2008

The College of William and Mary

Lori Beacham
Ellen Li
Adam Wasserman

[MEASURING THE PERFORMANCE OF EXTRAMURAL FUNDING AT THE NATIONAL LIBRARY OF MEDICINE]

The Thomas Jefferson Program in Public Policy
The College of William and Mary

Table of Contents

Executive Summary

The National Library of Medicine (NLM) has a long history of supporting medical research through the issuance of biomedical research grants. Given the inherent experimental and laborious nature of biomedical research, there has been little emphasis placed on measuring the effectiveness of awarded research grants. Because medical research is associated with the risk of a dead end, it can be difficult to quantify. Although some legitimately question whether there truly is an effective way to measure medical research grants, a solid effort needs to be made to determine if quantitative metrics can be used as a means to evaluate grant performance. One reason for doing so is that Congress has awarded over $27 million per year in appropriations to the National Institutes of Health (NIH). Quantitative metrics can serve a three-fold purpose: 1) to ensure that research grants are fulfilling their intended purpose, 2) to demonstrate to taxpayers and Congress that biomedical research grant awards are productive and are a wise investment, and 3) to hold the grant recipient accountable for grant awards.

The current metrics at NIH are primarily meant to assess the performance of individual researchers. To begin to understand the value of spending on NIH research, this study assumes quantitative metrics can be applied across research grants. In our analysis of a sample of data, our study focused on several quantitative metrics including publication rate, citation rate, H-index measures, and an additional "TJPPP score," which takes into account differences in publishing journals. While there are various approaches to measuring the effectiveness of research, each has limitations that must be acknowledged. For this reason, grant evaluators should qualitatively analyze research grants using a case study approach. By employing both quantitative and qualitative measures, we can begin to evaluate research grants based on overall performance.

To improve the NLM's ability to assess the quality and effectiveness of research grants, the Thomas Jefferson Program in Public Policy (TJPPP) proposes four recommendations to the NLM.

Recommendation 1: Require specific output from each research grant

Recommendation 2: Implement quantitative metrics

Recommendation 3: Establish an external validation process

Recommendation 4: Continue an in-depth application review with benchmarks

I. Introduction

Consisting of five sections, this study attempts to evaluate and/or develop metrics that can be used to measure the performance of extramural funding. Section 1 introduces the issue and provides background information. Section 2 examines the grant review process as well as current metrics used to measure performance of grant research. Section 3 undertakes quantitative analysis based on the information provided for a sample of grants while section 4 discusses the need for qualitative analysis using a case study approach. Section 5 describes our findings and their policy implications. Section 6 provides several recommendations to future grant evaluators.

i. National Library of Medicine and Extramural Funding

Comprised of 27 institutes and centers that address a wide range of medical conditions and diseases, the National Institutes of Health (NIH) is leading the federal government's efforts to promote medical research and, ideally, medical advances ("About NIH: Mission"). Recognizing the large role that the NIH plays in contributing to medical research, Congress doubled its appropriation to $28 million between 1999 and 2003 (McCarthy 2007).

The National Library of Medicine (NLM), which is one of the NIH's 27 entities, exemplifies the NIH's mission as it is the world's largest medical library and a significant funder of biomedical research grants. These research grants play a critical role in enabling medical researchers at universities, medical schools and research institutions to engage in biomedical research that could potentially lead to innovations in technology and medical equipment or improved understanding of diseases and other medical conditions ("Fact Sheet"). Such developments could benefit both the sick and society at large.

The NLM biomedical research grants appear to separate into three categories: bioinformatics research grants, grants supporting the academic community of bioinformatics, and social science-related topic grants.

Within bioinformatics research grants, there are several grants, which include NLM Express Research Grants in Biomedical Informatics and Bioinformatics, Research Project Grants, Exceptional Unconventional Research Enabling Knowledge Acceleration (Eureka) Grants, Exploratory/Development Grants in Biomedical Informatics and Bioinformatics, Predictive Multi-

scale Models of the Physiome in Health and Disease Grants, and Innovations in Biomedical Computational Science and Technology Grants (BISTI). These grants are differentiated by their level of funding and their focus (e.g., preliminary research, long-term sustained efforts, proof-of-concept work, etc.). For purposes of this study, the issue at hand is to utilize a metric or metrics that can measure their effectiveness despite differences in level of investment and purpose.

The second category of grants is those that support the academic community of bioinformatics. The NLM Informatics Conference Grants and the Academic Research Enhancement Award (AREA) Grants exemplify this focus. AREA Grants support faculty who have not previously been NIH grant recipients. While AREA Grants can be evaluated using current metrics as faculty must engage in research and will likely publish articles, the effectiveness of Conference Grants will be more difficult to measure using standard metrics. Although conferences play a crucial role in the sharing of ideas and the dissemination of information, conferences tend not to be cited as frequently as the work of researchers.

The final grant category is social science-related topic grants. This category includes Understanding and Promoting Health Literacy Grants, Advancing Novel Science in Women's Health Research Grants, and Behavioral and Social Sciences Research on Understanding and Reducing Health Disparities Grants. Although certainly related to bioinformatics, these grants, as compared to Bioinformatics Research Grants, serve a distinct purpose. Rather than attempting to build a new model for x or develop a new way to do y, these grants encourage the examination of existing real world situations, similar to the work engaged in by a social scientist ("Grants and Funding").

ii. Policy Importance

Beginning with the Clinton administration, there has been a newfound emphasis on government accountability and performance. The Bush administration continued this focus by developing the Program Assessment Rating Tool (PART), which seeks to eliminate waste in government by promoting efficiency and results in government agencies and programs (Gueorguieva et al. 2008). In part, due to the inherent nature of medical research, which largely begins with hypothesis testing, the NIH and, specifically, the NLM, have not required grant recipients to produce any output.

Given the need for government accountability, it is necessary to determine whether the large amount of taxpayer money that is being awarded in the form of grants in hopes of promoting medical innovations and breakthroughs really is beneficial and outweighs the initial costs. Given the stakeholders' interests, including those of Congress and taxpayers, and the vast resources that are allocated to support the NIH's efforts, analysis needs to be undertaken to determine whether current metrics aimed at evaluating the performance of individual researchers can be used to measure the efficiency of research grants or whether new metrics can be developed. Furthermore, it would be useful to analyze whether such metrics can be applied across various biomedical fields (i.e., cancer, heart disease, informatics, etc.). Especially in times such as these when the economy is poor and NIH research budgets are not increasing as they did in the recent past, quantitative metrics would be useful in demonstrating the effectiveness of research grants and may provide Congress with the information needed to support the NIH's mission and to supply the resources that provide clearly understood benefits for the costs incurred.

II. Literature review

i. What Defines Grant Success?

As part of our study, we interviewed three NIH grant reviewers in an effort to gain a better understanding of the process that is used evaluate grant applications. Although the reviewers mentioned various issues that they felt would lead to a successful research grant, there were seven common themes that each reviewer considered to be essential elements of a potentially successful research proposal.

(1) Feasibility – Is the objective of the research proposal possible given the methods and process? While there should be some room for high payoff, low-probability research, it must be within reason. Space must still be allocated to significant research which, for example, may not cure cancer, but will lead to important medical advances in other less visible areas, such as allergy prevention.

(2) Significance – This factor addresses a wide variety of issues. Most importantly, what is the impact of the particular research on the medical community? How widely will the results be utilized and what advancement can be expected in the relevant fields of science? Another measure of significance is the degree of innovation exemplified in the

research. While research measuring the effects of beets on digestion may be important, such research would not have the same widespread effect as that of a new technological advancement in electronic health records.

(3) Researcher Development – Each reviewer mentioned the importance of using the grant process not only as a method of solid research, but also as a means to develop new researchers and to train the research corps of the future. Because statistics reveal that the average age of researchers is increasing, there needs to be an effort to ensure that this vital work continues well into the future.

(4) Priority Areas – It is important to recognize that medical research has priorities which change over time; different diseases and treatments seem to demand our attention at a given time. Reviewers of grant applications should recognize shifting priorities because such priorities are likely to be valued by the general public. As such, medical research should devote adequate attention to these areas.

(5) Longevity – Will the results have long-lasting effects or will the results merely have short-term implications? Recognizing grant proposals that contain the elements of longevity is especially important in areas such as informatics since technology is changing at such a rapid pace. It would be useless to focus research in areas that are becoming obsolete.

(6) Environment – The location in which the research is being conducted can be an important signal of the grant's likelihood of success. The facilities, equipment, and even proximity of field experts can give proposed research a solid beginning. However, a researcher's abilities may be heightened if he or she is able to successfully complete a research grant in the absence of a productive environment.

(7) Approach – The methodology of a particular research proposal is important when assessing the probability of success. There are proven methods in medical research and, although there is always room for innovation and improvement, some approaches have higher chances of succeeding.

