THE HENRY J. KAISER FAMILY FOUNDATION

See No Evil: How Internet Filters Affect the Search for Online Health Information

Appendices

sex education
safer sex
sexual heath

A Kaiser Family Foundation Study
December 2002

By Paul Resnick, Ph.D.
Caroline Richardson, M.D., and
Derek Hansen
University of Michigan

Table of Contents

Appendix A: Categories, Topics, and Search Strings

In order to insure 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 (e.g., diabetes)
2. Health topics involving sexual body parts, not sex related (e.g., breast cancer)
3. Health topics related to sex (e.g., pregnancy prevention)
4. Controversial health topics (e.g., abortion)
5. Pornography

For each of the first four categories, we chose three topics. Popular health topics relevant to adolescents were chosen based largely upon the results of the Kaiser Family Foundation survey *Generation Rx.com* (Rideout, 2001). Two search strings were then chosen for each topic making a total of 24 health related search strings. For the pornography category we selected 6 pornography related search strings. Frequency data for each search string was obtained from two different search engine logs of search string use, one from Overture.com and the other from Excite (Jansen, Spink, & Saracevic, 2000). When selecting search terms we attempted to find popular terms that described a given topic. Table A-1 includes the final search terms, topics, and categories along with the number of searches performed at Overture and Excite.

Table A-1

Category	Topic	Search String	Overture[1]	Excite[2]	Search String	Overture[1]	Excite[2]
Non-Sex related searches	Diabetes	diabetic diet	36,385	3	diabetes	295,524	88
	Addiction	ecstasy	105,069	22	alcohol	120,312	31
	Depression	suicide	109,452	43	depression	212,018	54
Sexual Body Part, not sex related health problem	Breast Cancer	breast cancer	136,215	39	cancer	208,514	58
	Jock Itch	jock itch	10,173	6	yeast infection	85,027	15
	Breast Feeding	breast feeding	59,904	9	breast pump	19,694	0
Sexual Body Part, sex related	STD	STD	107,433	30	herpes	214,759	62
	Safe Sex	safe sex	11,928	5	condoms	79,376	34
	Pregnancy Prevention	pregnancy	651,765	274	birth control	79,838	18
Controversial Issues	Abortion	RU486	35,827	2	abortion	302,627	93
	Homosexuality	gay	379,590	217	lesbian	345,951	97
	Sexual Assault	rape	N/A	70	date rape	N/A	8
Pornography	Porn 1	blowjob	564,048	0	free sex	657,180	374
	Porn 2	teen porn	75,387	22	hardcore porn	1,048,985	40
	Porn 3	porn	3,819,082	1,005	XXX	674,192	599

[1]Frequency of search terms taken from Overture "Search Term Selection Tool." Represents number of searches for the entire month of March 2002 for the Overture (formerly GoTo) search engine. This database is an edited database that does not allow for certain search terms (e.g., rape) and in some cases combines word variations (e.g., condoms and condom). Details of the editorial process can be found at the site.

[2]Frequency of search terms taken from September 1, 1997.

Appendix B: Search Engines

Selection criteria
Search engines were chosen based upon popularity and certain unique characteristics of search engines (See Table 7).

Popularity
The primary source for the most widely used search engine for adolescents was the Kaiser Family Foundation's *Generation Rx.com* report. We focused on the 15-19 age group in particular. Other sources that measure the most popular search engines (i.e., Jupiter Media Metrix, Nielson NetRatings) were also consulted, although they were not weighted as heavily since they are not specific to adolescents or health searches.

Table B-1: Search engines selected

Search Engine	Generation Rx.com (age 15-19)	Jupiter Media Metrix (March 2002)[1]	Nielson NetRatings (Dec. 2001)[2]
Yahoo!	44%	34%	60%
Google	9%	29%	20%
AOL	7%	22%	50%
MSN	6%	37%	50%
Ask (Ask Jeeves)	4%	16%	4%
Alta Vista	5%	Not Available	6%

[1] See http://searchenginewatch.com/reports/mediametrix.html for explanations and caveats. Data reflects % of unique users that performed a "search specific" activity.
[2] See http://searchenginewatch.com/reports/netratings.html for explanations and caveats. Data reflects % of unique users that visited a particular web site.

