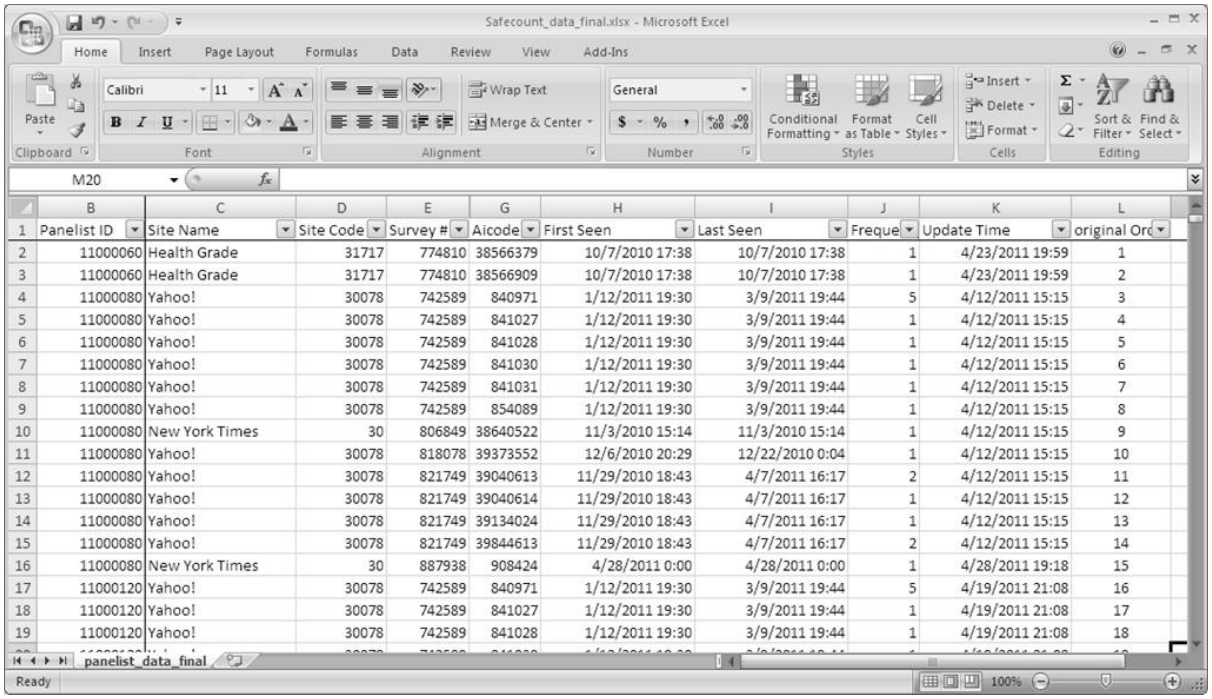
Out of the 5,007 respondents exposed to the survey tag, Safecount returned cookie-based passively observed data for 2,094 respondents (2,094 ÷ 5,007 = 42%) for at least one of the 22 websites Safecount covered. Due to limitations of the research design, if no passively observed data were recorded Kantar Media cannot know if there existed technological hurdles inhibiting the collection of data; if the respondent had opted-out of Safecount tracking; or if the respondent simply had no activity among the 22 websites. Kantar Media asked Safecount to cover the website properties of *The New York Times*, *The Wall Street Journal*, and *USA Today*, but neither MARS nor OBS collected website usage information for the national newspapers. Kantar Media does not have a basis for comparing this information against the same websites collected via respondent recall, so these observations were not included in the analysis. The analysis will focus on passive and respondent recall observations among 19 primarily health-related websites. Out of the 2,094 respondents with Safecount data, 2 respondents had only data associated with the website properties of *The New York Times*, *The Wall Street Journal*, or *USA Today* and no respondent recall among the 19 remaining websites. The 2 were excluded leaving 2,092 respondents remaining for the analysis.