Having had the opportunity to speak with several grant reviewers, it is readily apparent that the review process for grant applications is comprehensive (Hripcsak, Lehmann, and Mandl, personal communication, 2008). There is no doubt that the existing grant application process is carefully designed to ensure that only the most deserving and qualified applications receive

grant funding. Despite such an extensive grant review process, there is no guarantee that the grants will be effective. For this reason, steps need to be taken to devise metrics that can be employed when seeking to evaluate the effectiveness of biomedical research grants. However, it should be acknowledged that there are some inherent difficulties in accurately measuring the efficiency of biomedical research grants.

ii. Program Assessment Rating Tool (PART)

One area we looked to for guidance on measuring research grants was the Federal Government's Program Assessment Rating Tool (PART). PART evaluates hundreds of government programs and requires each to develop metrics to assess their performance (PART Website). There are several research programs involved with the PART process, but they were not helpful in our evaluation. This finding was due again to the uniqueness of medical research. The grants assessed through PART were in fields of study with very clear and measurable outcomes. Outcomes in medical research, however, are extremely difficult to measure because of the time lag from research breakthrough to measurable medical outcome. In some medical research, such as drug and clinical trials, it is easier to assess success and failure. These trials, however, are largely the result of successful research that occurred years and sometimes decades prior. In all, we found little value in the metrics measured in PART as a guide to our project.

iii. Bibliometrics

Bibliometrics is a process by which to set of methods used to assess studies. A significant approach to research evaluation, citation analysis is a commonly used bibliometric method ("Bibliometrics"). This study will highlight the various applications of bibliometrics in the evaluation of scientific research. Specifically, this section will analyze the strengths and weaknesses of four primary metrics, (i.e., publication rate, citation rate, the journal impact factor and the h-index), which are currently utilized when evaluating the performance of individual research scientists, and assess their applicability to biomedical research grants. By applying the current metrics to a sample of research grants from the NLM and the National Cancer Institute (NCI), a determination can be made concerning their usefulness in measuring the productivity of biomedical research grants.

a. Publication

Having a basic understanding of the multiple categories of the NLM research grants, efforts can be undertaken to assess current quantitative metrics. Although each varies in terms of the output being measured as well as the degree of objectivity, there are four bibliometrics that are currently employed when attempting to evaluate the performance of individual researchers. These metrics are currently used to evaluate the performance of individual scientists, but they can also be applied to the evaluation of biomedical research grants since these grants are the financial vehicles by which scientists and researchers are able to fund their work. The importance of bibliometrics in evaluating research performance is growing because such quantitative metrics appear to provide some degree of objectivity and can be easily used to advance or deny grant awards to researchers (Butler 2008). The two most objective outcome measurements are publication rate and citation rate.

The number of publications is simply the number of articles that the researcher was able to publish as a result of the research funded through the grant. It is generally thought that grant efficiency is positively correlated to the publication rate because as the publication rate increases so does grant efficiency. On its face, the number of publications would appear to be a strong indicator of research because a high publication rate may indicate more work on the part of the researcher.

While the publication rate is a commonly used metric, there are several limitations that should be addressed. First, publication rate would likely be an inaccurate reflection of the research attempted in instances of “failed” research because researchers are hesitant to publish articles that demonstrate their shortcomings. Even if researchers were to write such articles, they would likely have difficulty publishing their work because journals, which receive hundreds of articles and can only publish a limited number, tend to favor articles that demonstrate positive results. On the other hand, it can even be difficult for researchers to publish successful, quality work if the subject matter is lackluster or unpopular. Third, publication requirements may prompt researchers to sacrifice quality for quantity (Butler 2008). For example, should a researcher who published a 50-page, high quality article be counted as having two publications when a comparable researcher published two, 25-page articles possibly in an effort to increase his publication number and, quite possibly, citation rate?

Additionally, questions of equity may arise concerning the grant amount and the number of researchers committed to a grant. Should evaluators consider the actual grant award when looking at publication rate since a researcher working from a smaller grant may not have resources and capabilities similar to that of a researcher who was given a larger grant award and, thus, will not have as many articles to publish? If so, a more accurate metric might be publication rate per grant award total. The grant award might also affect the number of researchers involved in a biomedical research project as a larger grant award will likely equate to a larger project and, thus, more researchers. Are grants supporting more researchers more likely to produce more publications because there may be more findings to discuss or sections of the article can be more evenly distributed?

b. Citation

From a researcher's perspective, one motivation for increasing publication rate is its possible influence on citation rate, another metric. Citation rate refers to the number of times a researcher's publication is referenced in the publications of other researchers (Butler 2008). George Hripcsak, a former chair of a review committee at Columbia University**,** noted that it is not difficult for researchers to publish papers that are cited, especially if the citations occur in article reviews. However, such papers might not be influential, thus detracting from the validity of citations as an accurate indicator of the effectiveness of grant research (Hripcsak, personal communication, 2008).

Citation rates can also influence publication rates, however, given that citations tend to be unevenly distributed (i.e., a large percentage of a researcher's publications are infrequently cited while only a few publications are frequently cited). To minimize the potential for a researcher to cease publication efforts if his initial writings happen to be heavily cited, the Australian government, which invests in medical research through its large appropriation to two research councils that promote medical research and discoveries through research grants, also adopted a percentile distribution approach that is used in conjunction with citations per publication. Utilizing the Thomson database, which provides an annual listing of cited publications, Australian evaluators were able to determine the number of publications per year and their respective rankings (i.e., the top 1%, 10%, 20% and 50% of cited publications) (Butler 2008).

c. H-index

Incorporating both publication rate and citation rate, the h-index is meant to provide an overview of the quality and extent of a researcher's work. A scholar with an index of h has published h papers, each of which has been cited by others at least h times. It is a balanced metric between the number of published articles and the number of citations, reflecting both the number of publications and the number of citations per publication. A single number, the h-index arguably improves ease when evaluating performance by representing the highest number of publications with, at the very least, the exact number of citations in each publication.

Hirsch asserts that significant differences in the number of publications or the number of citations between two researchers could still result in similar h-indexes and, therefore, similar research performance levels by offsetting few publications that are highly cited with many publications that receive low-to-average citations. Some would argue, though, that such a result could mask quality because high performance would be more accurately reflected in publications with higher citation rates. Because an argument could be made that such high citation rates and, thus, possibly a high h-index might not reflect the quality of an individual researcher's work since there may have been a large number of co-authors for a publication, it is advised that this metric not be used when comparing a grant with one researcher to a grant with many researchers. Furthermore, when calculating one's h-index, comparisons should be limited to the researcher's specific field since citation rates vary among disciplines and, more specifically, topics (Hirsch 2005).

d. Journal Impact Factor

The impact factor, often abbreviated IF, is a measure of the citations to science and social science journals. It is frequently used as a proxy for the importance of a journal to its field ("Impact Factor"). Given its usefulness, the impact factor should be considered when evaluating the outputs of research grants.

The impact factor has a complex relationship with the citation rate. Citation rate is a determinate in a journal's impact factor, but journals also play a crucial role in the rate of citations. In combination with the publication rate, the citation rate can influence a journal's impact factor since a journal's impact factor is calculated for a given year by determining the number of citations for that year from articles published by the journal, for example, in the

previous two years and dividing that figure by the number of "citable," published articles in the specified journal during the two-year time frame (The PLoS Medicine Editors 2006).

Rather than simply focusing on the total number of articles written by a researcher, Lewison et al. (2003) found that peer reviewers of grant applications tend to value articles in highly ranked publications. Describing this view as "elitism" as opposed to "egalitarianism," the authors attempted to measure the geographic equality of grant distribution in Great Britain. In theory, there should be a direct relation between an area's ability to perform research, measured in total number of publications originating from that area, and the amount of grants received, measured by the number of papers citing specific grantee organizations over a five-year period. Measuring the 'quality' of the journal through the use of the impact factor, the authors found that the elitism theory was consistent with the majority of data. The value of the journals rather than the shear number of journal articles best explained the variation in allotted grants. To directly quote the authors, "There are lessons here for would-be grant applicants: papers in highly cited journals count much more than ones in 'ordinary' journals" (Lewison et al. 2003). It is important to note that, although articles published in prestigious journals with a strong impact factor are considered to be superior in quality, thus prompting researchers to focus their efforts on publishing in these journals, the journal impact factor is not an accurate reflection of an individual scientist's research performance because it represents citations to all of the journal's articles (The PLoS Medicine Editors 2006).

Journal readership and the specific biomedical field being reviewed can also affect the journal impact factor. Some journals may be widely noted for the scholarly knowledge they impart, but may not be heavily cited because their readership is primarily professionals, not researchers. Alternately, the type of biomedical research will play a large role in determining the type of journal in which the publication will be available. Clinical-based research, as compared to other types of biomedical research, is less likely to be heavily cited because clinical journals are not cited as frequently as other types of journals.

A clinical-based researcher, however, should not feel pressured to publish in more recognized journals because his findings may not have as much of an impact on the specific field given the journal's audience. The National Eye Institute learned this lesson after it made a conscious effort to reach a larger audience by publishing their highly specialized articles in more general journals with high impact factors. Although the papers were cited at a higher rate than

before, their actual to expected citations ratio decreased because the topic-specific articles did not receive the average citation rate that was standard for the journal (Lewison 2002).

One possible means of correcting for the tendency to publish in "alternate" journals that may yield higher citation rates, but have a lower impact would be for evaluators to determine the top-ranking journals in a specific field. However, such an approach would be somewhat subjective. While a panel of experts in the specific field would be assigned the task of defining "top-ranking," there is room for biasness (Butler 2008).

e. Clinical Guidelines

Discussing the use of clinical guidelines as an intermediate outcome measure and seeking to evaluate the "payback" of biomedical research, Grant et al. (2000) concentrate their analysis on papers cited in clinical guidelines in the United Kingdom. Aimed at tracking the flow of knowledge from the laboratory to the clinic, this approach and its implications are mainly applicable to applied research evaluation. Focus is placed both on primary outputs (e.g., publications in the serial peer-reviewed literature) and secondary outputs (e.g., evidence-based clinical guidelines). The authors found that it usually takes three years for a paper to be cited in clinical guidelines while the median knowledge cycle time (i.e., the time between a paper's publication and its citation in a clinical guideline) is eight years.

Interestingly, Grant et al. (2000) propose a broader way to evaluate research: to disaggregate the research process and assess the "payback at each stage." They mentioned four levels of journals: basic, clinical investigation, clinical mix, and clinical observation. While a citation in a clinical guideline does not guarantee an impact on health and therefore a payback on the research investment, it could be considered an indicator of research utility and, thus, an intermediate outcome, or "secondary output" (Grant et al. 2000).

