

# Does Pornography-Blocking Software Block Access to Health Information on the Internet?

Caroline R. Richardson, MD

Paul J. Resnick, PhD

Derek L. Hansen, BS

Holly A. Derry, MPH

Victoria J. Rideout, MA

**T**HE INTERNET HAS BECOME AN important tool for many individuals with health concerns,<sup>1</sup> especially adolescents.<sup>2</sup> Teenagers grapple with sensitive health issues, including depression, substance abuse, and birth control. Concerns about confidentiality, accentuated by many teens not yet having their own health provider, make adolescents' access to information via the Internet particularly important. Given rapidly expanding Internet access, it is not surprising that more than 70% of 15- to 17-year-olds say they have used the Internet to look up health information.<sup>3</sup> Almost half have researched traditional health topics such as cancer or diabetes. About 40% of adolescents have searched for information on a sexual health topic such as pregnancy, birth control, human immunodeficiency virus/acquired immunodeficiency syndrome, or other sexually transmitted diseases; 1 in 4 have researched problems with drugs or alcohol; 17% have searched for information on depression or mental illness; and 11% have searched for information on sexual assault.<sup>3</sup>

In 2000, the US Congress passed the Child Internet Protection Act (CIPA) mandating that schools and libraries install pornography-blocking software on

**Context** The Internet has become an important tool for finding health information, especially among adolescents. Many computers have software designed to block access to Internet pornography. Because pornography-blocking software cannot perfectly discriminate between pornographic and nonpornographic Web sites, such products may block access to health information sites, particularly those related to sexuality.

**Objective** To quantify the extent to which pornography-blocking software used in schools and libraries limits access to health information Web sites.

**Design and Setting** In a simulation of adolescent Internet searching, we compiled search results from 24 health information searches (n=3206) and 6 pornography searches (n=781). We then classified the content of each site as either health information (n=2467), pornography (n=516), or other (n=1004). We also compiled a list of top teen health information sites (n=586). We then tested 6 blocking products commonly used in schools and libraries and 1 blocking product used on home computers, each at 2 or 3 levels of blocking restrictiveness.

**Main Outcome Measure** Rates of health information and pornography blocking.

**Results** At the least restrictive blocking setting, configured to block only pornography, the products blocked a mean of only 1.4% of health information sites. The differences between blocking products was small (range, 0.6%-2.3%). However, about 10% of health sites found using some search terms related to sexuality (eg, *safe sex*, *condoms*) and homosexuality (eg, *gay*) were blocked. The mean pornography blocking rate was 87% (range, 84%-90%). At moderate settings, the mean blocking rate was 5% for health information and 90% for pornography. At the most restrictive settings, health information blocking increased substantially (24%), but pornography blocking was only slightly higher (91%).

**Conclusions** Blocking settings have a greater impact than choice of blocking product on frequency of health information blocking. At their least restrictive settings, over-blocking of general health information poses a relatively minor impediment. However, searches on some terms related to sexuality led to substantially more health information blocking. More restrictive blocking configurations blocked pornography only slightly more, but substantially increased blocking of health information sites.

JAMA. 2002;288:2887-2894

www.jama.com

computers used by minors in order to be eligible for some forms of federal funding. While the CIPA requirement

for libraries was struck down by a circuit court on the grounds that it violates the First Amendment,<sup>4</sup> it is cur-

**Author Affiliations:** Department of Family Medicine, University of Michigan Medical School, Veterans Affairs Health Services Research and Development Service (Dr Richardson), School of Information (Dr Resnick and Mr Hansen), and Health Media Research Laboratory (Ms Derry), University of Michigan, Ann Arbor; Henry J. Kaiser Family Foundation, Menlo Park, Calif (Ms Rideout).

**Corresponding Author and Reprints:** Caroline R. Richardson, MD, Department of Family Medicine, University of Michigan Medical School, 1018 Fuller St, Ann Arbor, MI 48109-0708 (e-mail: caroli@umich.edu).

**Medicine and Media Section Editor:** Annette Flanagan, RN, MA, Managing Senior Editor.

rently being appealed to the US Supreme Court. Meanwhile, 73% of schools<sup>5</sup> and 43% of public libraries<sup>6</sup> already use filters of some kind.

Filtering software intended to limit minors' exposure to pornography and other controversial material may inadvertently reduce the usefulness of the Internet as a health information tool for adolescents. Web sites that address issues of health and sexuality might be particularly susceptible to erroneous blocking. For example, cases of filters blocking access to breast cancer sites were widely publicized beginning in 1995, although this particular error has largely been corrected in recent years.<sup>7</sup> The use of filtering software in public schools and libraries is of special concern, because adolescents' health concerns often focus on issues related to sexuality, and because those who do not have computers at home rely on schools and libraries for Internet access.