The following table shows the 19 websites and the information available on a respondent recall basis:

	Safecount	**MARS**	**OBS**
AARP.org	*		*
About.com: Health	*	*	*
AOL Health	*	*	*
Everyday Health	*	*	*
FitnessMagazine.com	*		*
Health.com	*	*	*
HealthGrades	*		*
Healthline	*		*
iVillage Health	*	*	*
Mayo Clinic	*	*	*
Medscape	*	*	*
MensHealth.com	*		*
MSN Health	*	*	*
Prevention.com	*		*
Weather.com	*	*	
WebMD	*	*	*
WeightWatchers.com	*	*	*
WomensHealthMag.com	*		*
Yahoo! Health	*	*	*

MARS has recall data for 12 of the 19 websites, and OBS has recall data for 18 of the 19 websites.

Safecount delivered the recorded data in a Microsoft Excel file. The following image shows an example of the database structure taken from the first several records in the file:



	B	C	D	E	G	H	I	J	K	L
1	Panelist ID	Site Name	Site Code	Survey #	Aicode	First Seen	Last Seen	Freque	Update Time	original Orc
2	11000060	Health Grade	31717	774810	38566379	10/7/2010 17:38	10/7/2010 17:38	1	4/23/2011 19:59	1
3	11000060	Health Grade	31717	774810	38566909	10/7/2010 17:38	10/7/2010 17:38	1	4/23/2011 19:59	2
4	11000080	Yahoo!	30078	742589	840971	1/12/2011 19:30	3/9/2011 19:44	5	4/12/2011 15:15	3
5	11000080	Yahoo!	30078	742589	841027	1/12/2011 19:30	3/9/2011 19:44	1	4/12/2011 15:15	4
6	11000080	Yahoo!	30078	742589	841028	1/12/2011 19:30	3/9/2011 19:44	1	4/12/2011 15:15	5
7	11000080	Yahoo!	30078	742589	841030	1/12/2011 19:30	3/9/2011 19:44	1	4/12/2011 15:15	6
8	11000080	Yahoo!	30078	742589	841031	1/12/2011 19:30	3/9/2011 19:44	1	4/12/2011 15:15	7
9	11000080	Yahoo!	30078	742589	854089	1/12/2011 19:30	3/9/2011 19:44	1	4/12/2011 15:15	8
10	11000080	New York Times	30	806849	38640522	11/3/2010 15:14	11/3/2010 15:14	1	4/12/2011 15:15	9
11	11000080	Yahoo!	30078	818078	39373552	12/6/2010 20:29	12/22/2010 0:04	1	4/12/2011 15:15	10
12	11000080	Yahoo!	30078	821749	39040613	11/29/2010 18:43	4/7/2011 16:17	2	4/12/2011 15:15	11
13	11000080	Yahoo!	30078	821749	39040614	11/29/2010 18:43	4/7/2011 16:17	1	4/12/2011 15:15	12
14	11000080	Yahoo!	30078	821749	39134024	11/29/2010 18:43	4/7/2011 16:17	1	4/12/2011 15:15	13
15	11000080	Yahoo!	30078	821749	39844613	11/29/2010 18:43	4/7/2011 16:17	2	4/12/2011 15:15	14
16	11000080	New York Times	30	887938	908424	4/28/2011 0:00	4/28/2011 0:00	1	4/28/2011 19:18	15
17	11000120	Yahoo!	30078	742589	840971	1/12/2011 19:30	3/9/2011 19:44	5	4/19/2011 21:08	16
18	11000120	Yahoo!	30078	742589	841027	1/12/2011 19:30	3/9/2011 19:44	1	4/19/2011 21:08	17
19	11000120	Yahoo!	30078	742589	841028	1/12/2011 19:30	3/9/2011 19:44	1	4/19/2011 21:08	18