Although it does not appear to be a frequently used metric in evaluating the performance of a grant, grant income or the number of grants received as a result of the specific grant being evaluated for effectiveness may be a useful metric. Commonly employed as a tool in assessing the strength of grant applications because it is a reflection of a larger acceptance of a researcher's intended work and goals, this metric could also be used as an output indicator since it would serve to validate a grant recipient's work and to recognize their contributions to the

field of study thus far. This metric should not be used as a sole indicator of performance, however, because it is difficult, even for highly qualified researchers, to receive grant funding, especially in light of a general decline in research funding given the economy and an increased focus on other priorities (Butler 2008).

f. Other Bibliometrics

In order for the above-mentioned metrics to perform their desired functions, which is to measure the effectiveness of research grants, articles must be published in a journal. Recognition of this universal shortcoming is significant because grant performance, if indicated solely by the aforementioned bibliometrics, will be minimized if the researcher contributed to sources such as books, magazines and the internet. Although not a traditional form of publication and likely to reach a different audience, these alternate forms of publication can, nonetheless, have an impact on policy and educational research (The PLoS Medicine Editors 2006).

- Hypertext

Recognizing the growing importance of and reliance on the internet, Falagas et al. (2006) advocate the use of hypertext metrics. Falagas et al. (2006) assert that hypertext metrics will more accurately measure the prestige of an article, as compared to existing traditional measures, because current metrics fail to take into account the difference in publication and citation habits in different scientific communities. Furthermore, there is no agreed-upon consensus on which publication should be considered in the metric and an advantage is given to someone who writes many articles, but may not receive any citations.

Falagas et al. (2006) claim that hypertext metrics can reflect what the average researcher in a particular field does and gives an advantage to a few well-cited articles. At the core of hypertext metrics is measuring the relative connectedness of an article within the field. This process is established through a converted distance matrix, which takes into account the structure of who cites whom and whom is cited by whom. Within a normal application of hypertext metrics, research focuses on compactness, which captures how well connected a document is, and stratum, which measures the depth or linearity of the link structure.

Compactness is used as a measure because it is a way to roughly estimate the citation behavior of a community.

Because Falagas et al. (2006) believe that publications should be considered bi-directionally, they attempt to actually measure the rank of an author within a given field by creating the metric of citation prestige, which is based on stratum. In defending relative citation prestige as a method for measuring rank, Falagas et al. (2006) assert that a researcher gains prestige by having multiple well-cited papers, but loses prestige when many papers are uncited. Citing one's own work has no effect as does a group of authors who often cite each other. While this approach is novel and considers the internet as a useful tool in disseminating medical research, its potential as an effective metric will not be fully recognized unless it can be adjusted for article citations rather than webpages (Falagas et al. 2006).

- World Scale Scores

Lewison et al. (2007) assert that the selection of metrics, such as the number of citations and the journal impact factor, usually results in relatively different scores since citation indices are biased when comparing basic and applied research. However, the authors believe that the aforementioned metrics could be used in a matrix, which would be capable of producing "World Scale" scores. These scores could evaluate groups of researchers, and, thus, guide decisions concerning the allocation of research funds within an institution or nationally or assist in academic promotions because it would measure what percentage of the output an entity receives in a citation score (or other metrics) sufficient to place it in the top 10% (or another specified level) of the world production in that particular domain.

The concept of "World scale" was borrowed from the oil tanker charter market, in which the output from an entity is compared to that of the world at different levels of excellence. For example, at a percentile level of excellence, the United States has b% of papers that reached that level according to the matrix selected and the score for the United States is (b/a)*100. Lewison et al. (2007) propose reference to a multidimensional profile composed of a selected array of metrics to assess merits of research. In addition to traditional metrics including publications, citations and impact factors, Lewison et al. (2007) employed kite diagrams showing matrices that included relative esteem value, counts of patents that cite references within the subfield, citations in clinical guidelines, rating journals by researchers on a scale from "excellent"

to “decidedly secondary,” citations in journals actually read routinely by clinicians, policy makers, health professionals, researchers, and the general public, citations in governmental and international policy documents, citations in textbooks, and presence of researchers on journal editorial boards (Lewison et al. 2007).

Because the biomedical field is vast, ranging from informatics to genomic research, each field will vary in terms of basic approaches and methods. Genomic research is likely to be more clinical in nature and involve laboratory testing and trials while informatics is primarily the application of theory and research. For example, instead of solely employing citation rate for the NLM grants that concentrate on communications technology and imaging technology, evaluators may want to consider patents applied for from the programs or another measure that recognizes how and to what extent the technology is used within the field.

When attempting to measure research and development (R&D) portfolios, specifically technology portfolios, Lin and Chen (2005) suggest the following four metrics: patent quality (citations received per patent), R&D efficiency (logarithm of the number of patents received per million of R&D expenses), R&D effectiveness (logarithm of the number of citations received per million in R&D expenses), and Intellectual Assets Intensity (logarithm of the number of patents received per total assets). Not disregarding the contribution of citation rate as an indicator of performance, this approach combines citation rate with the number of patents. Furthermore, in addressing the measure of patent citation, a Herindahl-type index was employed in an effort to examine the extent to which a patent cites previous patents and the extent to which a patent is cited (Lin and Chen 2005).

iv. Caveats to Bibliometrics

a. Limitations

Clearly, the aforementioned discussion concerning metrics focuses heavily on citation rate as a tool for evaluating the effectiveness of research grants. However, as INSERM, a French medical research organization, suggests, there may be flaws in relying on certain databases when indexing citation scores. While our study did not rely on the Thomson database when calculating the citation rate of published articles, INSERM, a French medical research organization, asserts that Thomson improperly indexed their articles and the articles of a British medical organization, MRC. Thomson’s Web of Science properly recognized all of the

publications by the organizations, but the information was indexed into the Essential Science Indicator improperly. Because a manual search of the top 1% of articles revealed that Thomson was off by as much as 80%, INSERM concluded that it is difficult to place complete confidence in Thomson's database for ranking article importance (Haeffner-Cavaillon 2005).

Similarly, Scopus, which is the largest abstract and citation research database, has gained recognition from the research community worldwide, including endorsements from two prestigious Australian universities, in part, because it contains information for over 15,000 peer-reviewed journals. Such endorsements are significant given Australia's commitment to its recently adopted Research Quality Framework, which emphasizes the development of indicators, specifically quantitative measures, which can be used to evaluate the quality and usefulness of research ("Two Australian Universities"). Despite such recognition, Meho and Yang (2007) assert that Scopus, while it tends to be more accurate than Google Scholar, contains fewer citations due to differences in database coverage (Meho and Yang 2007).

Despite the caveats that exist, the potential role of current metrics in evaluating the efficiency of research grants should not be discounted. Even J.E. Hirsch, a physicist at the University of California, San Diego and the creator of the h-index, realizes the potential benefits of bibliometrics despite his recognition that "Nobel prizes do not originate in one stroke of luck but in a body of scientific work" (Hirsch 2005).

b. Are Unsuccessful Grants "Failures?"

To a large extent, research in the medical field is largely conducted through experimentation and hypothesis testing. This trial and error approach can result in a researcher obtaining results contrary to what he or she had anticipated, thus leading a researcher to devote years to research before witnessing a significant breakthrough. Do such results indicate that the grant was not efficient and, therefore, a waste of taxpayer dollars? Many would argue that the research, despite unfavorable results, was not in vain because it could have given the researcher invaluable insight concerning possible future research improvements and/or prevented another researcher from making the same "mistake."

A primary example of this phenomenon is evident in the research behind Herceptin, a drug used in breast cancer therapy. Once the HER 2 gene was discovered, it took researchers another five years to determine the protein that was created by the gene and then an additional

three years to identify the relationship between the HER 2 gene and breast cancer. Only then were researchers able to devote resources to the development of the drug, Herceptin (Delaney 2002). Had this project been a NCI grant and been evaluated a few years after it commenced, it would have likely been deemed a failure because there would not have been enough time for researchers to determine what protein was created by the gene. In retrospect, however, it is clear that the Herceptin research was successful. Not only has the drug been proven to be a successful therapy in fighting breast cancer, but also the research may provide useful information to other researchers in their quest to cure other forms of cancer. Because metrics cannot fully account for such "false positives," critics argue that biomedical research grants cannot be validly measured.

III. Quantitative analysis

One of the first questions that arise is how can the performance of grants be quantitatively measured. Ideally, in order to measure the "rate of return" of research grants, we measure the monetary amount of net positive effects upon society based on the new knowledge generated by research grants. In reality, however, it is difficult to identify the positive/negative effect of research outputs, let alone attach monetary value to it. Therefore, in this report, we will focus on identifying and measuring the outputs of research grants and articles published in peer-reviewed journals.

i. Data

We were given a total random sample of 60 grants, issued by the National Library of Medicine and the National Cancer Institute in 1997 and 2002. Within each year and each institution, 15 research grants (R01) were selected.[1] The basic information that was provided included grant number, project start and end date, principal investigator's name, estimated age[2] and housing institute.

[1] Data source: HIST QVR, retrieved on Oct. 16, 2008.
[2] Birth year is estimated by subtracting 22 from the year the Bachelor's Degree was obtained.

a. Collection

To identify the articles published by each grant, we searched the grant number without its session number on PubMed. For example, for grant "LM007593-02" issued in 2002, we searched ""LM007593"[Grant Number]" on PubMed,[3] and retrieved a list of 32 articles that claimed the grant "LM007593" as one of its grant sources. If the grant was supporting a project that commenced before 2002 then there may be articles published before 2002 that were supported by the previous session of the grant. These articles are excluded from our analysis because the amount awarded in 2002 did not contribute to the research outputs (e.g., articles which were published before 2002).

We then searched each qualified article on Scopus, "the largest abstract and citation database of research literature," and determined how many times an article was cited each year since publication (Scopus Website). Knowing the journal in which each article was published, we searched the journal names using Journal Citation Reports on Thomson Reuters Website to locate the journals' impact factor. Most of the journal impact factors we used in our analysis were released in 2004. Because the sample grants were issued in 1997 and 2002 and the articles were published, on average, approximately three years after the grant's commencement, we deem it inappropriate to use the most recent releases of impact factors.