Despite the concerns about the potential impact of blocking software on access to health information, and prolonged and impassioned public debate, surprisingly little empirical evidence exists regarding blocking errors. Recent government-commissioned studies in the United States, Europe, and Australia<sup>8-10</sup> used methodologies similar to ours but had smaller samples of health information sites. Furthermore, most filtering software systems allow administrators to specify blocking configurations, providing individual schools or libraries with the ability to tailor the blocking to local community standards. The effect of different configurations on the accuracy of the blocking systems has not been sufficiently tested.

We developed a computer model to simulate information-seeking by adolescents. Using this model, we tested the ability of 6 different blocking software packages commonly used in schools and libraries, as well as 1 product commonly used on home computers, each under a variety of blocking configurations, to discriminate between health information Web sites and pornography Web sites.

## METHODS

### Study Design

We simulated adolescent searching and browsing on the Internet to compile lists of Web sites that adolescents might come across while looking for either health information or pornography. For the search simulation, trained raters then classified each of the sites in these lists as health information, pornography, or other. Finally, we tested each site against 7 blocking products, each configured at 2 or 3 different levels of blocking restriction, to determine blocking rates for health information and pornography.

### Search Simulation

To simulate searches, we submitted search terms to the 6 Internet search engines that are among the most popular with teens according to data from a Kaiser Family Foundation survey<sup>2</sup>: Yahoo, Google, America Online (AOL), Microsoft Network, Ask Jeeves, and Alta Vista. To ensure that we had some variety in our list of sites with respect to likelihood of being blocked, we selected search terms from the following categories: (1) health topics unrelated to sex (eg, diabetes); (2) health topics involving sexual body parts, but not sex related (eg, breast cancer); (3) health topics related to sex (eg, pregnancy prevention); (4) controversial health topics (eg, abortion); and (5) pornography.

For each of the first 4 categories, we chose 6 frequently used search terms for health topics relevant to adolescents.<sup>2</sup> Frequency data for each search term was obtained from 2 different search engine logs of search term use, one from Overture.com<sup>11</sup> and the other from Excite.<sup>12</sup> For the fifth category, we also used the Overture and Excite data to select 6 frequently used search strings: *blowjob*, *free sex*, *teen porn*, *hard-core porn*, *porn*, and *XXX*.

On May 9, 2002, we ran a custom JAVA computer program to conduct searches for the 30 search strings on each of the 6 search engines and to store the results in a database. The search procedure programmed into this simu-

lation program was based on data from an observational pilot study during which we observed 12 teens conducting a total of 69 health information searches. Because none of adolescents in the observational study clicked on advertisements or sponsored links, and they looked past the fourth page of results less than 5% of the time, our JAVA program also ignored ads and sponsored links and captured only the first 40 search results from each search. The list of search results was collapsed into a smaller list of unique uniform resource locators (URLs), and sites that were not available for classifying or blocking tests because they were offline or broken, or for other technical reasons, were not included in the analysis. We also screened each Web site for automatic redirect coding, and for most of these sites we were able to follow the redirect link in our blocking tests. If either the original URL or the redirected destination was blocked, we considered the site to be blocked.

### Web Site Classification

Research associates coded the Web sites following a detailed coding scheme according to whether or not they contained health information and then by whether or not they were pornographic. The raters explored each site by reading pages and following links, seeking both health information and pornography. If no health information was found within 2 minutes, the site was classified as nonhealth; the same was done for pornography. Any information about topics that might be discussed in a medical school or school of public health counted as health information, even if the source or quality of the information was questionable. Loosely following the definitions of obscenity in US law,<sup>13</sup> any text or graphics depicting genitals or a sexual act and designed to appeal to a prurient interest, and not of an educational or scientific nature, were considered pornography. Sites that contained both health information and pornography (n=14/3987 rated sites) were classified as pornographic for all analyses.

Two primary raters were each assigned 60% of the sites, and ratings were done independently. Sites were assigned to raters using a systematic sampling from the complete list with a random component to ensure that raters could not know which sites would be rated by the other rater. The 10% overlap for each allowed us to calculate interrater reliabilities for both the health information rating ( $\kappa = .84$ ) and for the pornography rating ( $\kappa = .92$ ). Primary raters also had the option of not assigning a classification to a site for which they were unsure of the proper rating. These sites, and those given 2 different ratings by the 2 primary reviewers, were subsequently discussed with a third rater and a consensus rating of health, pornography, or other was assigned.