Panelist ID holds the unique OBS respondent ID needed as a key for matching the Safecount data back to MARS and OBS. The relevant information provided by Safecount is Site Name, Survey #, First Seen, Last Seen, and Frequency. Exposure to the Safecount survey tag in Kantar Media's online instruments instructs the respondent browser to read and pass along exposure information stored on *all* Safecount-owned cookies. The data reflect two aspects of Safecount's business: One, as an "always on" presence on a host website for the purposes of web intercept interviewing of visitors and data collection benefiting media property owners. Two, tracking advertising campaigns across media properties in order to provide campaign evaluation intelligence to advertisers. As such, the data show as separate records exposure to the vehicle as well as exposure to advertising carried by the vehicle. As the focus of the research is website level exposure, the analysis required addressing this duplication. Note that Safecount confirmed that Site Name = Yahoo! represents only activity on the Yahoo! Health subdomain.

Deduplication involved a process of grouping cases by Panelist ID, Site Name, Survey #, First Seen, Last Seen, and Frequency. Within case groups, cases were sorted in ascending order by the same set of variables. Within case groups, defined by Panelist ID, Site Name, Survey #, First Seen, and Last Seen, the last case in the group was retained. This process produces a dataset of unique website exposures within the levels of respondent and domain/subdomain.

The deduplicated dataset was further restructured to a respondent and website level dataset. Within case groups defined by Panelist ID and Site Name, the minimum of First Seen and the maximum of Last Seen were determined and the sum of Frequency calculated. Loss of information is associated with this restructuring step as exposure frequencies are not uniformly distributed within the respondent's history with any of the websites.

One caveat: the test approach reports exposure within a range of dates between First Seen and Last Seen. Again, one source of this date information is through ad campaign tracking. In an unknown number of website exposure records, the date information refers to first exposure and last exposure to the advertisement, and these exposures may have taken place on some other website. The accurate description of the data recorded is that the approach captures website exposure and frequency with these exposures taking place in points of time within the range of dates bounded by First Seen and Last Seen. The technological solution exists to capture exact timestamps for each website exposure, but the bandwidth required to do this would drastically increase, altering the cost of production for this test.

The approach involves collecting a full history of exposures to the short list of health-related websites. The range of exposure dates across the entire dataset is August 11, 2009 to May 25, 2011. Safecount records roughly 5 billion impressions in any given week, but focusing on activity for just the set of 19 websites Safecount cannot guarantee full monitoring coverage was achieved over the entirety of this timeframe. In addition, the usual obstacles should be considered when interpreting results using this approach: identity issues, non-coverage of other devices the respondent used for online activities, respondent opting-out of Safecount tracking at some point in the tracking history, etc. Cookie deletion/blocking is a particular concern. Prior research shows 35% of US-based computers deleting third-party cookies in a month and over 13% of US-based computers deleting third-party cookies four or more times in a month. (comScore, Inc., 2011)

The following table shows the aggregated sums of unique visitors to each website property across all observations. That is, the frequency counts were not filtered by recent behavior or any other particular timeframe:

	Safecount	MARS	OBS
AARP.org	61		316
About.com: Health	237	82	231
AOL Health	30	221	54
Everyday Health	158	16	58
FitnessMagazine.com	4		91
Health.com	36	69	126
HealthGrades	9		36
Healthline	39		50
iVillage Health	182	11	30
Mayo Clinic	1	122	451
Medscape	5	22	177
MensHealth.com	22		133
MSN Health	1,010	293	199
Prevention.com	8		292
Weather.com	411	567	
WebMD	277	439	1,330
WeightWatchers.com	37	99	225
WomensHealthMag.com	12		288
Yahoo! Health	1,293	973	380
None of the above	0	483	233

Expressed as vertical percentages of the base:

	Safecount	MARS	OBS
AARP.org	3%		16%
About.com: Health	11%	4%	12%
AOL Health	1%	11%	3%
Everyday Health	8%	1%	3%
FitnessMagazine.com	0%		5%
Health.com	2%	3%	6%
HealthGrades	0%		2%
Healthline	2%		3%
iVillage Health	9%	1%	2%
Mayo Clinic	0%	6%	23%
Medscape	0%	1%	9%
MensHealth.com	1%		7%
MSN Health	48%	14%	10%
Prevention.com	0%		15%
Weather.com	20%	27%	
WebMD	13%	21%	67%
WeightWatchers.com	2%	5%	11%
WomensHealthMag.com	1%		15%
Yahoo! Health	62%	47%	19%
None of the above	0%	23%	12%
TOTAL SAMPLE	**2,092**	**2,092**	**1,981**