Appendix C: Recommended Teen Health Sites

Selection Criteria

Two online directories (Yahoo! and Google) were used in order to determine the most popular and widely recommended health sites for adolescents. These were chosen because of their high use among adolescents, as found in our observational study of teen health searches. Comparable directories were not available for the other search engines used in our study. Within these directories there are several categories that are relevant to this report. We chose only those that mention teens (or youth) and health issues related to our topics in the category header. All sites that were listed under these category headers were included except Yahoo! recommendations that linked to other Yahoo! categories.

Results

In all, this methodology resulted in 633 sites pulled from the categories listed in Table C-1.

Table C-1: Recommended Teen Health Site Categories

Yahoo! Category Header:

- Home > Health > Reproductive Health > Pregnancy and Birth > Teenagers
- Home > Health > Teen Health
- Home > Health > Teen Health > Teen Substance Abuse
- Home > Health > Teen Health > Teen Substance Abuse > Organizations
- Home > Society and Culture > Cultures and Groups > Teenagers
- Home > Society and Culture > Cultures and Groups > Teenagers > Girls > Puberty
- Home > Society and Culture > Cultures and Groups > Teenagers > Organizations
- Home > Society and Culture > Cultures and Groups > Teenagers > Puberty
- Home > Society and Culture > Death and Dying > Suicide > Youth
- Home > Society and Culture > Death and Dying > Suicide > Youth > Lesbian, Gay, and Bisexual
- Home > Society and Culture > Death and Dying > Suicide > Youth > Organizations
- Home > Society and Culture > Sexuality > Teen Sexuality

Google Category Header:

- Kids and Teens > Health
- Kids and Teens > Health > Conditions and Diseases
- Kids and Teens > Health > Conditions and Diseases > AIDS
- Kids and Teens > Health > Conditions and Diseases > Cancer
- Kids and Teens > Health > Conditions and Diseases > Diabetes
- Kids and Teens > Health > Drugs and Alcohol
- Kids and Teens > Health > Drugs and Alcohol > DARE
- Kids and Teens > People and Society > Psychology
- Kids and Teens > Teen Life > Suicide
- Kids and Teens > Teen Life > Teen Health
- Kids and Teens > Teen Life > Teen Health > Diseases and Disorders
- Kids and Teens > Teen Life > Teen Health > Drug Awareness
- Kids and Teens > Teen Life > Teen Health > Fitness and Nutrition
- Kids and Teens > Teen Life > Teen Health > Pregnancy
- Kids and Teens > Teen Life > Teen Sexuality
- Kids and Teens > Teen Life > Teen Sexuality > Abstinence
- Kids and Teens > Teen Life > Teen Sexuality > Contraception
- Kids and Teens > Teen Life > Teen Sexuality > Gay Lesbian and Bisexual
- Kids and Teens > Teen Life > Teen Sexuality > Gay Lesbian and Bisexual > Resources
- Kids and Teens > Teen Life > Teen Sexuality > Rape and Abuse
- Kids and Teens > Teen Life > Teen Sexuality > Sexually Transmitted Diseases

Appendix D: Classification of Sites

The classification system was designed to meet several goals. Because a large number of URLs would be rated, it was necessary to develop a rating system that did not require much time to rate each URL. In addition, the ratings needed to be as consistent as possible between multiple raters and based upon a written document so that future raters would come up with similar results. An attempt was also made to be as objective as possible, meaning that Americans with different political and religious beliefs would still be able to agree on which sites met the given criteria.

Classification scheme
Classification was done on two separate dimensions, whether health information was present and whether porn information was present.

Dimension #1: Health vs Not Health
Health information is health related information on any topic that would be discussed in a medical school or school of public health, including:

- lists of health providers or clinics
- alternative medicine,
- behavioral treatments, and
- lay people stories.