**Blocking Products**

We tested 7 different blocking products (TABLE 1), 6 of which were products commonly used in schools and libraries. All 6 of these products allow the network administrator to specify a custom blocking configuration by specifying topics or categories. The categories vary from vendor to vendor, though they tend to be roughly comparable. Some vendors provide one or more default configurations, but vendors have a wide range of customers, including corporations as well as schools and libraries, and most vendors were not willing to identify a “typical” school configuration. Calls to 20 school systems and libraries confirmed wide variability in their configurations and that none was using a vendor’s default setting. We defined 3 configurations for each product, to reflect extreme choices and a middle position. Our least-restrictive configuration, matching the configuration used in another recent test,<sup>8</sup> was designed to block only pornography. Our moderately restrictive configuration blocked pornography as well as a few other categories such as illicit drugs, nudity, and weapons. It was modeled on the configuration used by one major statewide school network that blocked fewer categories than some school districts but more than others. Our most restrictive

**Table 1.** Blocking Products

Product	Company	Market Share, %*
SmartFilter v3.0.1	Secure Computing	<5 (Education)
8e6 v4.5	8e6 Technologies	<5 (Education)
Websense v4.3.1	Websense	6 (Education) 6 (Library)
CyberPatrol (SuperScout v4.1.0.8)	SurfControl	10 (Education) 45 (Library)
Symantec Web Security v2.0	Symantec	6 (Education)
N2H2 v2.1.4	N2H2	40 (Education) <5 (Library)
AOL Parental Controls	America Online	Home

\*Data from Curry and Haycock.<sup>16</sup>

configuration for each product was set up to block all topics or categories that plausibly might be blocked in some school or library. For most products, all categories that the products offered were blocked except news, health, education, finance, search engine, and job search sites. The details of our product configurations are available on the study Web page (<http://www.kff.org>).

The seventh blocking product we tested was America Online Parental Controls (AOL PC). At the time of our study, this product, designed primarily for home use, allowed only 2 configuration options appropriate for teens. Parents could choose a moderately restrictive setting for mature teens or a very restrictive setting for young teens. We have chosen not to include AOL PC in the between-product comparisons. This is partly because AOL is not commonly used in schools and libraries and partly because the limited configuration options made it impossible to determine if AOL’s blocking was truly comparable to the configurations that we set for the other products in the study.

Most of the blocking tests were completed immediately after the searches, on the same day. Due to technical difficulties related to AOL’s proprietary browsing software, the AOL PC blocking test took several days to complete. Due to errors in the initial runs, CyberPatrol’s configurations and 2 of the configurations each for Symantec (least restrictive and moderately restrictive) and Websense (moderately restrictive and most restrictive) were rerun about 6 weeks later.

**Top Health Sites Recommended for Teens**

Our browsing simulation entailed compiling a list of recommended health information Web sites for teens (n=633). Two online directories (Yahoo and Google) were used to determine the most popular and widely recommended health sites for adolescents. Within these directories, there are several health categories (eg, *Kids and Teens > Health > Drugs and Alcohol*). We selected only those sites for which the category header mentioned teens or youth as well as health issues related to 1 of our 24 health search terms. These sites were assumed to be health information sites and were not independently rated. The sites were compiled from the directories in June 2002.

**Statistical Analysis**

As a measure of an individual blocking product’s tendency to block health information Web sites, we calculated the percentage of health information sites that were blocked by each of the blocking products at each blocking level. The denominator for all of these percentages was the number of unique health information sites in our list of search simulation results that were reachable at the time of the blocking test for each product and configuration. A similar analysis was done for the pornography sites in our search simulation results and for the recommended health sites list. We also calculated summary percentage results for all of the blocking products at a given configuration.

In order to identify statistically significant differences in product performance, we used a series of 6 multivariable logistic regressions to calculate odds ratios and 95% confidence intervals for tendency to block health information or pornography. Multivariate logistic regression was used because it allowed us to test statistical significance of product differences without having to do post-estimation adjustments for multiple comparisons. The regression model also allowed us to examine the effects of factors such as search term on likelihood of appropriately blocking pornography or inappropriately blocking health information while controlling for blocking product. Models were estimated independently at each blocking level and independently for health blocking and for pornography blocking. The dependent variable in all of these models was a dichotomous variable representing the results of a single blocking test by 1 product at 1 blocking level for 1 site, either blocked or not blocked. The independent variables were dummy dichotomous variables representing the 6 different school and library blocking products.