In MARS, 483 respondents (483 ÷ 2,092 = 23%) did not check any of the 12 websites available for the analysis. Out of the 483 MARS respondents, 126 did not select any of the 20 websites measured in MARS including high traffic websites such as Google, Facebook, Bing, MSN, and Yahoo! Note that 111 OBS respondents were never exposed to the recall question because these respondents claimed no usage of the Internet for health-related research. For the calculation of this table, the 111 respondents were removed from the OBS base. In OBS, 233 respondents (233 / 1,981 = 12%) did not select any of the 18 websites available for the analysis. Out of the 233 OBS respondents, 222 did not select any of the 21 websites measured in OBS.

Matching Safecount data to MARS and OBS at the respondent and website level allows for the counting of instances where any visitation was recorded in both Safecount and MARS and in both Safecount and OBS:

	Safecount	MARS	OBS		Safecount ∩ MARS	Safecount ∩ OBS
AARP.org	61		316			26
About.com: Health	237	82	231		12	28
AOL Health	30	221	54		13	3
Everyday Health	158	16	58		1	4
FitnessMagazine.com	4		91			3
Health.com	36	69	126		2	3
HealthGrades	9		36			0
Healthline	39		50			1
iVillage Health	182	11	30		6	9
Mayo Clinic	1	122	451		0	1
Medscape	5	22	177		3	4
MensHealth.com	22		133			5
MSN Health	1,010	293	199		242	149
Prevention.com	8		292			0
Weather.com	411	567			202	
WebMD	277	439	1,330		99	205
WeightWatchers.com	37	99	225		21	28
WomensHealthMag.com	12		288			5
Yahoo! Health	1,293	973	380		799	311

Expressing the intersection frequency information in the table above as a percent within Safecount's total frequencies then the intersection frequencies as a percent of MARS total frequencies and OBS total frequencies:

	∩ MARS as % of Safecount	∩ OBS as % of Safecount		∩ MARS as % of MARS	∩ OBS as % of OBS
AARP.org		43%			8%
About.com: Health	5%	12%		15%	12%
AOL Health	43%	10%		6%	6%
Everyday Health	1%	3%		6%	7%
FitnessMagazine.com		75%			3%
Health	6%	8%		3%	2%
HealthGrades		0%			0%
Healthline		3%			2%
iVillage Health	3%	5%		55%	30%
Mayo Clinic	0%	100%		0%	0%
Medscape	60%	80%		14%	2%
MensHealth.com		23%			4%
MSN Health	24%	15%		83%	75%
Prevention.com		0%			0%
Weather.com	49%			36%	
WebMD	36%	74%		23%	15%
WeightWatchers.com	57%	76%		21%	12%
WomensHealthMag.com		42%			2%
Yahoo! Health	62%	24%		82%	82%

MARS collects website exposure data in terms of usage in the last 30 days. By incorporating the paper survey return date recorded by the MARS paper survey scanhouse, the 30 day period in which any recalled exposure might have taken place may be isolated on the calendar. One day was taken away from the survey return date in order to account for some minimum delivery time while the physical paper survey was in transit from respondent to the scanhouse. The OBS record includes a timestamp indicating the date the respondent completed the online interview. As the timeframe for the OBS research is focused on health and wellness research done in the past year, the year long period in which any recalled exposure might have taken place may be isolated on the calendar. The analysis was refined to identify instances where there is any possible temporal overlap in exposures recorded passively and exposures recalled. Note that About.com, AOL, iVillage, MSN, and Yahoo! are asked at the domain level. To better ensure the isolation of the MARS data to health subdomains respondent selections for "To Obtain/Research Healthcare Information" were used instead of "Used In Last 30 Days":