The following additional guidelines were provided to raters:

- We are not judging the quality or source of the information. Information that is wrong or that does not fit in the mainstream medical paradigm could still be health information.
- Also included would be advertisements for health products as long as they contain some health information, journal articles and newspaper articles on health related topics.
- Health information does not necessarily have to be related to the search topic or term.
- Legal discussions of health issues, on their own, do not count as health information. If there's health information (including statistics about disease prevalence), however, they would count.
- Healthy recipes don't count, unless they say why they're healthy.
- Health information about animals does not count.

Non-Health – Anything that does not contain health information as defined above.

Dimension #2: Porn vs Not Porn
Our definition of pornography was based loosely on the definition of obscenity under U.S. law. Things were classified as pornographic if they depict or describe an actual or simulated sexual act or sexual contact, or exhibit genitals in a way that seems intended to appeal to the prurient interest, and that satisfied the following criteria:

- **Not Art:** the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.
- **Detailed and explicit:** Vague innuendo does not count. Hints, airbrushed images that suggest but do not show genitals also do not count.

Non-porn – Anything that does not contain porn as defined above.

Rating Process
The process of rating each URL according to the classification scheme included several steps. First, each URL was assigned an identifier. The URLs were then randomized using the JAVA collections shuffle command. The odd sites went to the first rater and the even ones to the second rater. In addition, one in every 10 sites was rated by both raters in order to measure the level of disagreement between the two raters. Raters did not know which sites were reviewed by the other rater. During the rating of the first 300 URLs, the raters discussed problem cases with each other in order to clarify the rating criteria. After that, they rated independently.

A rating program was developed that allowed each rater to input their ratings. The program displayed, for each URL, a copy of the web page that had been downloaded using that URL, on the day that the tests were run. When the previously downloaded pages (e.g., the first page and all pages linked to off of the first page) were not enough to determine a classification, the rater went to the then-current site to make a determination. All of the rating was completed within two weeks of the day the sites were tested in order to minimize the amount of change.

In following links from the initial URL, the raters were permitted to go up to three levels deep (i.e., the first page is the first layer, a page linked to from the first page is a second layer, and a page linked to from the second layer is the third layer). In looking for health information, the raters could follow any links; a URL was considered a source for health information even if it just led easily to health information actually hosted on another site. If a blocking software program blocks access to one of these URLs, such a block would make it more difficult to find health information. In looking for porn, the raters followed only links within the same web site (i.e., same domain name). Sites were not considered pornographic simply for including a link to pornographic content on a different web site unless the link itself contained pornography

(e.g., an advertisement with pornographic photographs). It was assumed that the original site should not be blocked in such cases, but rather the software should block the destination site containing the actual pornography. However, a large list of links to sites containing pornography was considered pornographic content in and of itself.

Raters spent up to two minutes exploring each site. If no health information was apparent within two minutes of exploration, then the site was classified as non-health. If no porn had become apparent within two minutes, then the site was classified as non-porn. After the two-minute allotted time was up, the program would display a dialogue box that informed the rater that they must make a selection immediately. The program interface is displayed in Figure D-1.

A group consisting of the two original raters and a third rater met to review the sites that were originally marked as "Unable to Classify: need to discuss." There were 172 of these sites marked by at least one of the two original raters. Three raters also reviewed the sites that were rated by both original raters but on which there was disagreement. There were 29 of these sites, mostly resulting from one rater finding health information in an obscure place on a site while the other rater did not find that information within the two-minute time limit. In total, a group of three raters classified 201 sites together, or 4.7% of the total URLs that were classified.

Inter-rater reliability was quite good overall. In classification of porn vs. not, the raters agreed on more than 98% of the sites they rated in common, leading to a kappa score of .92. In classification of health vs. not, the raters agreed on more than 92% of the sites they rated in common, leading to a kappa score of .85.

Figure D-1

Appendix E: Blocking Software & Configurations

Blocking Software

Products were selected based upon popularity within the school and library markets. AOL Parental Controls was also included, to represent the home market. The final product list is shown in Table E-1.

