Open Access to Cancer Research in PubMed: June 1, 2006- May 31, 2011

Melissa Williamson
Stanford University

Introduction

With the evolution and proliferation of the Internet, how information is disseminated and the way in which people access that information has changed radically. Publishers now offer electronic versions of many of their publications. Thus, as individuals search for scholarly literature, rather than referencing paper journals or obtaining personal copies, they primarily download digital copies and either read the material directly off the screen or as a printout (Björk, 2010). Although the cost of electronic media is usually less than that of the traditional hardcopy format and has thus proved to be quite useful to libraries, large organizations, governments, and institutions, it is still often prohibitively expensive for an individual who wishes to obtain material, especially material from multiple journals. The need to remedy this information divide was in part the inspiration for Open Access (OA), the sharing of peer-reviewed scholarly information online globally and free of charge. OA was also " a reaction to the so-called 'serials crisis' of subscription prices that seemed to be constantly rising faster than the rate of inflation" (Björk, 2010). By shifting the cost of publishing to the author rather than the reader, OA allows for broader access to articles by the general public, providing not only "one solution to the problem of restricted access to results of publicly funded research," but also upholding the basic human right to know (Matsubayashi, 2009).

Open access may be achieved in a variety of ways, including via OA journals, web repositories, and personal posting by authors, and it seems especially critical in areas concerned with people's health and wellbeing. According to the Pew Research Center, 72% of Americans use the Internet to look up health and medical information (Pew, 2014). Not only is there a high demand by the general public for such information, but also it is of utmost importance that such information be verifiable. OA is an effort to freely disseminate such verified research. The soundness of this idea has been lauded: "Members of the Science and Technology Committee in the United Kingdom have stated that 'it is better that the public should be informed by peer-reviewed research'" (Matsubayashi, 2009).

It is equally important that patients and healthcare providers have

verified health information to make decisions regarding cancer treatment options. In 2012, an estimated 1.6 million Americans will be diagnosed with cancer, and more than half a million Americans will die of the disease (ACS). In the general population, the most common cancers are, in order (Table 1): prostate, breast, lung and bronchus, colon and rectum and urinary bladder. Among men, cancers of the prostate, lung and bronchus, colon or rectum, urinary bladder, and melanomas of the skin are the most common. Among women, cancers of the lung and bronchus, breast, colon or rectum, uterine corpus, and thyroid occur most frequently (ACS). Men also suffer exclusively from cancers of the testis and penis, while women suffer from cancers of the uterine corpus, ovary, uterine cervix, vulva, and vagina (Table 1). The importance of informing these patients about their illness cannot be overstated. Eheman et al. summarizes the imperativeness of disseminating such information, highlighting the increased satisfaction and health benefits that result, as well as the frustration of being unable to find information: "For people diagnosed with cancer, both the opportunity to provide input about their care and having information about their diagnosis, prognosis, and options for treatment are vital" (2009). Eheman reviews studies showing that many patients find cancer diagnosis and treatment "confusing and frustrating," while informed patients are happier and show "increased involvement in decision making, greater satisfaction with treatment choices, improved coping skills, and reduced anxiety" (2009).

PubMed (PM) is one of the most popular of the available biomedical research databases (Matsubayashi, 2009). It searches references and abstracts on life science and biomedical topics to facilitate the lookup of scholarly information. The National Institute of Health (NIH) maintains PM.

TABLE 1. The Most Common Cancers in the American Population 2012

	Cases	Males (% of pop.)	Females (% of pop.)
Penis*	1,570	1,570 (0.2)	N/A
Vagina*	2,680	N/A	2,680 (0.3)
Vulva	4,490	N/A	4,490 (0.5)
Testis	8,590	8,590 (1)	N/A
Uterine cervix	12,170	N/A	12,170 (2)
Ovary	22,280	N/A	22,280(3)
Uterine corpus	47,130	N/A	47,130 (6)
Thyroid	56,460	13,250 (1.6)	43,210 (5)
Urinary bladder**	72,330	55,600 (7)	16,730 (2)
Melanoma of skin	76,250	44,250 (5)	32,000 (4)
Colon and rectum	143,460	73,420 (9)	70,040 (9)
Lung and bronchus	226,160	116,470 (14)	109,690 (14)
Breast	229,060	2,190 (0.3)	226,870 (29)
Prostate	241,740	241,740 (29)	N/A
Total	1,144,370	557,080 (67.1)	587,290 (74.8)

Note. Data taken from American Cancer Society Cancer Facts and Figures (2012) based on estimate of 848,170 new male cases and 790,740 new female cases –collected May 28, 2012. [a] Available separately at www.cancer.org. [b] Source is http://www.emedicinehealth.com/bladder_cancer/article_em.htm.