For each model, we chose the blocking product that performed best (either least likely to block health information or most likely to block pornography) as the reference group

when specifying the independent variables. This allowed us to interpret the regression results such that odds ratios significantly different from 1 indicate that the product performed significantly worse than the best product in the category. We used STATA v7.0 SE<sup>14</sup> for this analysis and used STATA's "svy" commands, allowing adjustment for clustering by site. These models were estimated using a pseudo-log likelihood method. Goodness of fit was tested using the Hosmer-Lemeshow method on all 6 models unadjusted for clustering, and all had excellent fit across the range of probabilities.

We selected 1 of the search terms that resulted in a large number of blocked health information sites (*safe sex*) for more detailed analysis of the content of health information sites (n=45) that were blocked by at least 1 product at either the least restrictive or moderately restrictive settings. A research associate visited each of the sites and summarized the content in 1 or 2 sentences, with specific attention to content that might have triggered the blocking software. We then analyzed the summaries to determine patterns.

**RESULTS**

**Search Simulation Results**

Our search simulation yielded a total of 6760 Web sites. After eliminating du-

plicate sites (n=2501) and sites that were unreachable or could not be included for technical reasons (n=272), 3987 unique URLs remained. Of these unique sites, 2467 contained health information and not pornography, 516 contained pornography, and 1004 were rated as neither health information nor pornography.

Results of the blocking tests on the health information sites are shown in the first section of TABLE 2. Large differences are apparent with the 6 comparable products compared as a group across the 3 levels of blocking. At the least restrictive blocking configuration, the mean blocking rate of health sites was 1.4% (range for the 6 products, 0.6%-2.3%). The mean blocking rate of pornography sites was 87.2% (range, 84%-90%). As the level of blocking increased from least to moderate to most restrictive, the frequency of health blocking increased substantially while the improvement in pornography blocking was small. At moderate blocking settings, the mean blocking rate of health information sites was 5.2%; at the most restrictive settings, it was 24%. At the least restrictive configuration, 5% of all health information sites were blocked by at least 1 product. This compares with 16% of sites for moderate blocking settings and 63% of sites for the most restrictive settings.

**Table 2.** Blocking Results for All Products Across 3 Levels of Blocking Restrictiveness\*

Restrictiveness	No. (%)							Mean Blocking Rate, %†	AOL PC
	SmartFilter	8e6	Websense	CyberPatrol	Symantec	N2H2			
<b>Health Information URLs Blocked (n = 2467)</b>									
Least	56 (2.3)	27 (1.1)	15 (0.6)	39 (1.6)	48 (1.9)	20 (0.8)	1.4	NA	
Moderate	143 (5.8)	112 (4.5)	94 (3.8)	68 (2.8)	188 (7.6)	160 (6.5)	5.2	79 (3.2)	
Most	447 (18.2)	371 (15.1)	873 (35.4)	552 (22.4)	826 (33.5)	481 (19.5)	24.0	398 (16.1)	
<b>Pornography URLs Blocked (n = 516)</b>									
Least	450 (87.2)	460 (89.1)	433 (83.9)	442 (85.7)	453 (87.8)	462 (89.5)	87.2	NA	
Moderate	457 (88.7)	469 (90.9)	471 (91.3)	442 (85.7)	461 (89.3)	479 (92.8)	89.8	475 (92.1)	
Most	459 (89.0)	475 (92.1)	484 (93.8)	450 (87.2)	467 (90.5)	485 (94.0)	91.1	489 (94.8)	
<b>Recommended Health Information URLs Blocked (n = 586)</b>									
Least	0 (0)	3 (0.5)	3 (0.5)	3 (0.5)	8 (1.4)	2 (0.3)	0.5	NA	
Moderate	6 (1.0)	7 (1.2)	8 (1.4)	5 (0.9)	49 (8.4)	22 (3.8)	2.8	7 (1.2)	
Most	98 (16.8)	64 (10.9)	230 (39.4)	155 (26.5)	167 (28.5)	136 (23.2)	24.2	90 (15.0)	

\*AOL PC indicates America Online Parental Controls; URL, uniform resource locator; and NA, not applicable.  
 †For all products except AOL PC.

There were some statistically significant differences between products, as summarized in TABLE 3. In the least restrictive blocking configuration, Websense was the least likely to block health information, so Websense became the reference category (odds ratio, 1). SmartFilter, 8e6, CyberPatrol, and Symantec were all more likely to block health information than Websense, but N2H2 was not significantly more likely to block health information than was Websense. Within the margin of error for our study, Websense and N2H2 are both top products at not blocking health information at the least restrictive blocking configuration. Across all 3 blocking levels, N2H2 was the best at blocking pornography.