Upon completion of data collection, there are two datasets with regard to the 60 grants. The first dataset consists of 60 observations with variables including basic grant information, principal investigator's information, and total publications and citations received for each grant. The second dataset contains the detailed information for each published article. It has 588 observations covering publishing year, article name, authors, journal title, journal ISSN, journal impact factor, publishing volume and issue, and the number of citations received each year since publishing year.

b. Limitations

There are several limitations to our data collection methods. First, publications that did not claim receipt of NIH extramural funding were omitted. It is possible that articles submitted to PubMed failed to report that they were funded by certain extramural research grants. Second,

[3] To search grants see: http://www.ncbi.nlm.nih.gov/sites/entrez/

the citation database, Scopus, may not cover all of the publication and citation records. Third, the journal impact factors we collected for each article are all from the year 2004, or 2007, if earlier data was missing. The journal impact factors are used as weights of articles in calculating the "comprehensive impact" of articles. Since they represent the extent to which each article could have an impact upon other research, the article's potential impact should be associated with the impact of the journal for the year in which it was published. Ideally, the impact factor should be matched with an article by the publishing year (e.g., if an article was published in 1999, we should identify its journal impact factor in 1999 instead of 2004 or 2007). Currently, we do not have access to the journal impact factors for every year. In the future, if time permits and researchers are granted full access to journal impact factors from previous years, each article should be matched with the journal impact factor of its publishing year.

c. Other Options

Besides the journal impact factor, we could use comparable indexes to represent the weights of publishing journals. The potential candidates include: 1) trend line, which is calculated by dividing the number of citations received in a year by the total number of articles published in that year;[4] 2) immediacy index, the average number of "immediate citations" received by the published articles in a certain year; 3) Article Influence Score (AI), a measure of the average influence of each article during the first five years after publication; and 4) Eigenfactor Score(EF), a measure of the overall value provided by all of the articles published in a given journal in a year based on PageRank algorithm (Bergstrom, 2007).

[4] See Scopus website: http://www.info.scopus.com/journalanalyzer/

ii. Metrics Development

Table 1. Descriptive Statistics

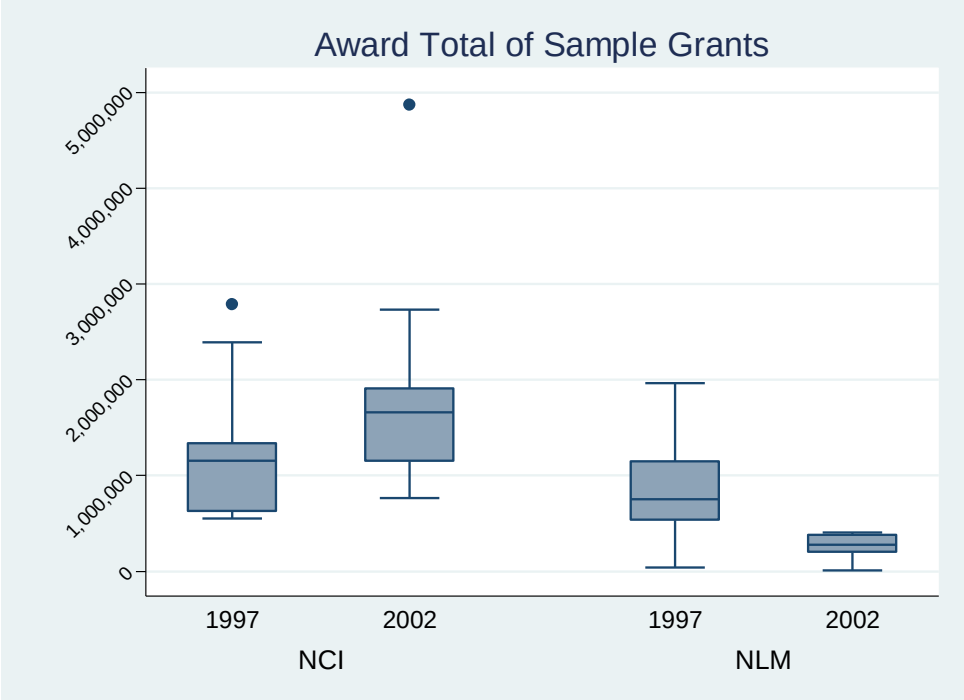
Variable	Mean	Std. Dev.	Min	Max
Total award ($)	1,017,207	838,950	8,620	4,871,752
Length of research (in months)	62.08	58.29	4	248
# of articles published	9.55	10.08	0	38
# of citations received	175.77	293.89	0	1,751
H-index	5.07	4.92	0	24
age of grantee at time of application	46.30	8.33	31	65
highest journal impact factor	6.37	6.15	0	32

a. Grants Information

- Total Award

The award amount is approximately $1 million on average for our sample of 60 grants (Table 1). However, the award amount differs greatly between institutions, even between years. As is shown in Figure 1, grants issued by the NCI are generally larger than those issued by the NLM, and the gap between the two institutions became larger in 2002 as the awards by the NCI increased and the awards by the NLM decreased.

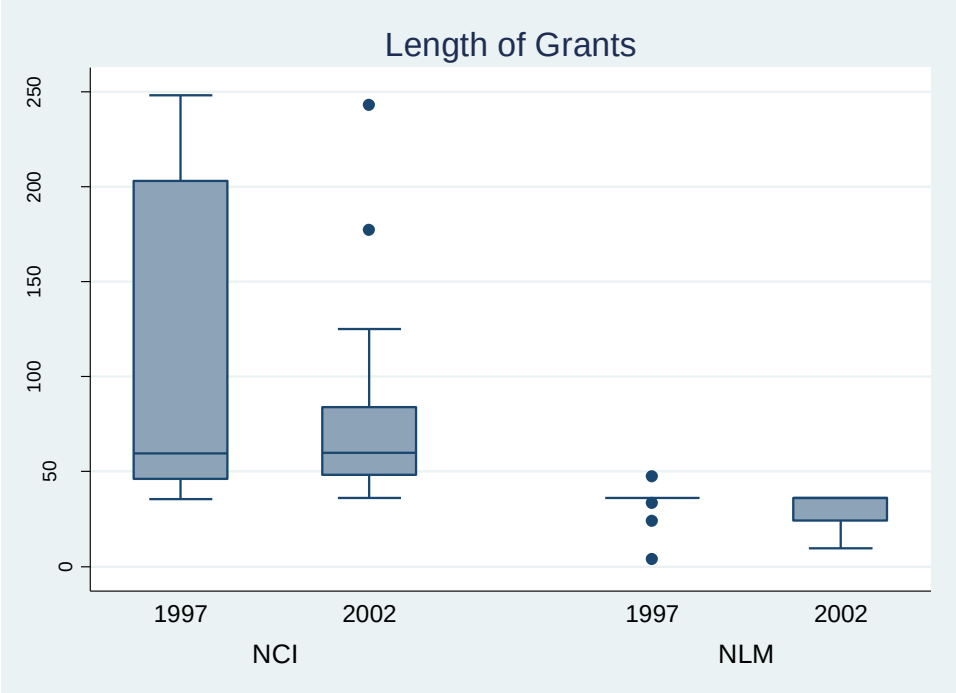
Figure 1. The Award Total of Sample Grants by Institution and Granting Year



- Length of Research

We were given the start date and the end date of each grant. For the renewal sessions of grants, the start date is the start date of the first session of the project. The length of the research is computed based upon the start date of the first session of the project and the end date of the current session of the project. As can be seen from Figure 2, grants issued by the NCI in 1997 are mostly renewal grants from previously issued grants and supported by research with a longer horizon, five of which were initiated in the 1980s.

Figure 2. Length of Research



On the contrary, grants issued by the NLM are primarily supporting bioinformatics research, which is of shorter duration and generally finishes within one grant session.

- Type and Location of Housing Institution

Most of the research projects are performed by researchers in university medical schools, while a small portion of the projects are housed by independent research institutes and hospitals, some of which are affiliated with medical schools.

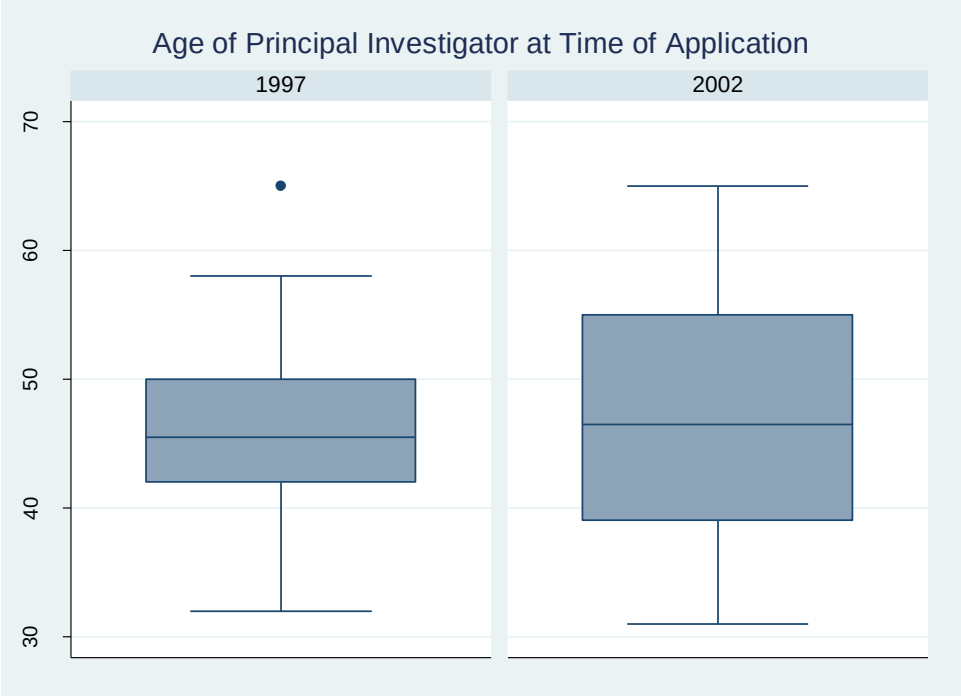
Figure 3 illustrates the geographic locations of the NLM sample grants issued in 1997 while Figure 4, provides the same information for the NLM sample grants issued in 2002. It appears that grants issued in 2002 are more concentrated with regard to housing institution. Among the 15 sample grants issued by the NLM in 2002, eight grants are issued to support research projects housed by three institutions: Mayo College of Medicine, Rochester, Columbia University Health Sciences, and Brigham and Women's Hospital. Also, 2002 grants are more concentrated geographically. All but two grants are issued to researchers in the Northeast.