Previous Research

A review of recent literature on OA (Table 2) shows that the percentage of OA literature varies greatly by discipline and to a lesser degree by search method but seems to be on a general increase.

In a study of OA from 2001-2002, Antelman surveyed 10 journals from different fields and reported an OA of 17% in philosophy, 29% in political science, 37% in electrical engineering, and 69% in mathematics (2004). In a study of articles published between 1992 and 2003, Hajjem found OA to be between 5% and 16% from his analysis of 1.4 million article records from 10 academic fields, including biology, psychology, sociology, and health. He found OA in biology to be 15%, and in health, 6% (2005). In 2005, Matsubayashi, using PubMed, Google Scholar, Google, and OAIster to look for articles from PM, found the OA percentage in biomedicine to be 27% (2009). Searching for the OA percentages using PM's entire database of articles published in 2005, Björk found an OA percentage of 23.1%; in 2008 he found an OA percentage of 23.3% (2010). Way, researching Library and Information Science articles, found an OA percentage of 27% in 2007 (2010). Björk further found the overall OA percentage in 2008 to be 20.4% and in medicine to be 21.7% using the Google search engine to find peer-reviewed titles (2010).

Aim of This Study

My objective in this study was to make an assessment of the growth of OA and ease of access to research in areas where access to information may be vital to patient diagnosis, treatment, and well-being by comparing the percentage of OA cancer research in PM from June 1, 2006- May 31, 2011 to PM's general OA percentage during that period as well as to previous calculations of PM OA.

TABLE 2. Open Access Studies Covering 1992 to 2008

Researcher	Year(s) covered	Methods	Overall OA percentage(s)	Medicine	PubMed OA
Antelman	2001-2	Four fields /Ten journals each	17-69%	----	----
Hajjem	1992-2003	Web of Science database 1.3 million articles from 10 fields	5-16%	15%-bio 6%-health	----
Matsubayashi	2005	Biomedical field articles in PubMed, Google Scholar, Google, and OAIster	27% (biomedicine)	----	----
Björk	2005	PubMed (general)	----	----	23.1% (overall)
Way	2007	20 Top Journals Library and Information Science	27% (library sciences)	----	----
Björk	2008	PubMed (general)	----	----	23.8% (overall)
Björk	2008	1837 titles / Google first	20.4%	21.7%	-----

Research Design

This study began with a comparison of the availability of current OA articles written in English from June 1, 2006- May 31, 2011 to the findings of previous studies and PM's general journal article OA to determine the rate of increase of OA. The study was specifically focused on medical research due to the high percentage of people in the US (80% of adults according to the 9/1/2010 PEW survey) who seek medical information online each year. Cancer was then chosen as the particular focus of the PubMed search due to its high incidence in the population (1

in 2 for men and 1 in 3 for women) as well as its high mortality rate (1 in 4 people die of the disease) and because cancer could be divided into multiple distinct subcategories (American Cancer Society 2012, p. 1). The five most prevalent US cancers in the general population, the five most prevalent by gender, and the five most prevalent gender-specific cancers as listed by the National Cancer Society were studied. The male-specific cancers were edited to three due to a lack of available statistics on any cancers less common than the top three.

I used PubMed as the database to query availability of articles on each type of cancer and the subset of those articles that is available "full free-text." Note that although there are a variety of definitions of OA, this study accepts the PM "free full-text" classification of an article as synonymous with OA. Also note that because articles available by other means, such as author's websites, institutional repositories, discipline-specific archives, journal websites and platforms, and other portals from third parties outside of PM were not counted, the count given here is less than or equal to the actual number of articles published during the period and available with full free-text. Although this means PM does not have 100% coverage of all the journals that were published from June 1, 2006-May 31, 2011, especially journals published in languages other than English, it was the sole database queried for the sake of simplicity and because the data collected could then be compared with Björk's assertion that in 2005 23.1% of the journal articles in PubMed was OA and in 2008 23.3%. PM was also chosen because of its status as one of the most popular databases in biomedicine (Matsubayashi, 2009). The rate of growth of OA in PM was determined by linear regression of percentages by year.