Three sources were utilized to select the most popular filters. First, a January 2001 School Library Journal article includes the most specific data and has been used in congressional testimony more recently (Curry & Haycock, 2001). A discussion with a public relations staff member at N2H2 yielded the same list of products. In addition, we have included all of the major products that have been included in a recent study by the Department of Justice (US Department of Justice, 2001). Furthermore, several third party products including FamilyClick, American Family Filter, and Family Safe Viewing are based on these products.

While all of these products rely on at least some form of black-list, they rely on different lists. In addition, Symantec has a textual analysis tool called Dynamic Document Review (DDR) that can be enabled at varying levels. None of the products default to a corresponding filtered search engine, although some of them block access to search engine results pages when "inappropriate" words are entered.

Blocking Configurations

The tables below show the categories blocked in each of the three configurations we tested for each product. Where applicable, we also provide the manufacturer's default configuration or configurations, for comparison purposes. These configurations were not tested.

Table E-1

Product (version)	Company	Market Share (Curry & Haycock, 2001)
N2H2 (2.1.4)	N2H2	Education (40%), Library (< 5%)
Cyber Patrol (1.2.0.6)	SurfControl	Library (45%), Education (10%)
Symantec Web Security	Symantec	Education (6%)
Smartfilter (3.0.1)	Secure Computing	Education (< 5%)
8e6 (4.5)	8e6 Technologies	Education (< 5%)
Websense (4.3.1)	Websense	Education (6%), Library (6%)
AOL Parent Controls	America Online	Home

*Some products (e.g., Websense) also heavily target the commercial market.

N2H2	Tested Configurations						Vendor's default configurations					
Categories	**Most**		**Intermediate[1]**		**Least[2]**		**Maximum**		**Standard**		**Minimum**	
adults only	X		X				X		X		X	
alcohol	X		X				X					
auction	X						X					
chat	X						X					
drugs	X		X				X		X			
electronic commerce	X						X					
employment search	X						X					
free mail	X						X					
free pages	X						X					
gambling	X		X				X		X			
games	X						X					
hate/discrimination	X		X				X		X		X	
illegal	X		X				X		X		X	
jokes	X						X					
lingerie	X						X					
message/bulletin boards	X						X					
murder/suicide	X		X				X					
news							X					
nudity	X		X				X		X			
personal information	X						X					
personals	X						X					
pornography	X		X		X		X		X		X	
profanity	X		X				X					
recreation/entertainment	X						X					
school cheating info	X		X				X					
search engines							X					
search terms	X						X					
sex	X		X				X		X		X	
sports	X						X					
stocks							X					
swimsuits	X						X					
tasteless/gross	X		X				X					
tobacco	X		X				X					
violence	X		X				X		X		X	
weapons	X		X				X					
Exceptions												
education			X		X				X		X	
for kids			X						X		X	
history			X		X				X		X	
medical			X		X				X		X	
moderated											X	
text/spoken only					X						X	

[1]Used by Utah Education Network

[2]Categories match those used in the DOJ report

Cyber Patrol	Tested Configurations							
Categories	**Most**		**Intermediate[1]**		**Least[2]**		**Vendor's Default**	
adult/sexually explicit	X		X		X		X	
advertisements	X							
arts/entertainment	X							
chat	X							
computing/internet								
criminal skills	X		X					
drugs/alcohol/tobacco	X		X				X	
finance/investment								
food/drink	X							
gambling	X		X					
games	X							
glamour/intimate apparel	X							
government/politics								
hacking	X		X					
hate speech	X		X					
hobbies/recreation	X							
hosting sites	X							
job search/career development	X							
lifestyle/culture	X							
motor vehicles	X							
news								
personals/dating	X							
photo searches	X							
real estate	X							
religion	X							
anonymous proxies	X						X	
search engines								
shopping	X							
sports	X							
streaming media	X							
travel	X							
usenet news	X							
violence	X		X				X	
weapons	X		X					
web-based email	X							
sex education	X							
Exceptions								
education			X		X			
reference			X		X		X	
kid's sites			X		X		X	
health/medicine			X		X			
sex education			X		X		X	

