Research proposal

The impact of medical knowledge accumulation and diffusion on health: evidence from Medline and other N.I.H. data

Frank R. Lichtenberg

Columbia University and National Bureau of Economic Research

frank.lichtenberg@columbia.edu

27 May 2010

The objective of this research is to provide reliable econometric evidence about the impact of medical knowledge accumulation and diffusion on health. This evidence will be based on analyses of the relationship across diseases between the change in the cumulative number of Medline publications pertaining to a disease and the change in the burden of the disease, controlling for the change in incidence of the disease. 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:[1]

$$\ln(\text{mort\_rate}_{st}) = \beta_1 \ln(\text{cum\_pubs}_{s,t-k}) + \beta_2 \ln(\text{inc\_rate}_{s,t-k}) + \alpha_s + \delta_t + \varepsilon_{st} \quad (1)$$

where

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
α_s = a fixed effect for cancer site s
δ_t = a fixed effect for year t
ε_{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, weighting by the mean mortality rate of cancer at site s during the entire sample period ($(1 / T) \sum_t \text{mort\_rate}_{st}$). 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

[1] 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

they are subject to other errors, i.e. errors in reporting cause of death and age at death.[2] 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 (Medical Literature Analysis and Retrieval System Online), the U.S. National Library of Medicine's (NLM) premier bibliographic database that 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.[3]

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.

---

[2] During the period 1979-1998, cause of death was coded using ICD9 codes. Since 1999, cause of death has been coded using ICD10 codes. An advantage of the National Cancer Institute's Cancer Query Systems is that the mortality data from the two periods have been linked together.

[3] The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences vital to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science as well as biophysics and chemistry. Increased coverage of life sciences began in 2000. The great majority of journals are selected for MEDLINE based on the recommendation of the Literature Selection Technical Review Committee, an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews, e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine, that are special priorities for NLM or other NIH components. These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has special collaborative arrangements. The majority of the publications covered in MEDLINE are scholarly journals; a small number of newspapers, magazines, and newsletters considered useful to particular segments of NLM's broad user community are also included. For citations added during 2000-2005: about 47% are for cited articles published in the U.S., about 90% are published in English, and about 79% have English abstracts written by authors of the articles.

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.[4] 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

[4] The mortality rate was more strongly related to the contemporaneous stock of publications than it was to the lagged stock of publications.

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.[5]

Extensions

The approach outlined above can[6] and should be extended in a number of ways. Feasible extensions include:

- Distinguishing between publications receiving U.S. government support (primarily via the Public Health Service) and other publications
- Distinguishing between different types of publications within a disease area (e.g. distinguishing between publications about diagnosis of a disease and publications about drug therapy for a disease)
- Using weighted rather than unweighted counts of publications, where the weights could reflect the impact factor of the journal in which the article was published, or the number of citations to the article after it was published
- Using alternative measures of disease burden, e.g. the number of hospital bed days
- Using measures of disease burden from different countries[7]
- Exploring the relationship between cumulative publications and other disease-specific indicators of medical innovation (e.g. drug vintage, and utilization of advanced imaging procedures)

Below I elaborate on the first two of these proposed extensions.

Publications receiving U.S. government support. MeSH descriptors indicate sources of financial support of the research that resulted in the published paper when that support is mentioned in the

[5] Kevin M. Murphy and Robert H. Topel, "The Value of Health and Longevity**,"** *Journal of Political Economy*, 2006, vol. 114, no. 5**.**

[6] Performance of some of these analyses would be facilitated by improved access to the complete, or nearly complete, Medline database, which the NLM could presumably provide to me.

[7] For example, Australia has cancer incidence and mortality data, by cancer site and year, similar to the U.S. data. See Frank R. Lichtenberg, "Are Increasing 5-Year Survival Rates Evidence of Success against Cancer? A reexamination using data from the U.S. and Australia," *Forum for Health Economics & Policy*, forthcoming.

article. The following table shows the number of Medline publications indicating three types of research support, by period, during 1975-2008.

Period	Research Support, U.S. Gov't, P.H.S.	Research Support, U.S. Gov't, Non-P.H.S.	Research Support, Non-U.S. Gov't
1975-1979	140,141	47,279	2,035
1980-1984	181,989	46,587	256,692
1985-1989	212,173	57,937	397,216
1990-1993	195,094	52,167	406,251
1994-1997	198,794	51,545	468,175
1998-2008	151,883	41,228	408,218
Total	1,080,074	296,743	1,938,587

Source: 2008 ASCII MeSH Descriptor file, http://www.nlm.nih.gov/mesh/2009/download/asc_abt.html

Moreover, in 1981, NLM began to carry the actual grant number for PHS grants, just as it appears in the original article, in the MEDLINE citation.

Different types of publications within a disease area. In Medline, publications are classified by subheadings as well as headings. Subheadings are used to retrieve frequently discussed aspects of a topic. For example, a publication may have the heading/subheading "Breast Neoplasms/Drug Therapy." The following table shows the most frequent subheadings associated with the heading "Neoplasm."[8]

Subheading	Frequency
Therapy	35,007
Drug Therapy	32,654
Pathology	21,998
Genetics	19,815
Diagnosis	19,347
Metabolism	18,509
Complications	17,294
Epidemiology	16,980
Immunology	15,832
Radiotherapy	14,670

Subheading	Frequency
Etiology	13,418
Prevention & Control	11,806
Mortality	10,750
Psychology	10,750
Physiopathology	8,280
Blood	6,702
Chemically Induced	6,438
Enzymology	5,062
Surgery	4,806
Nursing	4,380

Source: Ovid Medline

[8] There are currently 218,018 publications with the heading "Neoplasm."