Data Sources

PM is a database of citation and abstracts that also contains links to full-text articles, some of which are free, and some of which are accessed by subscription. PM is searched automatically with the aid of a special tool called Medical Subject Headings (MeSH) that expands upon search terms to include all logical subsets of the term entered to improve search results.

It is also important to note that because my choice of subject matter, cancer, is the second most common cause of death in the United States, it is possibly represented in PM more than other illnesses.

Creating the Samples of Articles

The five calendar years I studied were June 1, 2006- May 31, 2011. These dates were selected in light of the up-to 12-month embargo period after official publication that publishers frequently impose on posting an article to OA and the NIH policy mandating release of NIH funded research to OA within twelve months of publication. There exists no standard metric of when in or after a year's time an article will be released to OA;

therefore, because I searched articles in June of 2012, the necessary search end date to ensure an accurate count of the full-free text articles funded by the NIH was May 31, 2011. This sampling method cannot account for issues of non-compliance, late submissions, and policies regarding non-NIH funded research.

The search terms used were derived from the American Cancer Society's (ACS) 2012 Prevalence of US Cancers list. The exact search terms used appear in Table 4 as the "Cancers." To simplify lookup, I edited two of the search terms given on the ACS website to their most prevalent type, e.g., "Penis & other genital, male" and "Vagina & other genital, female" are shortened to "Penis "and "Vagina" as Table 4 depicts. Search terms were then entered into MeSH in PM to see if expansion of the term would result in a higher number of hits, and the terms with highest number of hits were chosen because it was not possible to distinguish the nature of the reference without reading each article and it was impossible to rule a reference containing the search term as "unacceptable" to all potential researchers.

The terms were then searched for in PM using the restriction of the years to June 1, 2006- May 31, 2011. I recorded the total number of articles, as well as the number of free full-text articles.

Results

The percentage of OA in PM of general journal articles and of cancer articles from 2005-2011 are shown in Table 3. Note that these OA values are almost universally higher for 2009, 2010, and for June 1, 2006- May 31, 2011 than those established by previous research on OA, summarized in Table 2. The italicized rows signify the numbers found repeating Björk's analysis of PM in his 2008 report. These numbers are also directly comparable to Matsubayashi's 2005 and Björk's 2008 overall findings. This study's finding of 22.34% OA in PM for 2005 is slightly less than Björk's findings of 23.1% OA in PM for 2005, and is lower than Matsubayashi's finding of 27.1% in biomedicine. This study's finding for 2008 of 27.39% OA in PM is however much higher than Björk's 23.8% finding for OA in PM for 2008 and 20.4% finding for general OA and 21.7% for medicine (Table 2).

TABLE 3. Rate of Growth of Open Access in PubMed

Journal Year	All Articles	Open Access	%OA
2005	639,438	142,823	22.34%
2006	683,004	157,695	23.09%
2007	719,707	178,353	24.78%
2008	765,065	209,523	27.39%
2009	746,981	237,352	31.77%
2010	806,954	257,691	31.93%
2011	924,677	209,257	22.63%
6/01/06-5/31/11	*3,634,082*	*1,025,647*	*28.22%*

Journal Year	Cancer Articles	Open Access	%OA
2005	94,907	24,393	25.70%
2006	100,583	26,123	25.97%
2007	108,021	29,596	27.40%
2008	114,871	34,052	29.64%
2009	120,941	38,282	31.65%
2010	131,538	42,470	32.29%
2011	136,666	31,096	22.75%
6/01/06-5/31/11	*547,880*	*167,560*	*30.58%*

Also note there is an increase in the rate of growth of OA across all years except for 2011. This is likely because much of the 2011 research has not yet been released from the twelve-month embargo period imposed by many publishers. Thus, the 2011 percentages were not used in calculating the rate of growth of OA for each category. Using linear regression, the average OA growth from 2005- 2010 in PM for all for journal articles was calculated to be 2.2% per year[1], and cancer articles 1.5% per year[2]. Extrapolating from these growth rates, it would take 36-50 years to achieve complete access to the previous year's literature. Given that PM is currently at a higher level of yearly OA than most other disciplines, this time estimate is probably longer for other disciplines. No attempt was made to analyze whether the articles in PM were reviews or investigative articles as there is no evidence that review articles are more or less likely to be OA and or to be of use to a researcher.