[1]Matches N2H2 Intermediate as closely as possible

[2]Categories match those used in the DOJ report

Symantec	Tested Configurations					
Categories	**Most**		**Intermediate[1]**		**Least[2]**	
alcohol/tobacco	X		X			
anonymous proxies	X					
humor	X					
prescription medicine						
real estate	X					
religion	X					
travel	X					
crime	X		X			
drugs/advocacy	X		X			
drugs/non-medical	X		X			
entertainment/games	X					
entertainment/sports	X					
finance						
gambling	X		X			
interactive/chat	X					
interactive/mail	X					
intolerance	X		X			
job search	X					
news						
occult/new age	X					
sex/acts	X		X		X	
sex/attire	X					
sex/personals	X					
sex/nudity	X		X			
sex education/basic	X					
sex education/advanced	X					
sex education/sexuality	X		X			
violence	X		X			
weapons	X		X			

[1]Matches N2H2 Intermediate as closely as possible

[2]Categories match those comparable to other products used in the DOJ report

Notes:
No default configuration is provided.
The DDR (dynamic document review) is turned on under the most restrictive and intermediate settings and turned off on the least restrictive setting.
The sensitivity settings for the DDR are kept at the default levels.

Smartfilter	Tested Configurations							
Categories	**Most**		**Intermediate[1]**		**Least[2]**		**Vendor's Default**	
anonymizers/translators	X							
art & culture								
chat	X							
criminal skills	X		X				X	
cults/occult	X							
dating	X							
drugs	X		X				X	
entertainment	X							
extreme	X		X		X		X	
gambling	X		X				X	
games	X							
general news								
hate speech	X		X				X	
humor	X							
investing								
job search	X							
lifestyle	X							
mature	X		X					
MP3 sites	X							
nudity	X		X				X	
online sales and merchandising	X							
personal pages	X							
politics/opinion/religion	X							
portal sites								
self-help								
sex	X		X		X		X	
sports	X							
travel	X							
usenet news	X							
webmail	X							

[1]Matches N2H2 Intermediate as closely as possible

[2]Categories match those used in the DOJ report

8e6 Technologies	Tested Configurations					
Categories	**Most**		**Intermediate[1]**		**Least[2]**	
alcohol	X		X			
alternative journals	X					
anarchy	X		X			
automobile	X					
banner ads	X					
chat	X					
criminal skills	X		X			
cults/gothic	X					
drugs	X		X			
employment	X					
entertainment	X					
financial						
free hosts	X					
gambling	X		X			
games	X					
hate/discrimination	X		X			
humor	X					
lifestyle	X					
magazines						
news						
obscene/tasteless	X		X			
opinion/politics/religion	X					
personals/dating	X					
pornography	X		X		X	
R-rated	X		X			
search engines						
self-help						
shopping	X					
sports	X					
tickets	X					
travel	X					
web-based e-mail	X					
web-based proxies/anonymizers	X					
web-based newsgroups	X					

[1]Matches N2H2 Intermediate as closely as possible

[2]Categories match those comparable to other products used in the DOJ report

Notes:

No default configuration is provided

Websense (Page 1 of 2)	Tested Configurations							
Categories	**Most**		**Intermediate[1]**		**Least[2]**		**Vendor's Default**	
abortion advocacy	X							
*pro-life	X							
*pro-choice	X							
advocacy groups	X							
adult material	X		X				X	
*adult content	X		X				X	
*nudity	X		X				X	
*sex	X		X		X		X	
*sex education	X							
*lingerie/swimsuit	X							
business/economy								
*financial data/services								
drugs	X		X				X	
*abused drugs	X		X				X	
*prescribed medications								
*supplements/unregulated compounds								
*marijuana	X		X					
education								
*educational institutions								
*cultural institutions								
*educational materials								
entertainment	X							
*mp3	X						X	
gambling	X		X				X	
games	X						X	
government								
*military								
*political groups	X							
health								
illegal/questionable	X		X				X	
information technology								
*hacking	X		X					
*proxy avoidance systems	X		X				X	
*search engines/portals								
*web hosting	X							
*URL translation sites	X							
internet communication	X							
*web chat	X						X	
*web-based email	X							
job search	X						X	
militancy/extremist	X		X				X	
news/media								
*alternative journals	X							