Overall, for the 24 health search strings only about 1% of search results were pornography, but the software blocked fewer of these pornography sites (62%) than those resulting from pornography searches (89%). Adding a dummy variable for a pornography vs not pornography search in the logistic regression reported in Table 3 confirmed that this difference is statistically significant ( $P < .001$ ).

When comparing health information blocking rates across the 24 different health search terms, there were some notable differences in performance as summarized in TABLE 4. At the least restrictive setting, where products were supposed to block pornography only, about 10% of nonpornographic health information sites

returned from searches using the terms *safe sex*, *condom*, and *gay* were blocked, while for most other searches less than 1% of health sites were blocked. At the moderately restrictive setting, these search terms again yielded a larger percentage of health results blocked, as did *ecstasy*, presumably because the moderately restrictive setting was supposed to block access to sites about illegal drugs. At the most restrictive blocking setting, most strings yielded a health information blocking rate of at least 10%, and half of the more controversial topics had rates above 40%.

When we tested the blocking products against a list of 633 top health information sites, we found similar results. After excluding 29 sites that were unreachable and eliminating duplicates, we ran our blocking test on 586 unique recommended health sites. At the least restrictive blocking setting, 0.5% (range, 0%–1.4%) of recommended teen health information sites were blocked. This compares with 2.5% (range, 0.9%–8.4%) at the moderately restrictive blocking settings and 23% (range, 10.9%–39%) at the most restrictive blocking settings.

**What Kinds of Health Sites Were Blocked?**

Of the 86 unique health sites resulting from searches using the term *safe sex*, 28 were blocked by some product at the least restrictive configuration and 42 were blocked by some product at the moderately restrictive configuration. Of

those blocked at the least restrictive configuration, the vast majority contained at least moderately specific descriptions of condom use and/or alternatives to intercourse. Four of these sites contained pictures and graphic depictions of sexual acts and 2 contained nudity that seemed to be artistic in nature. Three required users to confirm that they were older than 18 years before visiting the site. Four sites sold condoms. The additional health sites blocked at the moderately restrictive configuration did not appear qualitatively different than those blocked at the least restrictive level, ie, they did not contain more offers for condoms or more explicit information on safer sexual practices.

**COMMENT**

For all 7 of the filtering products we tested, access was blocked to only a small percentage of health information Web sites when the blocking configurations were set to the least restrictive settings. With only 1.4% of health information sites that we tested blocked, a teenager whose access to a particular health information site is inadvertently blocked will probably be able to easily find an unblocked site with similar information. This suggests that filtering software set to block pornography will not necessarily have a serious impact on access to general health information. Compared with other factors that may limit teenagers' access to health information when searching the

**Table 3.** Tendency to Block Health Information Sites vs Pornography Sites\*

Restrictiveness	OR (95% CI)					
	SmartFilter	8e6	Websense	CyberPatrol	Symantec	N2H2
<b>Health Information</b>						
Least	3.8 (2.3-6.2)	1.8 (1.1-3.0)	Reference†	2.6 (1.6-4.3)	3.2 (1.9-5.4)	1.3 (0.8-2.4)
Moderate	2.2 (1.7-2.7)	1.7 (1.3-2.1)	1.4 (1.1-1.8)	Reference†	2.9 (2.3-3.8)	2.4 (2.0-3.1)
Most	1.3 (1.1-1.4)	Reference†	3.1 (2.7-3.5)	1.6 (1.4-1.8)	2.8 (2.5-3.2)	1.4 (1.2-1.6)
<b>Pornography</b>						
Least	0.80 (0.58-1.1)	0.96 (0.70-1.3)	0.61 (0.45-0.82)	0.70 (0.51-0.96)	0.84 (0.61-1.2)	Reference†
Moderate	0.61 (0.42-0.88)	0.77 (0.51-1.2)	0.81 (0.54-1.2)	0.46 (0.32-0.66)	0.65 (0.44-0.95)	Reference†
Most	0.51 (0.35-0.77)	0.76 (0.47-1.2)	0.97 (0.61-1.5)	0.44 (0.29-0.65)	0.61 (0.40-0.93)	Reference†

\*OR indicates odds ratio; CI, confidence interval.

†The reference product for each group was chosen to be the best product. For pornography blocking, the best product is the one most likely to block pornography; for health information blocking, the best product is the one least likely to block health information.

