# Fundamental stocks of knowledge and longevity growth: evidence from Medline and other NIH data

**Frank R. Lichtenberg**
Columbia University and National Bureau of Economic Research
frank.lichtenberg@columbia.edu

A number of authors have expressed the view that a substantial portion of recent gains in the longevity and health of Americans is due to biomedical research and innovation. Cutler, Deaton and Lleras-Muney (2006) "tentatively identif[ied] the application of scientific advance and technical progress (some of which is induced by income and facilitated by education) as the ultimate determinant of health." Kramarow et al (2007) suggested that "medical advances have had an important role in the better health of older Americans." Fuchs (2010) stated that "since World War II…biomedical innovations (new drugs, devices, and procedures) have been the primary source of increases in longevity." And the NIH (2012) says that "thanks in large part to NIH research, Americans are living nearly 30 years longer than they did in 1900."

But empirical support for the hypothesis that most of the recent gains in the longevity and health of Americans is due to biomedical research and innovation is somewhat limited. The statements cited above are primarily based on case studies, and as noted by the Kaiser Family Foundation (2007), "while case studies can explain the impact of certain medical advances…it is difficult to generalize from them to an aggregate or national level." In 2005, the Australian Productivity Commission (2005) published a comprehensive, 605-page review of the literature on the benefits and costs of medical innovation. One of its conclusions was that "while there have been measurable improvements in various indicators of health and mortality…it is difficult to attribute these to health spending, let alone particular technologies, with any degree of precision. This is not to say that benefits of advances in medical technology have not been large, simply that measuring and attributing benefits is challenging." The Commission noted that "most formal studies…have focused on the expenditure impacts of medical technology, partly because costs are more easily identified and quantified than are benefits." Moreover, "most quantitative research relating to the impact of broad categories of technology [on health outcomes] has focused on pharmaceuticals."

The objective of this study is to provide new, robust evidence about the impact of biomedical research and innovation on U.S. longevity growth during the last several decades. This evidence will be based on analyses of the relationship across diseases between the change in the cumulative number of Medline[1] publications pertaining to a disease and the change in mortality from the disease, controlling for the change in incidence of the disease. My approach will be similar to that pursued by Adams (1990),

[1] MEDLINE (Medical Literature Analysis and Retrieval System Online), the U.S. National Library of Medicine's (NLM) premier bibliographic database, contains over 17 million references to articles published (generally since 1949) in more than 5200 current biomedical journals from the United States and over 80 foreign countries.

who studied the relationship between measures of academic science and multifactor productivity growth in U.S. manufacturing industries, using article count data in each science as measures of knowledge. He found that knowledge was a major contributor to productivity growth.

I hypothesize that (1) controlling for the change in incidence, the greater the increase in knowledge about a disease, the greater the reduction in the burden of the disease, and (2) the increase in the cumulative number of publications about a disease is a useful indicator of the increase in knowledge about the disease. Controlling for disease incidence is important, because diseases with large exogenous increases in incidence are likely to have larger increases in knowledge (cumulative publications) and smaller reductions in disease burden. Hence, failure to control for incidence would lead to underestimates of the effect of medical knowledge accumulation and diffusion on health.

Although reliable incidence data are not available for many diseases, they are available for many types of cancer. Hence, this project will assess the impact of medical knowledge accumulation and diffusion on the burden of cancer, using longitudinal, annual, cancer-site-level data on over 45 cancer sites (breast, colon, lung, etc.) during the period 1978-2006.

The burden of cancer can be measured in a number of ways. The measure I will initially use is the age-adjusted mortality rate, which some investigators have argued is the best available measure (i.e., preferable to the 5-year relative survival rate), because it is not subject to lead-time bias.