Websense (Page 2 of 2)	Tested Configurations							
Categories	**Most**		**Intermediate[1]**		**Least[2]**		**Vendor's Default**	
~productivity management	X							
~*advertisements	X							
~*freeware/software download	X							
~*instant messaging	X							
~*message boards/clubs	X							
~*online brokerage & trading	X							
~*pay-to-surf	X							
~Bandwidth management	X							
~*internet radio & TV	X							
~*streaming media	X							
~*peer-to-peer file sharing	X							
~*personal network storage/backup	X							
~*internet telephony	X							
racism/hate	X		X				X	
religion	X							
*non-traditional religions	X							
*traditional religions	X							
shopping	X							
*internet auctions	X							
*real estate	X							
Society & lifestyle	X							
*alcohol/tobacco	X		X					
*gay & lesbian issues	X							
*personals/dating	X							
*restaurants & dining	X							
*hobbies	X							
*personal web sites	X							
special events	X							
sports	X							
tasteless	X		X				X	
travel	X							
vehicles	X							
violence	X		X				X	
weapons	X		X				X	

*Subgroups within broader category

~Categories that are available for extra cost

[1]Matches N2H2 Intermediate as closely as possible

[2]Categories match those used in the DOJ report

Appendix F: Details on Excluded URLs & Testing Dates

Thirty searches were conducted against each of six search engines, for a total of 180 searches. Eleven of these searches did not yield any results, due either to transient errors or search engines rejecting the search strings as unacceptable. The remaining 169 searches yielded 7,800 URLs (40 per search). After eliminating duplicates and inaccessible Web sites, we were left with 3,987 distinct URLs that were rated. Of these, 2,983 were found to be either health information or pornography sites, and were therefore tested against the filters, along with an additional 586 health sites that were "recommended" in several online directories, for a total of 3,569 sites that were tested against the filters (3,053 health sites and 516 pornography sites)—details follow.

Unacceptable Search Strings

Some of the search engines that we utilized do not return any results for certain search strings. For example, when searching for the search string "hardcore porn" at Ask.com, a page is displayed that asks if the searcher would like to be redirected to another search engine specializing in adult content. Instead of going to the other search engine, we attempted to view the results available at Ask.com by clicking on the "view search results on Ask.com" link. For some search strings this link displayed a list of results that we used in our analysis but for other search strings there was no list of results available.

Of the eleven failed searches, seven were related to our pornographic search terms. The remaining four resulted from search engines not returning results for the search terms "jock itch", "safe sex", "gay", and "date rape." All of these health searches except "date rape" returned results when tested on the same search engines at a later date, leading us to believe that the errors may have been related to temporary unreachability of the search engines. In retrospect, it would have been a good idea for us to manually check such anomalies on the day the tests were run, and reissue the appropriate search requests, but that was not done. The breakdown among search engines was six failed searches at Ask, four at MSN, and 1 at AOL.

Inaccessible URLs

Each of the 7,800 URLs was tested to make sure that it was accessible. There are a number of reasons that a site may not be accessible either temporarily or permanently. An attempt was made to access each site and when problems occurred the sites were not included in our analysis. 249 URLs were eliminated due to inaccessibility, leaving 6,511.

In addition to the URLs derived from the search engines, there was a list of 633 unique recommended URLs as provided by Google and Yahoo! directories. Of these sites, 29 were inaccessible, leaving 604.

Redirect Handling

When a URL redirected to another URL, both were tested against the blocking software. If either of the two were blocked then the original URL was considered to be blocked. However, a few sites that utilize JavaScript redirects (rather than HTTP or HTML redirects) were excluded from the analysis, since their use of redirects was discovered after testing the blocking software. 23 URLs were dropped for using JavaScript redirects, leaving 6,488 URLs from searches.