Internet, including spelling errors, limited search skills, and uneven quality of search engines, overblocking by filtering software set at the least restrictive blocking settings poses a relatively minor barrier for most of the health topics we studied.

However, the blocking rates were noticeably higher for some topics. For example, with searches on *safe sex*, almost 10% of attempts to access health results were blocked, and 33% of health sites were blocked by at least 1 of the products, even on the least restrictive setting. More than 20% of attempts were blocked at the moderate setting. These

blocking rates may be enough to make blocking software a serious impediment to searching for this type of health information. This is particularly concerning given that 80% of teens identify sexual health as very important.<sup>2</sup> The conventional wisdom that the presence of words mentioning sexual body parts fools blocking software appears not to be true (no *breast cancer* search results were blocked at the least restrictive configuration). There do seem to be patterns, however, in the types of blocking errors. To the extent that these blocked health information sites represent errors and not intentional block-

ing of controversial sites, further research and product development should be devoted to improving the ability of products to discriminate between pornography and health information in sites related to safe sex, condoms, and homosexuality.

We also found that configuration of the products can have a large impact on access to health information. The moderately restrictive configurations that we believe approximate many schools' settings led to more than 3 times as much blocking of health information as the least restrictive, pornography-only blocking settings. Overall, the most re-

**Table 4.** Blocking of Health Information by All Products During Health Searches

Search String	Sites Returned From Searches, No. (%)*				Health Information Sites Blocked, %†		
	Health	Pornography	Other	All	Least Restrictive	Moderately Restrictive	Most Restrictive
<b>Health Topics Unrelated to Sex</b>							
Diabetic diet	202 (88)	0 (0)	28 (12)	230	0.1	0.2	13.9
Diabetes	196 (85)	0 (0)	34 (15)	230	0.1	0.4	10.0
Ecstasy	171 (72)	2 (1)	64 (27)	237	0.3	24.9	36.2
Alcohol	204 (88)	0 (0)	27 (12)	231	0	7.1	12.7
Suicide	200 (85)	2 (1)	32 (14)	234	0.2	1.7	13.7
Depression	189 (80)	0 (0)	46 (20)	235	0	1.0	11.2
<b>Health Topics Involving Body Parts, Not Sex-Related</b>							
Breast cancer	224 (97)	0 (0)	8 (3)	232	0	0.2	6.9
Cancer	221 (95)	0 (0)	12 (5)	233	0	0.3	3.7
Jock itch	180 (94)	0 (0)	11 (6)	191	0.6	1.4	15.4
Yeast infection	206 (94)	1 (0)	11 (5)	218	0.1	1.1	18.4
Breast feeding	217 (92)	0 (0)	18 (8)	235	0.2	1.2	18.6
Breast pump	163 (70)	0 (0)	69 (30)	232	0	0.9	26.3
<b>Health Topics Related to Sex</b>							
Sexually transmitted disease	112 (49)	0 (0)	118 (51)	230	1.4	3.4	23.4
Herpes	203 (88)	0 (0)	29 (12)	232	1.0	1.8	23.0
Safe sex	131 (68)	11 (6)	51 (26)	193	9.3	20.5	50.0
Condom	152 (65)	9 (4)	72 (31)	233	9.1	27.7	55.4
Pregnancy	219 (94)	0 (0)	15 (6)	234	0.4	0.6	31.6
Birth control	210 (89)	0 (0)	27 (11)	237	1.8	5.0	34.7
<b>Controversial Health Topics</b>							
RU486	177 (76)	1 (0)	54 (23)	232	0.3	2.1	25.8
Abortion	184 (79)	0 (0)	48 (21)	232	0.2	3.2	44.6
Gay	74 (38)	10 (5)	113 (57)	197	11.1	24.6	59.9
Lesbian	79 (34)	7 (3)	149 (63)	235	3.8	17.1	59.0
Rape	151 (67)	2 (1)	74 (33)	227	1.2	3.3	22.0
Date rape	152 (83)	1 (1)	30 (16)	183	1.7	5.2	21.1
Total	4217 (78)	46 (1)	1140 (21)	5403	1.3	5.1	24.1

\*Totals reflect all search results; if a site is suggested by more than 1 search engine, it may be counted multiple times. A maximum of 240 results are possible for each search, but some search engines did not return results for some searches, either because of transitory errors or because they refuse to return results for some search strings, and some sites returned by searches were broken or unreachable.