The OA percentages for specific cancers in PM are given in Table 4.

[1] Regression line equation: y=21.411904761905+2.1885714285714x—36 years
[2] Regression line equation: y=25.044285714286+1.4922857142857x—50 years

TABLE 4. Percentage of open access (OA) articles in PubMed by leading types cancer (June 12, 2012)

Cancers	Total Articles	OA Articles (%)
Vulva	1,244	196 (15.8)
Penis	1,179	226 (19.2)
Vagina	4,080	894 (21.9)
Urinary bladder	8,530	1,982 (23.2)
Uterine corpus	290	69 (23.8)
Uterine cervix	10,754	2,634 (24.5)
Melanoma of skin	7,221	1,949 (27.0)
Colon & rectum	2,345	639 (27.2)
Thyroid	10,270	2,822 (27.5)
Lung & bronchus	11,164	3,172 (28.4)
Ovary	16,192	4,708 (29.1)
Testis	4,256	1,340 (31.5)
Prostate	31,353	10,088 (32.2)
Breast	64,637	22,447 (34.7)

Statistical Significance of Findings

The statistical significance of this assessment using difference of proportions, or a z-test, is shown in Table 5. This test is an assessment of how likely it is that the difference between two data values is due to chance. A high level of significance signifies that the occurrence was not due to chance and is denoted by a large negative or positive z value. Positive values signify growth in the table below and negative values signify a lag in the amount of data available. A very high level of significance—above .001 for the majority of the data was found—signifying an overall positive growth in OA publishing of literature in PM since Bjork's 2008 findings, but the large number of negative z values denote a substantial dearth in the research published OA for the majority of cancers here assessed in comparison to other OA publishing in PM.

TABLE 5. Statistical Significance of Findings (February 23, 2014)

Individual cancer research in PM OA % 6/01/06-5/31/11		**Number of articles assessed**	**PM's Journal OA % 6/01/06-5/31/11**	**Number of articles assessed**	**Difference statistically significant at the .001 level? (-4.4>z or z>4.4)**	
Vulva	15.8	1,244	28.22	3,634,082	-9.731887581	Yes
Penis	19.2	1,179	28.22	3,634,082	-6.880601929	Yes
Vagina	21.9	4,080	28.22	3,634,082	-8.965125567	Yes
Urinary bladder	23.2	8,530	28.22	3,634,082	-10.29067401	Yes
Uterine corpus	23.8	290	28.22	3,634,082	-1.672342997	No
Uterine cervix	24.5	10,754	28.22	3,634,082	-8.559676692	Yes
Melanoma of skin	27.0	7,221	28.22	3,634,082	-2.301222023	No
Colon & rectum	27.2	2,345	28.22	3,634,082	-1.097120575	No
Thyroid	27.5	10,270	28.22	3,634,082	-1.618952714	No
Lung & bronchus	28.4	11,164	28.22	3,634,082	0.421923343	No
Ovary	29.1	16,192	28.22	3,634,082	2.48238473	No
Testis	31.5	4,256	28.22	3,634,082	4.751408251	Yes
Prostate	32.2	31,353	28.22	3,634,082	15.58540897	Yes
Breast	34.7	64,637	28.22	3,634,082	36.23934021	Yes