**Preliminary econometric model**

I propose to estimate difference-in-difference models of the age-adjusted mortality rate using longitudinal, cancer-site-level data on over 45 cancer sites. The equations will be of the following form:[2]

$$\ln(\text{mort_rate}_{st}) = \beta_1 \ln(\text{cum_pubs}_{s,t-k}) + \beta_2 \ln(\text{inc_rate}_{s,t-k}) + \alpha_s + \delta_t + \varepsilon_{st} \qquad (1)$$

where

$\text{mort_rate}_{st}$ = the age-adjusted mortality rate from cancer at site s (s = 1,…, 46) in year t (t=1978,…,2006)

$\text{cum_pubs}_{s,t-k}$ = the cumulative number of Medline publications associated with cancer at site s by the end of year t-k (k=0,1,…)

$\text{inc_rate}_{s,t-k}$ = the age-adjusted incidence rate of cancer at site s in year t-k

$\alpha_s$ = a fixed effect for cancer site s

$\delta_t$ = a fixed effect for year t

$\varepsilon_{st}$ = a disturbance

I hypothesize that $\beta_1 < 0$ and that $\beta_2 > 0$: the log change in the age-adjusted mortality rate is inversely related to the log change in the number of Medline publications and positively related to the log change in the age-adjusted incidence rate. This equation will be estimated via weighted least-squares,

[2] The cancer sites are those included in the National Cancer Institute's SEER Cause of Death Recode shown here: http://seer.cancer.gov/codrecode/1969+_d09172004/index.html

weighting by the mean mortality rate of cancer at site s during the entire sample period ((1 / T) $\sum_t$ $mort_rate_{st}$). The estimation procedure will account for clustering of disturbances within cancer sites. Eq. (1) includes the *lagged* value of cum_pubs (and inc_rate), since it may take several years for medical knowledge accumulation to have its peak effect on mortality rates.

**Data and descriptive statistics**

*Cancer incidence and mortality rates.* Data on age-adjusted cancer incidence and mortality rates, by cancer site and year, will be obtained from the National Cancer Institute's Cancer Query Systems (http://seer.cancer.gov/canques/index.html). Mortality data are based on a complete census of death certificates and are therefore not subject to sampling error, although they are subject to other errors, i.e. errors in reporting cause of death and age at death. Cancer incidence rates are based on data collected from population-based cancer registries, which currently cover approximately 26 percent of the US population; incidence rates are therefore subject to sampling error.

*Publications data.* Data on the cumulative number of publications associated with cancer at site s by the end of year t-k will be obtained from MEDLINE. A distinctive feature of MEDLINE is that the records are indexed with NLM's Medical Subject Headings (MeSH). MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as "Diseases" or "Neoplasms." More specific headings are found at more narrow levels of the eleven-level hierarchy, such as "Intestinal Neoplasms" and "Lymphoma, Non-Hodgkin." There are 25,588 descriptors in 2010 MeSH.

*Descriptive statistics.* The following table shows the cumulative number of Medline publications in 1975 and 2007 for selected cancer sites, ranked in descending order of cumulative publications in 1975.

| MeSH descriptor | cum1975 | cum2007 |
|---|---|---|
| Lung Neoplasms | 18,009 | 115,958 |
| Breast Neoplasms | 14,936 | 147,052 |
| Brain Neoplasms | 13,342 | 64,210 |
| Liver Neoplasms | 12,422 | 77,551 |
| Stomach Neoplasms | 12,240 | 54,832 |
| Uterine Cervical Neoplasms | 10,928 | 45,022 |
| Hodgkin Disease | 9,659 | 27,491 |
| Uterine Neoplasms | 7,690 | 28,684 |
| Melanoma | 7,677 | 50,004 |
| Leukemia, Lymphoid | 7,210 | 20,842 |

| | | |
|---|---|---|
| Lymphoma, Non-Hodgkin | 6,396 | 26,737 |
| Bone Neoplasms | 6,333 | 36,953 |
| Kidney Neoplasms | 6,271 | 40,138 |
| Multiple Myeloma | 5,706 | 23,571 |
| Leukemia, Myeloid | 5,665 | 20,776 |
| Ovarian Neoplasms | 5,484 | 44,612 |
| Colonic Neoplasms | 5,120 | 45,070 |
| Laryngeal Neoplasms | 4,961 | 18,622 |
| Thyroid Neoplasms | 4,873 | 27,683 |
| Urinary Bladder Neoplasms | 4,585 | 33,025 |

Source: Author's calculations from data contained in Unified Medical Language System, http://www.nlm.nih.gov/research/umls/

The growth rate of the cumulative number of publications varied considerably across cancer sites. For example, at the end of 1975, there had been 21% more publications about lung cancer than there had been about breast cancer. At the end of 2007, there had been 21% *fewer* publications about lung cancer than there had been about breast cancer. Also, the 1975-2007 growth rate of the cumulative number of publications about melanoma was much higher than the growth rate of the cumulative number of publications about uterine cancer.