Figure 3. Geographic Location of Housing Institution of 1997 NLM Sample Grants

Figure 4. Geographic Location of Housing Institution of 2002 NLM Sample Grants

- Age of Principal Investigator

Figure 5. Age of Principal Investigator at Time of Application

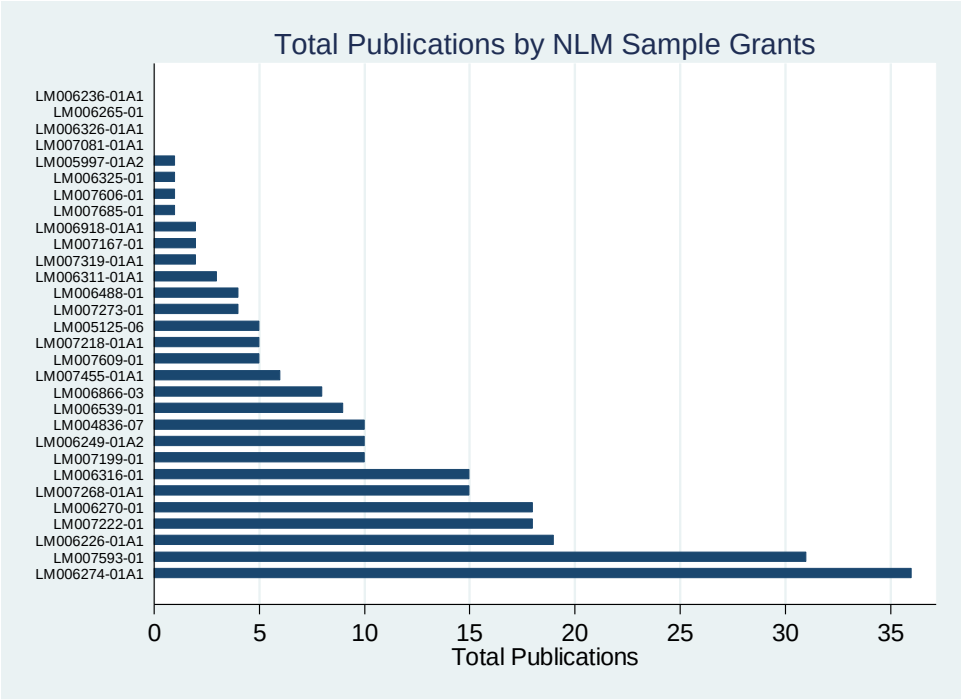


As Table 1 reveals, the average age of the principal investigator at the time of the application is 46.3 for our sample of 60. Furthermore, as can be seen from Figure 5, one-half of the researchers were between the ages of 42 and 50 at the time of grant applications. In 2002, the range varied between the ages of 39 and 55.

b. Unscaled Metrics

- Total Publications

Figure 6. Number of Published Articles Supported by NLM Sample Grants

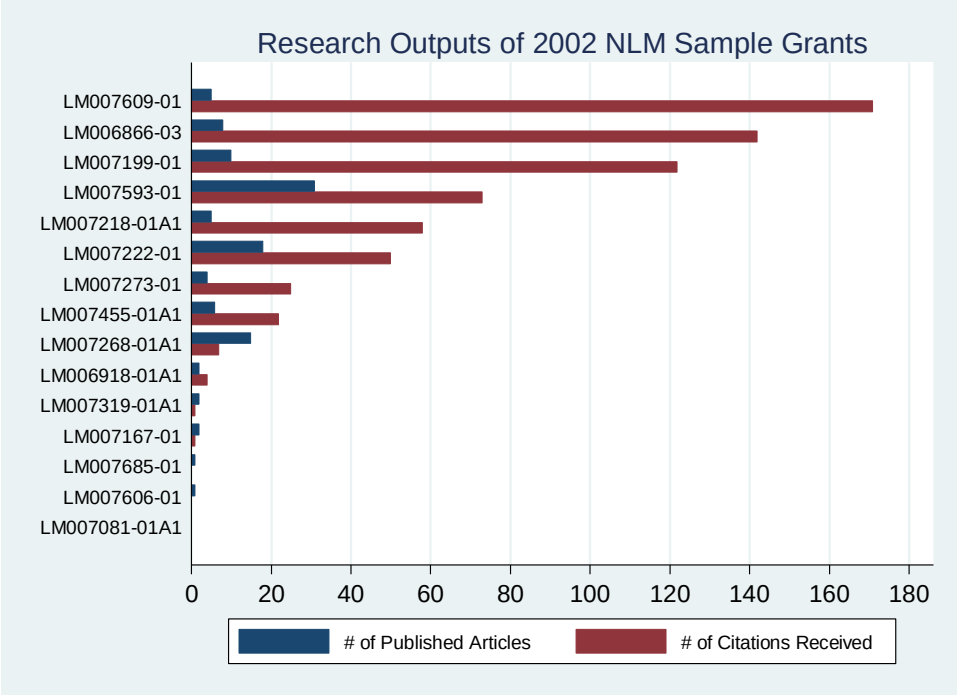


Of the 60 grants in the sample, each grant generated an average of 9.55 articles published in peer-reviewed journals. As we can see from Figure 6, the number of publications varied significantly from 0 to 38 while most grants are associated with less than 20 published articles.

- Total Citations

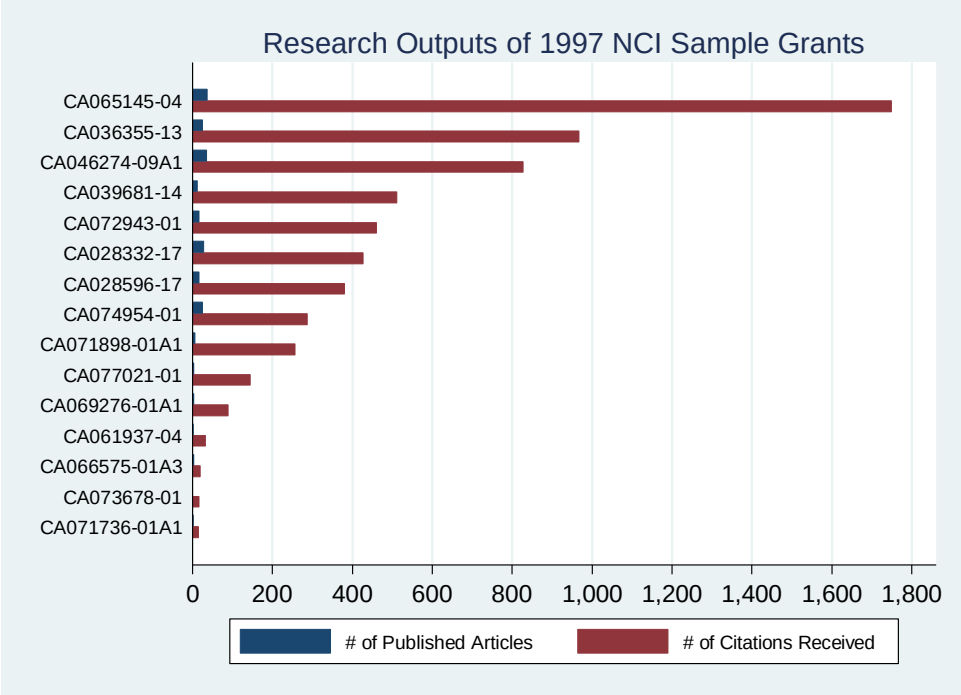
Published articles alone do not signal impact upon the academic community, but the number of citations received is a strong indicator of impact. As can be seen from Figure 7, among the 15 grants issued by NLM in 2002, the number of citations is not closely related to the number of published articles. There are research projects that published less than ten articles yet received more than 100 citations. On the flip side, there are instances when the number of citations is less than the number of publications.

Figure 7. Number of Publications and Citations for NLM Sample Grants issued in 2002



The loose relationship between the number of published articles and the number of citations may be strengthened as time passes since it takes time for some research to attract attention. As we can see from Figure 8, which represents the output for our sample of NCI grants issued in 1997, the number of citations is generally consistent with the number of publications. Besides, the variations in number of citations increase greatly as time spans. Among the 15 grants in the NCI 1997 sample, the number of citations received by published articles ranges from 16 to 1751.

Figure 8. Number of Publications and Citations for NCI Sample Grants issued in 1997



- H-index

The h-index is based on the set of the researcher's most cited papers and the number of citations that they have received in the publications of others. A scholar with an index of h has published h papers each of which has been cited by others at least h times. It is a balanced metric between the number of published articles and the number of citations, reflecting both the number of publications and the number of citations per publication. The index is designed to improve upon simpler measures such as the total number of citations or publications. Since citation conventions differ widely among different fields, the index works properly only for comparing scientists working in the same field (Hirsch 2005). Thus, we present the h-indexes of the sample grants issued by NLM and NCI in Figure 9 and Figure 10, separately.