	Safecount "last 30 days"	**MARS**		**Safecount "last year"**	**OBS**
AARP.org				54	316
About.com: Health	62	31		227	231
AOL Health	0	36		28	54
Everyday Health	31	16		142	58
FitnessMagazine.com				3	91
Health.com	5	69		35	126
HealthGrades				9	36
Healthline				35	50
iVillage Health	36	1		155	30
Mayo Clinic	0	122		1	451
Medscape	0	22		3	177
MensHealth.com				18	133
MSN Health	357	35		864	199
Prevention.com				7	292
Weather.com	165	567			
WebMD	50	439		240	1,330
WeightWatchers.com	13	99		35	225
WomensHealthMag.com				11	288
Yahoo! Health	506	153		1,148	380
None of the above	0	1,024		0	233
TOTAL	**854**	**2,092**		**1,783**	**1,981**

Expressed as vertical percentages of the base:

	Safecount "last 30 days"	**MARS**		**Safecount "last year"**	**OBS**
AARP.org				3%	16%
About.com: Health	7%	1%		13%	12%
AOL Health	0%	2%		2%	3%
Everyday Health	4%	1%		8%	3%
FitnessMagazine.com				0%	5%
Health.com	1%	3%		2%	6%
HealthGrades				1%	2%
Healthline				2%	3%
iVillage Health	4%	0%		9%	2%
Mayo Clinic	0%	6%		0%	23%
Medscape	0%	1%		0%	9%
MensHealth.com				1%	7%
MSN Health	42%	2%		48%	10%
Prevention.com				0%	15%
Weather.com	19%	27%			
WebMD	6%	21%		13%	67%
WeightWatchers.com	2%	5%		2%	11%
WomensHealthMag.com				1%	15%
Yahoo! Health	59%	7%		64%	19%
None of the above	0%	49%		0%	12%
TOTAL	**854**	**2,092**		**1,783**	**1,981**

Note that zero instances of AOL Health were observed in the Safecount “last 30 days” analysis. AOL Health was folded into Huffington Post in early February 2011 and from that point in time no longer exists as a subdomain called “AOL Health”.

Excluding AOL Health and calculating Pearson’s product-moment correlation coefficients at the respondent and website level, for the comparison against MARS observations “last 30 days”, $r = 0.17$; for the comparison against OBS observations “last year”, $r = 0.33$. Some linear relationship exists between the sets of observations, but the measures are not strong.

Counting instances where visitation was recorded in both Safecount and MARS and in both Safecount and OBS within intervals defined by MARS “last 30 days” and OBS “last year”:

	Safecount "last 30 days"	**MARS**		**Safecount "last year"**	**OBS**		**Safecount "last 30 days" ∩ MARS**	**Safecount "last year" ∩ OBS**
AARP.org				54	316			23
About.com: Health	62	31		227	231		0	27
AOL Health	0	36		28	54		0	3
Everyday Health	31	16		142	58		0	4
FitnessMagazine.com				3	91			3
Health.com	5	69		35	126		1	3
HealthGrades				9	36			0
Healthline				35	50			1
iVillage Health	36	1		155	30		1	9
Mayo Clinic	0	122		1	451		0	1
Medscape	0	22		3	177		0	3
MensHealth.com				18	133			4
MSN Health	357	35		864	199		12	125
Prevention.com				7	292			0
Weather.com	165	567					81	
WebMD	50	439		240	1,330		30	184
WeightWatchers.com	13	99		35	225		11	28
WomensHealthMag.com				11	288			4
Yahoo! Health	506	153		1,148	380		42	276

Expressing the intersection frequency information in the table above as a percent within Safecount’s total frequencies then the intersection frequencies as a percent of MARS total frequencies and OBS total frequencies:

	∩ MARS as % of Safecount “last 30 days”	**∩ OBS as % of Safecount “last year”**		**∩ MARS as % of MARS**	**∩ OBS as % of OBS**
AARP.org		43%			7%
About.com: Health	0%	12%		0%	12%
AOL Health	N/A	11%		0%	6%
Everyday Health	0%	3%		0%	7%
FitnessMagazine.com		100%			3%
Health.com	20%	9%		1%	2%
HealthGrades		0%			0%
Healthline		3%			2%
iVillage Health	3%	6%		100%	30%
Mayo Clinic	N/A	100%		0%	0%
Medscape	N/A	100%		0%	2%
MensHealth.com		22%			3%
MSN Health	3%	14%		34%	63%
Prevention.com		0%			0%
Weather.com	49%			14%	
WebMD	60%	77%		7%	14%
WeightWatchers.com	85%	80%		11%	12%
WomensHealthMag.com		36%			1%
Yahoo! Health	8%	24%		27%	73%

It is possible to further isolate within some 30 day time period or some yearly time period the passively collected observations by using Frequency to calculate probabilities of exposure within the time interval defined by the minimum of First Seen and the maximum of Last Seen in the Safecount data. However, even at the level of possible matching exposures through overlapping time intervals, there is not uniform agreement between observations collected passively and through respondent recall. Intersection frequencies for most websites are less than 10 observations.

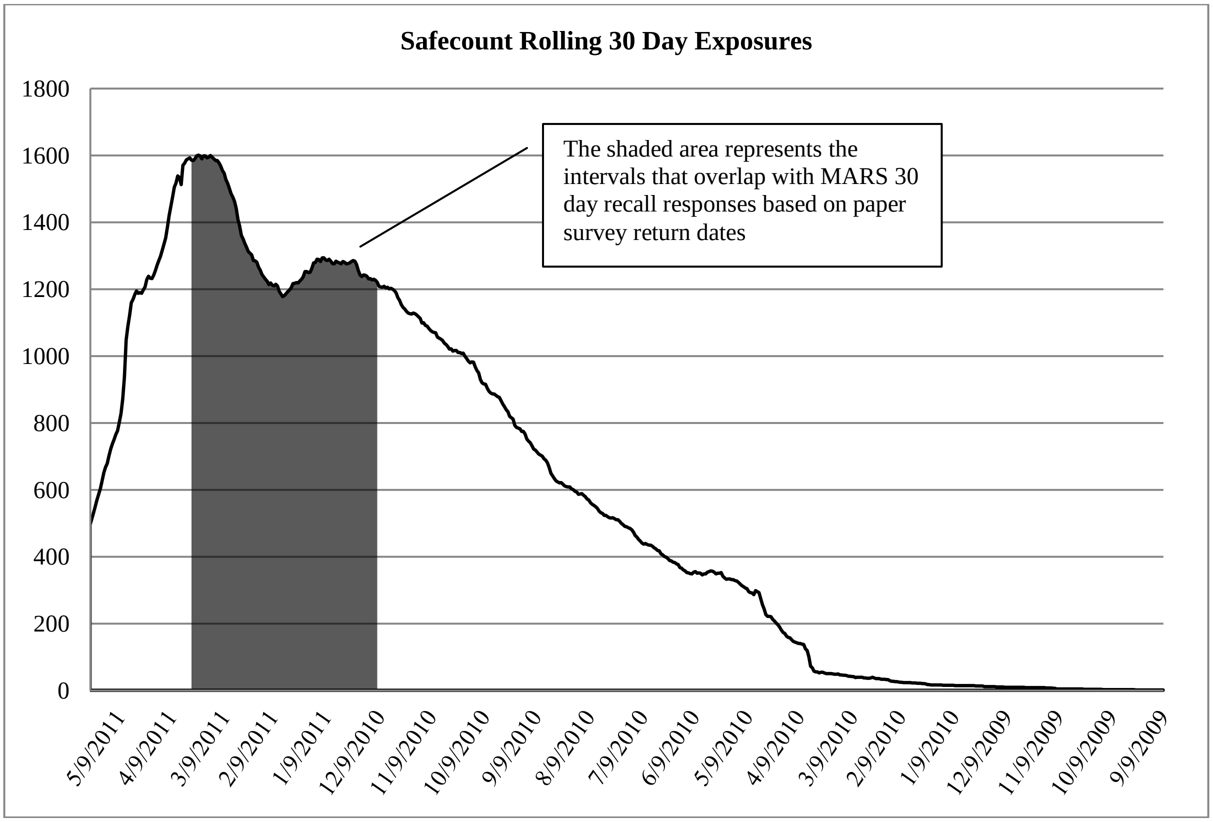
An extreme instance illustrates a fundamental difference between passive data collection and respondent recall. The top Safecount passively collected website is Yahoo! Health (1,148 observations). Yahoo! Health is an example of a website Kantar