†Blocking percentages reflect the percentage of unique sites returned by all searches for that search term that were blocked. If a site is suggested by more than 1 search engine, it is counted only once. Percentages are averaged across the 6 products for the given setting (least, moderate, most).

strictive configurations blocked more than 17 times as often as the least restrictive configurations. However, these more restrictive settings led to only slight improvements in blocking of pornography: their main effect was to block other potentially controversial types of information, including some types of health information.

There may be principled reasons why some schools or libraries choose to block more than pornography, including some kinds of health information. These decisions, however, should be viewed as important policy decisions and not mere technical configuration issues to be left to network administrators. The choice of configurations should get at least as much public and managerial scrutiny as the initial decision about whether to install filters at all.

Comparing among the products, the blocking rates for health information varied by a factor of 2 or more. At the least restrictive settings, for most health searches the overall blocking rates were small enough that erroneous blocking was rare for all the products. For more restrictive settings, and for searches on topics such as *safe sex*, differences among products would become more noticeable.

The products each blocked 80% to 90% of the pornography sites at minimal blocking levels. Health searches generated links to pornography sites only about 1% of the time, so that accidentally stumbling across pornography while searching for health information is a rare occurrence, and even rarer with blocking software. However, it is interesting to note that the filters were far more effective at blocking pornography sites resulting from pornography searches than at blocking pornography sites resulting from health searches. One possible explanation is that the same characteristics, such as particular text appearing in the content or links to and from other sites, that caused pornography sites to appear in search engine results also caused the blocking software to classify them incorrectly. We do not know exactly what text content or link patterns might

be the source of the errors, or whether the sites were deliberately designed to induce such errors.

Some simple industry-wide actions might reduce error rates even further and aid in product selection and configuration. For example, it would be helpful if creators of health or pornography sites could provide hints to the vendors about how the site should be classified. One solution might be the more widespread use of embedded labels<sup>15</sup> or the creation and use of domain names such as .health and .xxx.<sup>7</sup> Conversely, it would be helpful if vendors informed operators of Web sites about whether their sites were blocked, so that errors could be identified and corrected more quickly. This could be accomplished through an electronic clearinghouse, operated by a non-profit organization or government agency, where people could submit a URL and find out immediately whether the site was blocked on any of the configurations of the major vendors.

Vendors may have commercial reasons for not fully disclosing their blocking strategies. However, providing the ability to check if a specific URL is blocked would not require vendors to divulge the trade secrets of their classification methods or publish their entire blocking lists. Some vendors voluntarily provide sites allowing users to check for blocking of specific URLs. Legislation or regulation could mandate vendor participation or provide incentives such as certifying vendors for government contracts if they allow these blocking checks. Moreover, if a publisher does find that its site is blocked and feels that it is a mistake, the software vendor may not be responsive to an inquiry asking for a reevaluation. One possible solution would be to establish an appeal process that the vendor would have to respond to within a fixed period of time. Finally, to aid in product and configuration selection, tests of the form reported in this article should be conducted on a regular basis, using a different set of search topics each time.

While the rigorous sampling methods and the large sample size lend weight

to the results, there are several limitations to our study. First, while we simulated searches on topics that previous surveys indicate interest teenagers, our simulations were still fairly basic. We did not attempt to model how teenagers react to the short summary text for each site that a search engine returns, and how that influences their choices of which links to follow. Similarly, we did not attempt to model how having some sites blocked would affect the progress of a search. Second, we made no attempt to rate the quality of health information or the relevance of health sites to the search topics. Third, when we counted blocked health information sites, we made no attempt to check whether alternative sources of the same health information were available and not blocked. Thus, this study measures the percentages of health and pornography items that are blocked, but was not designed to give a detailed picture of how the presence of blocking software would affect the quality of health information a teenager would find when searching. Fourth, some of the product configurations were tested at a later date due to technical difficulties. Since the search results from an earlier date were used, it is possible that the product vendors had revised their blocking decisions for those URLs, perhaps reducing the number of blocked health sites or increasing the number of blocked pornography sites. However, results for product configurations tested later were roughly consistent with the overall pattern of results, both for individual products and across products.

Another important limitation of the study is that it focused only on the categories of pornography and health information. Some individuals may think that teenagers should be prevented from accessing information on controversial topics such as condoms, homosexuality, and abortion. Our analysis treated sites discussing these topics as health information sites. Depending on one's opinion about accessibility of information on these controversial topics, the more restrictive blocking rates for health information found in some

of the software configurations may or may not be problematic. While it was fairly easy to achieve interrater reliability in classifying pornography and health information, it is less clear what the objective criteria for more controversial topics would be, and we deferred that to future research. For those who are interested in rerating our sample, running their own statistics, or simply examining our ratings, the database is available on the study Web page (<http://www.kff.org>).