Individual cancer research in PM OA % 6/01/06-5/31/11		Number of articles assessed	PM's cancer article OA % 6/01/06-5/31/11	Number of articles assessed	Difference statistically significant at the .001 level? (-4.4>z or z>4.4)	
Vulva	15.8	1,244	30.58	547,880	-11.30483612	Yes
Penis	19.2	1,179	30.58	547,880	-8.473610604	Yes
Vagina	21.9	4,080	30.58	547,880	-11.99590897	Yes
Urinary bladder	23.2	8,530	30.58	547,880	-14.69489077	Yes
Uterine corpus	23.8	290	30.58	547,880	-2.50534184	No
Uterine cervix	24.5	10,754	30.58	547,880	-13.56667744	Yes
Melanoma of skin	27.0	7,221	30.58	547,880	-6.562399932	Yes
Colon & rectum	27.2	2,345	30.58	547,880	-3.545331932	No
Thyroid	27.5	10,270	30.58	547,880	-6.715336085	Yes
Lung & bronchus	28.4	11,164	30.58	547,880	-4.951061396	Yes
Ovary	29.1	16,192	30.58	547,880	-4.029910561	No
Testis	31.5	4,256	30.58	547,880	1.297535102	No
Prostate	32.2	31,353	30.58	547,880	6.050096212	Yes
Breast	34.7	64,637	30.58	547,880	21.41702376	Yes

Individual cancer research in PM OA % 6/01/06-5/31/11		Number of articles assessed	Björk General OA 2008	Number of articles assessed	Difference statistically significant at the .001 level? (-4.4>z or z>4.4)	
Vulva	15.8	1,244	20.4	1837	-3.223483354	No
Penis	19.2	1,179	20.4	1837	-0.804972631	No
Vagina	21.9	4,080	20.4	1837	1.300930007	No
Urinary bladder	23.2	8,530	20.4	1837	2.59855491	No
Uterine corpus	23.8	290	20.4	1837	1.324238648	No
Uterine cervix	24.5	10,754	20.4	1837	3.807945516	No
Melanoma of skin	27.0	7,221	20.4	1837	5.782733163	Yes
Colon & rectum	27.2	2,345	20.4	1837	5.094730216	Yes
Thyroid	27.5	10,270	20.4	1837	6.356518022	Yes
Lung & bronchus	28.4	11,164	20.4	1837	7.134583113	Yes
Ovary	29.1	16,192	20.4	1837	7.85214258	Yes
Testis	31.5	4,256	20.4	1837	8.840852374	Yes
Prostate	32.2	31,353	20.4	1837	10.57785548	Yes
Breast	34.7	64,637	20.4	1837	12.73095475	Yes

PM's journal article OA % 6/01/06-5/31/11	Number of articles assessed	Björk General OA 2008	Number of articles assessed	Difference statistically significant at the .001 level? (-4.4>z or z>4.4)	
28.22	3,634,082	20.4	1837	7.445430481	Yes
PM's cancer article OA % 6/01/06-5/31/11	**Number of articles assessed**	**Björk General OA 2008**	**Number of articles assessed**	**Difference statistically significant at the .001 level? (>4.4)**	
30.58	547,880	20.4	1837	9.45692109	Yes

Discussion

The differences in the percentages of OA between this study and previous studies can be explained by a variety of factors, including the following: the share of OA is changing with time due in part to 2008 NIH mandate, the number of years that have lapsed between the studies, the number of articles counted/ subcategories used, use of American Cancer Society as a source for article data, use of PM as search engine vs. Google or other search engines, a further release of articles in general and to OA in three the years since Björk's study, and a greater awareness and interest in open access.

However, even at the highest level of OA found in this study, that of breast cancer, still only approximately a third (34.7%) of the cancer research more than a year old is generally accessible. Moreover, as breast cancer is a cancer whose research is benefitted by a particularly successful fundraising campaign, its OA far outpaces that of most other cancers (Townsend 2010). This is a troubling statistic when one considers the cost to patients and healthcare provides who would need to subscribe to a multitude of sites or be a member of a purchasing institution to access information that may be vital to their healthcare decisions. Thus, OA to cancer research, even allowing for the best available statistics, in actuality means access to only one third of the information and research at least one year old. Thus, for both patients and doctors who wish to understand the disease by accessing the most up-to-date research, there is no way to avoid subscription costs with approximately 70% of the research more than a year old not available due to copyright restrictions, and 100% of the current year's research not published OA inaccessible for the same reason.