Figure 9. H-index of 1997 NLM Sample Grants

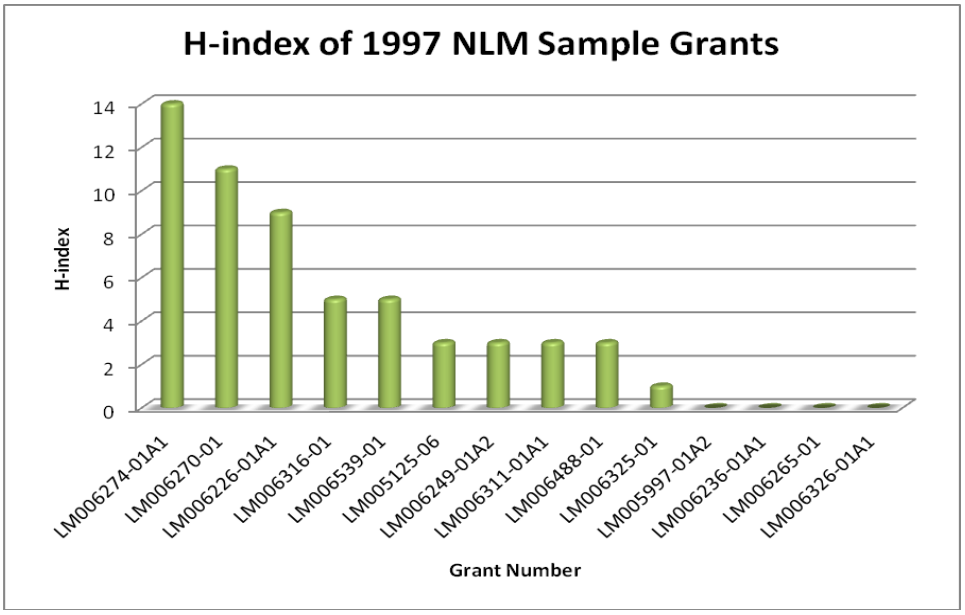
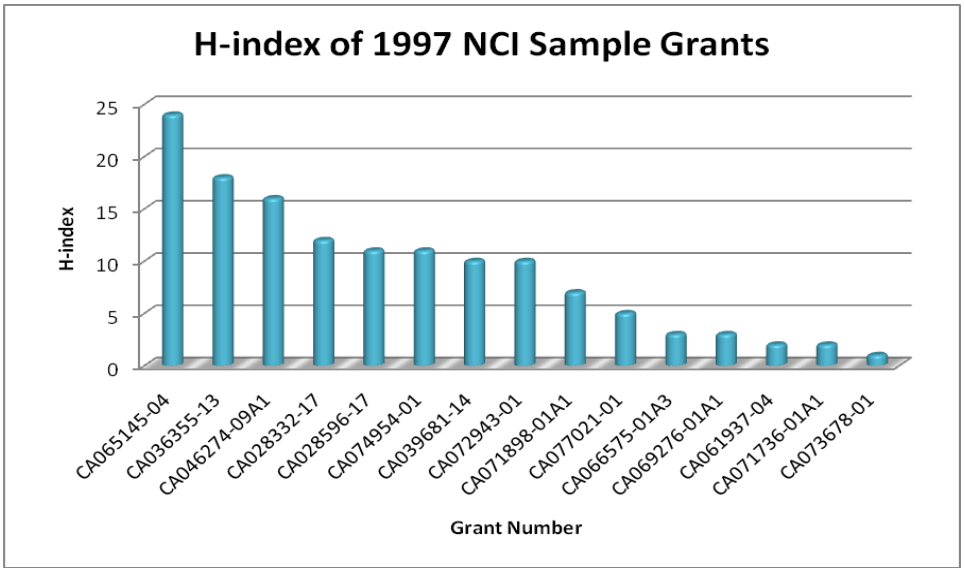


Figure 10. H-index of 1997 NCI Sample Grants



As we can see from Figures 9 and 10, given that all of the grants were issued in 1997, the h-indexes of the NCI grants are much higher than those of the NLM grants. This confirms the argument that h-index works properly only for comparing scientists working in the same field.

- TJPPP Score

Even though the h-index is a balanced metric between the number of publications and the number of citations, it fails to take into account the difference in journals. When the h-index is computed, it does not count publications in highly circulated journals differently than publications in low-circulated journals. We believe that this approach is flawed because, even though they have been cited the same number of times, an article published in a journal with a larger readership has a higher potential impact than that of an article published in a low-ranking journal with lower readership levels. Therefore, we developed a metric using the journal impact factor as a weight. First, we computed a number for each grant, which we call the "citation score":

$$citation\ score = \sum_i journal\ impact\ factor_i \times \#\ of\ citations_i$$

In short, the citation score for grant j is the summation of products of the journal impact factor of article i and the number of citations received by article i.

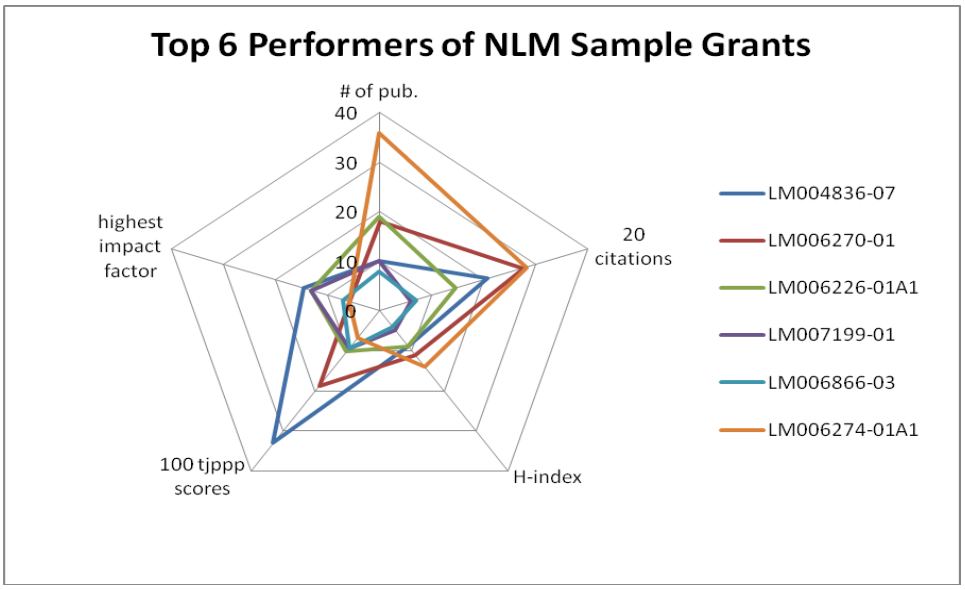
It was apparent that citation scores for grants with few publications that received no citations are zero, which is the same as those grants with no publications at all. To address the differences, we deem it appropriate to add the number of publications to the citation score. In this way, we obtained another metric, which we named the "TJPPP score," in recognition of our program:

$$TJPPP\ score_j = citation\ score_j + \#\ of\ published\ articles_j$$

There are other possible methods to address the differences between grants with no publications and grants with few publications, but which received no citations. For example, we can add the square or the cube of the number of published articles to the citation score in an effort to give more credit to grants with more publications and thus higher potential to be cited in the future. Our approach using the TJPPP score is a preliminary means by which to explore the possible metrics that take the difference in publishing journals into account. So, by combining all of the unscaled metrics, it is apparent that grants' performance is ranked quite differently.

As Figure 11 demonstrates, some grants may be exceptional in terms of the number of publications, while others outperform in terms of number of citations received or TJPPP scores.

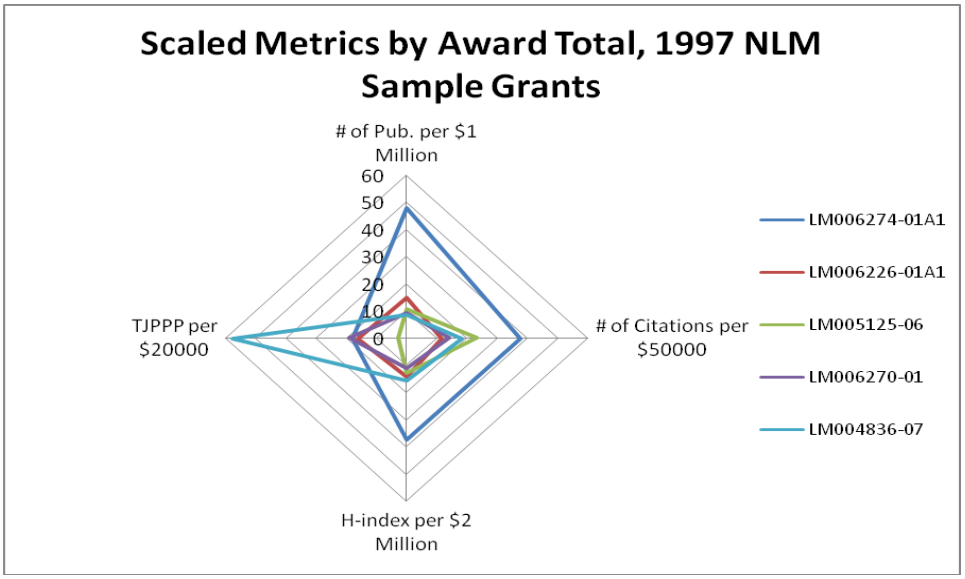
Figure 11. Comparison Among Unscaled Metrics, NLM Sample Grants



c. Scaled Metrics

To account for the differences in size of the grants or the length of research, we can compare the performance of grants with scaled metrics. We can divide their output metrics, such as number of publications, number of citations, h-indexes and TJPPP scores, by their size or length, such as the total amount of award or the number of months since the project's commencement. Also, we can divide the size of grants by the output as another way to compare their performance. It would be interesting to compare the grants by the dollar amounts awarded per published article or the dollar amounts per citation received.