The differences between products were much smaller than the differences between settings within each product. For general health information searches, at their least restrictive settings, overblocking by filtering soft-

ware poses a relatively minor risk. However, for searches for some sexually related health information and for homosexuality, the blocking of health information sites was around 10% even on the least restrictive setting, suggesting that blocking software is less effective at distinguishing pornography sites from those discussing these health topics. Moreover, more restrictive blocking configurations substantially increased health information blocking with only slight improvement in pornography blocking: the main effect of the more restrictive settings is to block other categories of controversial material besides pornography.

**Author Contributions:** *Study concept and design:* Richardson, Resnick, Rideout.

*Acquisition of data:* Richardson, Hansen, Derry.  
*Analysis and interpretation of data:* Richardson, Resnick, Hansen, Rideout.

*Drafting of the manuscript:* Richardson, Resnick, Hansen.

*Critical revision of the manuscript for important intellectual content:* Richardson, Resnick, Hansen, Derry, Rideout.

*Statistical expertise:* Richardson, Resnick.

*Obtained funding:* Richardson, Resnick, Rideout.

*Administrative, technical, or material support:* Hansen, Derry.

*Study supervision:* Richardson, Resnick, Derry, Rideout.

**Funding/Support:** This study was supported by a contract from the Kaiser Family Foundation. Ms Rideout is employed by the Kaiser Family Foundation and had substantial input into the study design, analysis, and presentation of results.

**Acknowledgment:** For project conceptualization and management we thank Victor Strecher, PhD, MPH, and Ed Saunders, MS. For programming and data management we thank Mike Nowak, MS, and Ian Jones, BA. For statistical consultation we thank Rod Little, PhD, MSc, and Ying Yuan, MS. Joel Howell, MD, PhD, Rod Hayward, MD, and Mack Ruffin, MD, MPH, provided helpful comments on manuscript drafts.

REFERENCES

1. Fox S, Rainie L. *The Online Health Care Revolution: How the Web Helps Americans Take Better Care of Themselves*. Washington, DC: Pew Internet and American Life Project; 2000:23.
2. Rideout V. *Generation Rx.com: How Young People Use the Internet for Health Information*. Menlo Park, Calif: Henry J. Kaiser Family Foundation; 2001:37.
3. *Generation Rx.com Survey*. Menlo Park, Calif: Henry J. Kaiser Family Foundation; 2001.
4. *American Library Association Inc v United States*, 9537 US Dist Lexis (2002 ED PA).
5. Cattagni A, Westat EF. *Internet Access in US Public Schools and Classrooms: 1994-2000*. Washington, DC: National Center for Education Statistics; 2001:20.
6. Oder N. The new wariness. *Libr J*. 2002;127:55-57.

7. Thornburgh D, Lin H, eds. *Youth, Pornography, and the Internet*. Washington, DC: National Academy Press; 2002.
8. US Department of Justice. *Web Content Filtering Software Comparison*. Morrisville, NC: eTesting Labs; 2001:18.
9. Greenfield P, Rickwood P, Tran HC. *Effectiveness of Internet Filtering Software Products*. Sydney, Australia: CSIRO Mathematical and Information Sciences; 2001:90.
10. Brunessaux S, Isidoro O, Kahl S, Ferlias G, Soares ALR, for MATRA Systemes & Information, Matra Global Netservices, Sail Labs, Hypertech, Red Educativa. *Report on Currently Available COTS Filtering Tools*. 2001. Available at: [http://www.saferinternet.org/downloads/netprotect\\_review\\_tools.pdf](http://www.saferinternet.org/downloads/netprotect_review_tools.pdf).

Accessibility verified October 30, 2002.

11. Tools page. Overture Web site. *Search Term Selection Tool*. Available at: <http://www.overture.com/d/USm/adcenter/tools/index.jhtml>;2002. Accessibility verified October 30, 2002.
12. Jansen B, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing Manage*. 2000; 36:207-227.
13. *Miller v California*, 413 US 15 (1973).
14. StataCorp. *Stata Statistical Software: Release 7.0 SE*. College Station, Tex: Stata Corp; 2001.
15. Resnick P. Filtering information on the Internet. *Sci Am*. 1997;273:106-108.
16. Curry A, Haycock K. Filtered or unfiltered? *Sch Libr J*. 2001;1:42-47.