Media terms a Portal Health Subdomain. The top OBS recall website is WebMD (1,330 observations). WebMD is an example of a website Kantar Media terms a Healthcare Destination Website. With Safecount, WebMD is a distant third in rank (240 observations), and with OBS, Yahoo! Health is third in rank (380 observations). The set of respondents who Safecount observed passively visiting Yahoo! Health is not necessarily different from the set of respondents selecting WebMD as a website they would "regularly use to find health and wellness information"; 749 respondents have both passive observations of Yahoo! Health and respondent recall of WebMD. However, as seen in the table above, the results demonstrate that there is not much agreement between passively observed data and respondent recall data for Yahoo! Health and WebMD (24% of Safecount's Yahoo! Health observations also have OBS Yahoo! Health observations, and 14% of OBS WebMD observations were validated with Safecount WebMD observations). Within the 749 respondents, 195 respondents (195 ÷ 749 = 26%) have OBS Yahoo! Health observations and 117 respondents (117 ÷ 749 = 16%) have Safecount WebMD observations.

The implication is that with the OBS recall question, asked in terms of websites where respondents may find health and wellness information, respondents select Healthcare Destination Websites, such as WebMD and Mayo Clinic, known to carry such information. Further, respondents recall these sites but many perhaps do not "regularly use" these sites: either not use these sites at all or do not use them regularly enough to be validated with passive observations. Conversely, the test approach passively detects usage of Portal Health Subdomains, such as Yahoo! Health and MSN Health, which many respondents either do not recall or do not consider sources for health and wellness information. That is, respondents may be reaching Portal Health Subdomains by following links from the main portal website or by visiting them in the course of casual reading apart from activities respondents consider serious health and wellness research.

As the OBS is served after the completion of the MARS paper-base survey, a byproduct of the research design implies that passive observations more recent than the last 30 day interval defined by the MARS paper survey return date were recorded. Similarly, the use of the mini online recontact study means that passive observations more recent than the last year interval defined by the OBS interview completion timestamp were recorded.

As mentioned earlier, the range of exposure dates across the entire Safecount dataset is August 11, 2009 to May 25, 2011. For each day in this range, Kantar Media determined the date 30 days earlier and then calculated the total number of observations in the 30 day period. That is, the number of Safecount observations was counted for the period of April 26 to May 25, 2011, then the number of Safecount observations were counted for the period of April 25 to May 24, 2011, and working through until the time period August 11 to September 9, 2009. The results were plotted using the latest date in each 30 day period as the x-axis point:



The chart demonstrates a general trend of data attrition looking into the past. The most relevant portion of the chart is where the rolling 30 day periods overlap with the responses captured with MARS via respondent recall. At these points on the calendar, the number of observations available from Safecount was not uniform but declined from a maximum of 1,600 observations for the 30 day period ending on March 24, 2011, to a minimum of 1,177 for the 30 day period ending February 3, 2011, a decrease of 26%. Loss of information due to loss of cookie data reduces the opportunity of validating respondent recall observations with passively collected observations. A known cause for this loss of cookie data is cookie deletion.