**Some preliminary evidence**

I have estimated two versions of the following special case of eq. (1), using annual data for the period 1978-2006 on 46 cancer sites:

$$\ln(\text{mort_rate}_{st}) = \beta_1 \ln(\text{cum_pubs}_{st}) + \beta_2 \ln(\text{inc_rate}_{s,t-5}) + \alpha_s + \delta_t + \varepsilon_{st} \qquad (2)$$

In this specification, the mortality rate in year t is a function of cumulative publications in year t and the incidence rate in year t-5. In the first model I estimated, I imposed the restriction $\beta_2 = 0$, i.e. I did not control for incidence. This restriction was not imposed in the second model. Estimates of both models are shown in the following table.

| Model | Regressor | Estimate | Standard Error | Z | Pr > \|Z\| |
|---|---|---|---|---|---|
| | | | | | |
| 1 | $\ln(\text{cum_pubs}_{st})$ | -0.109 | 0.113 | -0.97 | 0.334 |
| | | | | | |
| 2 | $\ln(\text{cum_pubs}_{st})$ | -0.347 | 0.143 | -2.43 | 0.015 |
| 2 | $\ln(\text{inc_rate}_{s,t-5})$ | 0.513 | 0.113 | 4.56 | <.0001 |

When incidence is not controlled for (Model 1), the coefficient on the stock of publications is not statistically significant. However, when incidence is controlled for (Model 2), the coefficient on the stock of publications is negative and statistically significant (p-value = .015), and the coefficient on lagged incidence is positive and significant (p-value < .0001). These findings are consistent with our hypothesis that, controlling for the change in incidence, the greater the increase in the stock of publications about a disease (an indicator of the stock of knowledge about the disease), the greater the reduction in the mortality burden of the disease.

Model 1 implies that, if cancer incidence (but not the stock of publications) had remained constant, the cancer mortality rate would have declined at an average annual rate of 0.81%. Model 2 implies that, if both cancer incidence and the stock of publications had remained constant, the cancer mortality rate would have *increased* at an average annual rate of 0.43%. Hence the estimates imply that the increase in the stock of publications about cancer reduced the age-adjusted cancer mortality rate by about 1.24 percent per year during the period 1978-2006. Murphy and Topel (2006) estimated that a 1 percent reduction in cancer mortality is worth nearly $500 billion.

**References**

Adams JD (1990), "Fundamental Stocks of Knowledge and Productivity Growth," *Journal of Political Economy*, Vol. 98, No. 4 (Aug.), pp. 673-702.

Australian Productivity Commission (2005), *Impacts of Advances in Medical Technology in Australia*, Research Report, Melbourne, http://www.pc.gov.au/__data/assets/pdf_file/0003/17193/medicaltechnology.pdf

Cutler D, Deaton A, Lleras-Muney A (2006). "The Determinants Of Mortality," *Journal of Economic Perspectives* 20(3,Summer), 97-120.

Fuchs, VR. (2010), "New Priorities for Future Biomedical Innovations," *N Engl J Med* 363:704-706 August 19, http://www.nejm.org/doi/full/10.1056/NEJMp0906597

Kaiser Family Foundation (2007), *How Changes in Medical Technology Affect Health Care Costs*, March, http://www.kff.org/insurance/snapshot/chcm030807oth.cfm

Kramarow E, Lubitz J, Lentzner H, Gorina Y (2007),"Trends in the Health of Older Americans, 1970–2005," *Health Affairs 16 (5),* September, 1417-1425.

Murphy KM, Topel RH (2006), "The Value of Health and Longevity**,"** *Journal of Political Economy*, vol. 114, no. 5**.**

National Institutes of Health (2012), *Impact of NIH Research on Our Health,* http://www.nih.gov/about/impact/health.htm