On comparing the lowest average percentage of general OA availability (28% of journal articles from June 1, 2006- May 31, 2011) with the OA on information about specific cancers during the same time period, this study found that only five of the cancers studied—prostate, breast, lung and bronchus, ovary, and testis cancer—had a higher percentage of OA. This signifies that a great deal of cancer research areas

have a lower-than-average PM OA percentage, which is concerning due to cancer's prevalence and mortality rates. A second troubling observation is that the volume of literature on each cancer seems to decrease with prevalence, so this lack of OA literature, combined with an overall dearth of research into such cancers, creates a definite scarcity of available information for those without subscription access. For example, in a case like uterine corpus cancer, this means there are only 69 articles available from the past five years as opposed to the 64,637 articles available for breast cancer from the same time period (Table 4).

Further Study

These results should be of interest to the general public, cancer sufferers, their support network, or anyone who would like to find more information about the disease online. These results provide a compelling look at the growth of OA studies to date and should also interest academic publishers who monitor OA for their business strategies and copyright policies regardless of their currently held opinion of the sustainability or value of OA. Lastly, research funders like the NIH, who promote the availability of the results from research projects they fund, will also find these conclusions meaningful.

There are numerous ways to extend this study, such as researching the lack of availability of OA articles on gender-specific cancer information and the less prevalent cancers. It would also be interesting to compare the quality of OA articles to non-OA articles as well as domestic OA articles to those from different countries, because medical practices often differ radically in different locations due to culture and regulation. Additionally, it would be interesting to track the growth rate of OA access in other disciplines for comparison.

References

American Cancer Society. (2012, January 1). *Cancer Facts & Figures 2012*. Retrieved June 10, 2012, from http://www.cancer.org/Research/CancerFactsFigures/CancerFactsFigures/cancer-facts-figures-2012

Antelman K. (2004). Do Open-Access Articles Have a Greater Research Impact? College & Research Libraries, 65(5), 372-82. Retrieved June 10, 2012, from http://crl.acrl.org/content/65/5/372.full.pdf+html

Björk, B-C., Welling, P., Laakso, M., Majlender, P., Hedlund T., et al. (2010). Open Access to the Scientific Journal Literature: Situation 2009. PLoS ONE 5(6), e11273. doi:10.1371/journal.pone.0011273. Retrieved June 10, 2012, from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0011273

Eheman, C., Kirshner, J., Johnson, D., Roscoe, J., Rodriguez, E. M., Purnell, J., et al. (2009). Information-Seeking Styles Among Cancer Patients Before and After Treatment By Demographics And Use of Information Sources. *Journal of Health Communication*, *14*(5), 487-502. DOI: 10.1080/10810730903032945. Retrieved June 10, 2012, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024551/

Hajjem, C., Harnad, S., Gingras, Y. (2005). Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. IEEE Data Engineering Bulletin 28(4), 39–47. Retrieved June 10, 2012, from http://eprints.ecs.soton.ac.uk/11688/

Home - PMC - NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved June 10, 2012, from http://www.ncbi.nlm.nih.gov/pmc/

Home - PubMed - NCBI. (n.d.). *National Center for Biotechnology Information*. Retrieved June 10, 2012, from http://www.ncbi.nlm.nih.gov/pubmed

Matsubayashi, M., Kurata, K., Sakai, Y., Morioka, T., Kato, S., et al. (2009). Status of open access in the biomedical field in 2005. Journal of the Medical Library Association 97: 4–11. DOI: 10.3163/1536-5050.97.1.002 Retrieved June 10, 2012, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605039/

Townsend, A. Fundraising Flourishes for Breast Cancer Research and Treatment. (2010, Sep 07). *The Plain Dealer*. Tuesday, September 07, 2010. Retrieved June 30, 2012, from http://www.cleveland.com/healthfit/index.ssf/2010/09/fundraising_flourishes_for_bre.html

Way, D. (2010). The Open Access Availability of Library and Information Science Literature. College & Research Libraries, 71(4), 302-309.

Retrieved June 10, 2012, from http://crl.acrl.org/content/71/4/302.full.pdf+html

Health Fact Sheet | Pew Research Center's Internet Project. (n.d.). *Pew Research Center's Internet Project*. Retrieved March 30, 2014, from http://www.pewinternet.org/fact-sheets/health-fact-sheet/