Conclusions

Respondent recall via self-administered interviewing has its downsides. By definition, control of the interview is entirely in the respondent's hands, and there is no opportunity to prompt or to clarify questions. A result is item non-response: 8% of MARS Internet Users left the recall question blank even though the response list includes high traffic websites such as Google, Bing, Facebook, MSN, and Yahoo!; and 13% of OBS respondents who stated that they had used the Internet for health and wellness research did not select any named websites in the recall question. It is not plausible that the true incidences of non-usage of these high traffic websites is this high.

The test results show no substantial agreement between passively observed data and respondent recall. The OBS question asks respondents to select from a set of websites they would "regularly use to find health and wellness information". Respondents select websites which are familiar as sources of health and wellness information, but which they either do not actually visit or visit regularly enough to validate with passive observations. At the same time, respondents do not select other websites for which the results show passive observations and are carriers of health-related advertising because respondents do not perceive such websites as places they would turn to for serious health and wellness research. The drawbacks and risks of respondent recall inspire a wish to get away from respondent perceptions and potential misinterpretations and simply observe actual behavior. Yet, passively collected information says little about respondent intent. The OBS survey delves more deeply, asking about which characteristics of the websites selected make those websites attractive as sources for health and wellness information. Survey data are still needed to reveal what is happening that drives the respondent to visit (and if not to visit then to recognize the attributes of) one website over another.

The choice of this particular passive, direct, and involuntary approach has the benefit of collecting a respondent's usage history in one single read executed through a simple survey tag incorporated into an online survey program. The approach is relatively less disruptive than recruiting respondents for MARS-related research and then making appeals to join a user-based panel. Other than recruitment to an online interview no other cooperation is technically required. All respondents exposed to the Safecount survey tag have all Safecount-owned cookies read.

The test approach has its downside as well. There exists attrition of cookie-based usage history as that history moves farther into the past. A large contributing factor to this loss is due to cookie deletion. In making a comparison against the MARS recall data, it should be noted that the collection of passively collected Safecount data occurs on a recontact and recruitment of MARS respondents to participate in a separate online interview. The average gap in time from the collection of MARS data to the collection of Safecount data was 75 days. The situation was exacerbated by the use of a mini online survey instrument to expose OBS respondents to the Safecount survey tag who had completed the interview prior to incorporation of the survey tag in the OBS survey instrument. Excluding the results of the mini online survey, the average gap time was 64 days. Further streamlining processes to reduce the gap between MARS data collection and Safecount data collection may improve the comparability of the data from the two sources.

The test focused on the results of data collected for 19 primarily health-related websites, but to implement the approach in a production capacity the list of websites must be expanded. Safecount has existing relationships with hundreds of website publishers/media owners who carry Safecount's research tags for the purposes of intercept sampling and surveying visitors. The test approach reports exposure within a range of dates, but this range is imprecise. The exposure information is drawn from ad campaign tracking activities, and date information may refer to first exposure and last exposure to the advertisement with these exposures possibly taking place on some other website. The technology exists to capture exact timestamps for each exposure at domain, subdomain, and even page level yielding standard online metrics such as frequency, time spent, and page views. The bandwidth required to cover all exposures to all available websites would increase exponentially, greatly altering the cost of production from that of this small test. But the focus of MARS is providing data of use in supporting the pharmaceutical and healthcare industries, and so when it comes to the efforts of Kantar Media to passively capture the online behavior of MARS respondents, the coverage would be limited to measures primarily of health-related websites. It remains to be determined what is economically feasible to measure for users of MARS data.

Reference

comScore, Inc. (2011, February 3). *The Impact of Cookie Deletion on Site-Server and Ad-Server Metrics in Australia: An Emperical comScore Study*. Retrieved August 16, 2011, from comScore: http://www.comscore.com/Press_Events/Presentations_Whitepapers/2011/The_Impact_of_Cookie_Deletion_on_Site-Server_and_Ad-Server_Metrics_in_Australia_An_Empirical_comScore_Study