Duplicate URLs

Frequently, the same URL was returned from searches at more than one search engine, or from searches on different terms at the same search engine. We eliminated duplicates from the test set. We counted two URLs as the same if they were exactly the same text string, or differed only in that one of them had a trailing '/' character at the end. Thus, two otherwise equivalent URLs that used different domain names to pick out the same server were treated as distinct in our test set. A total of 2,501 duplicate URLs were eliminated, leaving 3,987. Of these, 1,004 were found to be neither health information nor pornography, and were eliminated, leaving a total of 2,983 sites resulting from the searches that were tested against the filters. Of the 604 recommended URLs from the Google and Yahoo! directories, 18 were duplicates, leaving a total of 586 distinct recommended URLs. These 586 sites were also tested against the filters, for a grand total of 3,569 sites tested against the filters.

Unclassifiable URLs

During the rating process (i.e., while sites were being classified by reviewers) 176 URLs were determined to be unusable due to the fact that the corresponding web sites were in a foreign language (and not obviously pornography), did not display any information, etc. In reporting results, these sites were treated as "other" (i.e., not health, not porn).

Other Considerations

Each of the relevant URLs was tested against each filtering product. However, because the sites were tested against the various filters throughout a 24-hour period there were some instances where a site that was originally available was unavailable later while testing a certain blocking product. On average, only about ten search result URLs that were originally accessible were inaccessible during each blocking test. Of the 587 recommended sites that were originally accessible, only one was inaccessible during a blocking test. When results for the various filter/configurations are presented, the percent of sites blocked is calculated as a percent of the total available sites when the particular product/configuration was run.

Dates of Testing

On May 9th, 2002 all search engine URL results were captured, checked for redirects and inaccessibility, cached, and tested against the blocking software. Product blocking lists were updates on May 9th with the exception of N2H2, which was inadvertently last updated on April 30, 2002. Because of difficulties with automating the entire process for AOL, the blocking portion of the test was performed throughout May 9th-11th, 2002.

Initial results showed that CyberPatrol's performance was far worse than other products, which raised our suspicions that the product was not configured properly. Company representatives confirmed that the demo version of the Cyber Patrol blocking product that we had used had an incomplete blocking list, so all of the sites were retested against a new version of Cyber Patrol in mid June 2002 using a complete list, also downloaded in June. Other technical problems on our part made it necessary to retest Symantec Least Restrictive and Intermediate configurations and Websense Intermediate and Most Restrictive configurations on June 28th, 2002 and July 1, 2002 respectively. For this last test, the products' blocking lists from June 8th were used in order to match prior runs as closely as possible.

The 586 recommended sites were tested against the filters on June 10th, 2002 after having their blocking lists updated on June 8th, 2002. The recommended sites were rerun for Cyber Patrol, Symantec, and Websense on the same dates as the reruns for the search result URLs.

Appendix G: Blocked Searches

Some of the blocking products, under certain configurations, blocked some searches, so that a user would be unable to use a search engine to get any results. Table G-1 summarizes the results for each product that did not allow searching on certain search terms when tested on the most restrictive configuration.

Table G-1

SearchTerm	N2H2	Symantec	8e6 Technologies
abortion		X	X
alcohol			X
birth control			
blowjob	X	X	X
breast cancer			
breast feeding			
breast pump			
cancer			
condoms		X	
date rape			X
depression			
diabetes			
diabetic diet			
ecstasy			X
free sex		X	
gay	X	X	X
hardcore porn		X	X
herpes	X	X	
jock itch			
lesbian		X	X
porn		X	X
pregnancy	X		
rape			X
RU486			
safe sex		X	
STD	X		
suicide			
teen porn	X	X	X
xxx	X	X	X
yeast infection		X	

Appendix H: Proportion of Health and Pornographic Results Returned, by Search Engine

Although not the primary focus of this study, the results allow comparison among the six search engines in their propensity to return health, porn, or other sites in response to the search strings on health topics. Alta Vista brought up the most porn. Ask brought up the highest percentage of health sites,though it was not statistically significantly different than Google or Yahoo!. The lowest rate of health information per search was on AOL (76%).