Figure 12. Scaled Metrics by Award Total, 1997 NLM Sample Grants

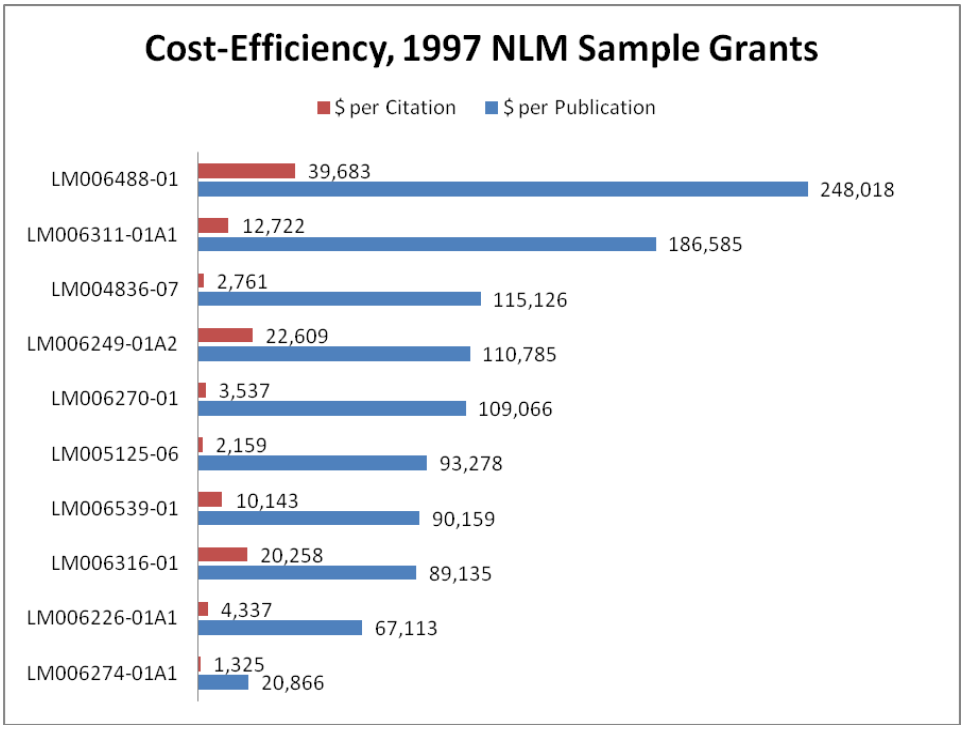


As it pertains to grant outputs scaled by the length of research, it may be inappropriate to compare the grants in our sample with scaled outputs since we only identify the outputs released in or after the granting year (e.g., a project commenced in 1984 and the grant was renewed in 1997). Because the award money was granted in 1997 and lasted, hypothetically, three years, the articles published before 1997 are not associated with the award money and are not counted as outputs of the grant in our analysis. However, if we scale the outputs of grants with length of research, we will underestimate the performance of long-term renewed projects since their previous outputs are not counted by our dataset. To scale the outputs of grants with length of research, we must include outputs of grants from the beginning of the first session of a project to determine whether long-term research outperforms short-term projects.

- Cost-Efficiency

As we can see from Figure 13 below, the cost-efficiency of grants varied significantly. Among the ten grants that published more than one article in the 15 sample grants issued by the NLM in 1997, the most “cost-efficient” grant spent $20,866, on average, for each published article, and $1,325 for each citation received. However, the least “cost-efficient” grant spent $248,018, on average, for each published article, and almost $40,000 for each citation received.

Figure 13. Cost-Efficiency, 1997 NLM Sample Grants



d. Ordered Metrics vs. Comprehensive Metrics

In conclusion, given the variation in emphasis of metrics, it is more appropriate to evaluate research grants based on a series of metrics, unscaled and scaled, rather than simply rely on one or two metrics. Also, it is useful to differentiate ordered metrics and comprehensive metrics.

Using "ordered metrics," grant evaluators may determine different thresholds for different metrics which are utilized in a certain order. For example, we may first examine the number of publications and for those grants that published, say, two or more articles, one can determine how many citations the articles have received for each grant, respectively. If they passed the citation threshold, grants will then be evaluated based on their "TJPPP score" or comparable measures, and the differences in publishing journals will be taken into account.

"Comprehensive metrics" would allow grant evaluators to utilize some integrated metrics which contain most, if not all, of the information provided by different regular metrics. The TJPPP score is a simple version of what is meant by "comprehensive metrics," which

emphasizes the difference in publishing journals, but fails to account for the differences in publishing journals of the citing articles. To evaluate the impact of research more accurately, further improvement could be made by taking the citing articles into account. For example, new metrics could be developed using the PageRank algorithm ("Impact Factor").

iii. Regression Analysis

To investigate what factors determine the performance of grants, we used the TJPPP scores to create the TJPPP category of the NLM sample grants. The dependent variable (TJPPP category) is a categorical variable with four ordered values, which represent the relative performance of each grant. For example, "0" means that a grant's TJPPP score is below the 25th percentile of the sample, (e.g., "poor performance"), and "3" represents a TJPPP score that is above the 75th percentile of the sample, (e.g., "great performance"). A more complete description of the variables can be found in Table 2 below.

Table 2. Descriptive Statistics of Regression Variables, NLM Sample Grants

Variable	Description	Mean	Std. Dev.	Min	Max
TJPPP category	equals 0 if below 25%, equals 1 if between 25% and 50%, equals 2 if between 50% and 75%, equals 3 if above 75%.	1.5	1.2	0	3
total award	in $100,000	5.52	4.52	0.09	19.63
length of research	in years	2.69	0.77	0.33	3.96
age of principal investigator at time of application	in years	46.9	8.3	31	65
grant year dummy variable	equals 0 if grant year is 2002, equals 1 if grant year is 1997	0.5	0.5	0	1

Controlling for grant year, we used an ordered logistic model to regress the TJPPP categorical variable on the explanatory variables, including total award, length of research, and age of principal investigator. As illustrated in Table 3 and Table 4, we then calculated the marginal effects of the three explanatory variables of interest at different award levels and lengths of research.

Table 3. Marginal Effects at Different Award Levels

	TJPPP category=0 poor performance	TJPPP category=1 fair performance	TJPPP category=2 good performance	TJPPP category=3 great performance
award=400,000	-.0918***	.0085	.0472**	.0361***
length=2.5 years	-.2249	.0209	.1155	.0884
age=47	-.0165	.0015	.0085	.0065
award=500,000	-.0820***	-.0105	.0444	.0481
length=2.5 years	-.2009	-.0258	.1088	.1179
age=47	-.0147	-.0019	.0080	.0087
award=600,000	-.0689***	-.0270	.0340**	.0618***
length=2.5 years	-.1686	-.0661	.0833	.1514
age=47	-.0124	-.0049	.0061	.0111

Note: * represents statistical significance at 85% level, ** represents statistical significance at 90% level, *** represents statistical significance at 95% level.

Due to the small sample size, most of the estimates are not statistically significant, (e.g. we cannot assert with great confidence that the marginal effects are not zeroes. However, there is still some useful information that can be gained from our analysis. The marginal effects indicate the effect that each variable has on the probability of each TJPPP category of performance. Among the three independent variables, we can see that the total award is the most influential factor among the three, and age of principal investigator (PI) appears to be completely irrelevant to grant performance.

The higher the total award, the less likely a grant will perform poorly. For example, when the award amount is $400,000, the length of research is 2.5 years, and the age of the principal investigator (PI) is 47, a $100,000 increase in the total award is associated with a decrease of .0918 in the probability of performing poorly. Also, the marginal effect of money decreases as the award total increases. When the award amount is $600,000, a $100,000 increase in the total award is merely associated with a decrease of .0689 in the probability of performing poorly.

Table 4. Marginal Effects at Different Length of Research Levels

	TJPPP category=0 poor performance	TJPPP category=1 fair performance	TJPPP category=2 good performance	TJPPP category=3 great performance
award=550,000	-.0896***	.0033	.0471**	.0392***
length=2 years	-.2195	.0080	.1154	.0960***
age=47	-.0161	.0006	.0085	.0071
award=550,000	-.0586***	-.0353	.0219**	.0720***
length=3 years	-.1436	-.0864	.0536	.1763
age=47	-.0105	-.0063	.0039	.0129
award=550,000	-.0293	-.0380	-.0280**	.0953
length=4 years	-.0716***	-.0931*	-.0686	.2333
age=47	-.0053	-.0068	-.0050	.0171

Note: * represents statistical significance at 85% level, ** represents statistical significance at 90% level, *** represents statistical significance at 95% level.

Even though most of the marginal effects of length of research estimated by our model are not statistically significant, we may conclude from Table 4 that lengthy grants are more likely to outperform grants with shorter periods of duration. For example, when the total award is $550,000, the length of research is two years and the age of PI is 47, one more year of research is associated with an increase of .0960 in the probability of being classified as "great performance."

IV. Qualitative Measures and Case Studies

While quantitative measures are an essential step in evaluating the effectiveness of research grants, they are unable to provide a complete assessment for several reasons. Because medical research is undoubtedly unique and there are countless topics, subjects and iterations which can be studied, the impact of research may not be known for years to come. The dividing line between success and failure is not easily distinguishable in the area of medical research; one person's failure may be another researcher's breakthrough. A research grant which does not accomplish the expected results may uncover a process or methodology, which, in turn, enables another researcher the opportunity to discover a cure for cancer. At first glance, the initial research would be considered a failure, but a retrospective examination of the research may allow the evaluator to realize that the research was, in fact, truly valuable. Because quantitative measures are limited in their ability to completely account for some of the ancillary

effects which make research so important, qualitative analysis becomes an important part of grant evaluation and, thus, should be employed alongside quantitative measures.