Table H-1

	Google	Alta Vista	Yahoo!	AOL	Ask	MSN
Health	741 (84%)	662 (79%)	737 (83%)	630 (76%)	673 (88%)	588 (79%)
Porn	6 (0.7%)	18 (2.2%)	7 (0.8%)	3 (0.4%)	4 (0.5%)	7 (0.93%)
Other	139 (16%)	158 (19%)	139 (16%)	198 (24%)	92 (12%)	154 (21%)
Total	886	838	883	831	769	749

Appendix I: Filtered Search Engine Review

Some search engines provide a filtered search feature that returns only links to URLs that are consistent with the search engine's filtering criteria. Since this type of feature is fundamentally different from blocking technology, these filtered searches were not included in the study (a filtered search would not prevent a user from accessing any particular URL). However, we did test one such filtered search engine, Google's SafeSearch, and the results of that effort are presented here. The SafeSearch feature was set to "use strict filtering" which filters out pornography and explicit sexual content in both text and image searches.

This simplest performance comparison is the number of health, porn, and other sites returned by the two search engines. One interesting possibility is that a filtered search engine may actually help users find health information more efficiently. This would occur if pornographic sites were filtered out of the results and replaced with pertinent health sites. In the case of Google, however, only six pornographic sites appeared in the top 40 results from all 24 health search strings combined. Only one pornographic site was returned by Google SafeSearch from the health searches.

Table I-1

Search Topic	Percent of sites from regular Google search that were excluded from Google SafeSearch results
Diabetes	7%
Addiction	5%
Depression	5%
Breast cancer	13%
Jock itch	42%
Breast feeding	34%
STD	8%
Safe sex	78%
Pregnancy prevention	18%
Abortion	2%
Homosexuality	28%
Sexual assault	88%
Total (all health topics)	28%

For health searches the difference in the percentages of health, porn, and other sites that were returned by the unfiltered and filtered searches was not statistically significant. For porn searches, however, there were significant differences. Google SafeSearch refused to provide any results at all for four of the six search terms. On the two search terms for which it did return results ("hardcore" and "free sex"), only 9.5% were pornographic sites, compared with the unfiltered search engine's 81% pornography rate.

Because Google and Google SafeSearch start from the same database of sites, it is possible to determine which sites were filtered out of the original Google results. Table I-1 shows the percent of sites that were different by health topic.[Sorry about the mess following – it's just me moving the table up in the next for page break purposes.]

Table I-1 shows that an average of 28% of all sites were filtered out by the filtered search engine. When broken down by classification, 27.2% of all health items returned in unfiltered search on health terms were not returned in the filtered search, 28.0% of "other" items were not returned, and 5 out of 6 porn items (83%) were filtered out.

In conclusion, Google SafeSearch has a much higher filtering rate on health sites than blocking software packages have on their least restrictive settings. However, Google SafeSearch is no more likely to filter health than non-health sites. Since it replaces the filtered sites with more from the same pool, the overall percentage of health results from Google Safe Search is equivalent to that from Google unfiltered search (and higher for searches on porn terms). We have not attempted to determine whether the replacement health sites differ in quality or in point of view from those that were displaced.

References

Curry, A., & Haycock, K. (2001). Filtered or Unfiltered? School Library Journal(1).

Jansen, B., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing & Management, 36, 207-227.

Rideout, V. (2001). Generation Rx.com: How Young People Use the Internet for Health Information. Menlo Park, California: Henry J. Kaiser Family Foundation.

US Department of Justice. (2001). U.S. Department of Justice: Web Content Filtering Software Comparison. Morrisville, NC: eTesting Labs.

The Henry J. Kaiser
Family Foundation
2400 Sand Hill Road
Menlo Park, CA 94025
Phone: 650-854-9400
Fax: 650-854-4800

Washington Office:
1450 G Street N.W. Suite 250
Washington, DC 20005
Phone: 202-347-5270
Fax: 202-347-5274

www.kff.org

The study on which these Appendices are based appears in the Journal of the American Medical Association, December 11, 2002. Additional copies of the executive summary of this study (#3294) are available for free at www.kff.org or by calling the Kaiser Family Foundation's publication request line at 1-800-656-4533. Additional copies of these Appendices (#3295) are available online at www.kff.org.