Similar to the initial application process, in which peer reviewers evaluate the strength of applications by taking into consideration, among other factors, the research topic, the intended contribution to medical research and the probable success of the project, research grant evaluators, when assessing the efficiency of research grants, should employ qualitative metrics since it has been established that there are limitations to bibliometrics. While increased manpower would be required, the inclusion of qualitative metrics, although somewhat subjective, would provide a more comprehensive analysis.

i. The Use of Qualitative Measures

The National Science Foundation (NSF) undertakes a similar post-grant review process. Using a two-fold system to ensure quality within each of their various sections, the NSF relies heavily on the quality of their application process. Its screening procedures include peer review, with special emphasis placed on the potential contribution to the field and the abilities of those involved in the research. As part of its post-grant review, the NSF, every three years and for each area of study, convenes a Committee of Visitors (COV), which is comprised of leaders in academia who possess the subject specific knowledge necessary to judge the outcomes and process. By examining the merit review process (i.e., application) and the outcomes of the research, the COV aims to determine whether the process is functioning and generating results. Following the true spirit of qualitative analysis, there does not appear to be a formula that is applied annually ("The NSF Merit Review Process").

ii. Case Studies

Although a cursory glance at the quantitative results of the data can provide an indication of strong and weak grant performers, it is, nevertheless, necessary to examine the grants within a qualitative framework since such a framework will ensure a complete assessment of research effectiveness. Although it is acknowledged that such a process can be time-intensive, there is no better measure than an actual hands-on review of the research results.

After analyzing 60 sample grants according to several quantitative metrics, 16 grants were then selected for limited case studies. Primarily consisting of the detailed abstract of the grant, we examined the available information for each grant in an effort to see whether there were any trends that may confirm how the grants compared to our quantitative measures. After careful review of the information, a discernable pattern could not be found. Furthermore, from a qualitative standpoint, there was not a conceivable way to distinguish between the top, bottom and middle performers. Such a finding was largely due to our complete lack of medical training. An expert looking at this information may possibly find valuable information, which may confirm or reject the rankings of the grants from a quantitative standpoint. Although our case study revealed no discernable differences, we feel that qualitative analysis is, nonetheless, a valuable tool that should be employed when attempting to assess the effectiveness of research grants.

Case studies of our sample grants could not be completed due to our lack of medical knowledge. However, in our attempt to assess the impact of NIH grants on medical research and advances in the medical field, we chose to perform case studies of previous grants sponsored by NIH. To this end, we selected two NIH researchers and, based on the reviewed literature, attempted to demonstrate the significant impact of their work on the medical field despite the extended period of time required for research and development.

Dr. K. Frank Austen, who was sponsored by the National Institute for Allergies and Infectious Diseases (NIAID), was recognized in 1999 for his major breakthrough in the study of asthma. Receiving long-term funding from NIAID, Dr. Austen devoted over 35 years of NIH-sponsored research to understanding the molecular cause of asthma and discovered that the body, during an asthma attack, creates leukotrienes, which cause the blood vessels to constrict. Recognizing this phenomenon as a biological cause for the reaction in bronchial asthma, Dr. Austen was able to enlist support from pharmaceutical companies, which manufactured leukotriene inhibitors, a new class of medications used by over 3.5 million asthma sufferers worldwide. A major scientific advancement in this field, this drug was the first novel treatment for asthma since the 1970s. Since a majority of asthma cases are caused by excessive leukotriene production in the body, this drug therapy has proven to be largely effective in treating asthma and has implications for allergic rhinitis, bronchitis, inflammatory bowel

disease, ulcerative colitis, and rheumatoid arthritis. This breakthrough would not have been possible without funding from the NIH (Leibnitz, 1999).

Similar to Dr. Austen, who received long-term funding by NIH, Dr. Leland H. Hartwell received approximately $41 million in NIH funding over a 35-year span. In 2001, Dr. Hartwell was awarded the Noble Prize for Physiology or Medicine for discovering the gene that is responsible for cell division. His major finding has implications for improved understanding of normal cell cycle division as well as mutations in cell division, which can lead to cancer and birth defects. Because his research crossed field boundaries and has the potential to improve multiple diseases, he received support from the National Institute of General Medical Sciences, the NCI and the National Center for Research Resources. Dr. Hartwell is one of 79 American Nobel Prize recipients since 1945 and of those recipients, "60 either worked at or were funded by NIH before winning the prize" (NIH News Release, 2001).

From the aforementioned examples, it is clear that long-term medical research has led to significant advances in understanding and treating various conditions and diseases. What is questioned, however, is whether these advances are worth the costs that are embodied in the many research grants that are awarded each year. Do the benefits to society exceed the costs placed on taxpayers or would congressional appropriations be better spent on other programs which can provide clear quantitative results?

V. Conclusions

i. The Benefits of Medical Research

An examination of the literature exploring the connection between research spending and lower mortality reveals that medical research has resulted in positive outcomes. For example, Murphy and Topel (2003) found that throughout the twentieth century, average life expectancy increased by 30 years and heart disease death rates fell by two-thirds between 1954 and 2000. These advancements, combined with similar progress in developed countries, contributed to the largest gain in health in global history. Using such indicators as quality of life as well as monetary value of life, Murphy and Topel (2003) prove that the costs of medical research are clearly less than societal benefits.

Assuming that medical research equates to an increase in the level of health knowledge, which, in turn, leads to a reduction in death rates, Murphy and Topel (2003) measure the effectiveness of overall medical research spending. Combining economic theory and the life cycle consumption model to create an equation that reflects a spread of values for an additional year of life based on age and gender as well as one's expected utility and willingness to pay for another year of life based solely on health, they aggregated the data according to the population from 1970 to 1998 and calculated a total gain of $37 trillion.

While all Americans certainly consider longevity of life and quality of life to be a precious, unquantifiable gift, Murphy and Topel (2003) estimate that increases in longevity result in an annual gross increase of $2.6 trillion in U.S. wealth and an average annual net gain of $1.6 trillion once spending on health care is taken into consideration. Of the $2.6 trillion, $1.1 trillion is attributable to a decrease in heart disease-related deaths. An additional ten percent decline in heart disease-related deaths, they argue, would lead to a benefit of $5.1 trillion. Also, a decrease in cancer mortality rates would, they contend, result in an economic benefit of approximately $4.4 trillion.

Assuming medical research accounts for only ten percent of the annual net gain, medical research would equate to an annual benefit of $160 billion, which far exceeds the annual expenditure, which, in 1995, was $35 billion. Since US funding levels dedicated to medical research are clearly less than the benefits, taxpayers realize a return on their investment. Furthermore, as incomes rise, health status becomes an even stronger consideration not only to the worker who may lose their income should they become critically ill, but also to the employer, which relies on a healthy workforce in order to increase productivity. These statistics are staggering and appear to suggest that while the investment in medical research may be substantial, its return may be even greater.

ii. Policy Relevance

Accordingly, Congress might want to invest even more funding in biomedical research. Especially now in an era when a large percentage of the national GDP is devoted to rising health care costs, which will likely only increase, in part, because of the large "baby boomer" generation, investment in biomedical research and development may be particularly beneficial

as it can lead to advances in medical technology and equipment and the development of pharmaceutical drugs and vaccines. Biomedical informatics research grants, especially those focusing on integrated health information systems, may be particularly useful in the coordination of care since the growing elderly population tends to see multiple doctors for different conditions and take several medications. Such advancements have the potential to produce substantial gains in quality of life and quality of care. Besides the economic benefits of longevity, other economic benefits can range from efficiency and improved utilization of resources to early detection and the prevention of conditions and, thus, costly treatment (Murphy & Topel, 2003).

iii. Summary

The pivotal role that biomedical research grants serve in medical research and the future of health care are clear, but limited funding, especially in light of our current economic situation, dictates a wise allocation of resources. Consequently, great effort must be taken to evaluate the effectiveness of these grants. Although the metrics discussed in this paper were developed to assess the performance of individual researchers and scientists, they can be applied to research grants since research grants embody the work of the researcher(s). Despite this acknowledgement, it is clear that there are limitations to each of the quantitative metrics discussed. While one metric should not be used to assess grant efficiency, each metric, when employed as a set, can be a useful tool to evaluators as they attempt to assess the efficiency of biomedical research grants. That being said, each metric should not be weighted and subsequently combined to produce a single quantitative score since such a comprehensive figure could result in significant distortions of performance (Butler 2008). Instead, evaluators should separately consider the results within a broader context, namely qualitative criteria. Bibliometrics should enhance, not merely substitute qualitative judgments.

It is apparent that different types of research will dictate different performance benchmarks or possibly a stronger reliance on some indicators more than others. For example, researchers in the biological sciences typically have higher h-indices than those in physics, suggesting a higher rate of publications and/or citations, while clinical journals are typically less cited than other types of journals (Hirsch 2005; Lewison 2002). This distinction is not meant to inhibit the use of quantitative metrics in various biomedical fields or the medical field, in

general, but rather suggests that increased diligence will have to be shown when evaluating grants across fields since such grants are not readily comparable.

VI. Recommendations

Based on our findings, we developed four recommendations for the National Library of Medicine in evaluating funded research grants.

i. Recommendation 1: Require Output From Every Research Grant

The first recommendation is to require output from every grant issued. The most common output for medical research is some form of publication. Without some type of output, it is difficult to draw any connection to a successful outcome in the future. Research with no output also makes it difficult to assess the validity of the research results. We understand the concern of forcing a grant to produce some output, which may take away from the independence of the grant, but there is no reason why the researcher should not be able to at least produce a report of their findings and results to the NLM. Researchers, who have performed work of value, want to share it with their peers; that is, after all, the essence of research. It appears that the only researchers who would not want to share their results are those who view their research as a failure. They need to be reminded that even "unsuccessful" research has significance and may very well lead to a breakthrough elsewhere. Medical researchers need to be reminded that they share a common goal and are, thus, all on the same team.

ii. Recommendation 2: Implement Quantitative Metrics

Our next recommendation is to implement the quantitative metrics outlined in this proposal. "Ordered metrics" may be easier to implement than "comprehensive metrics", the development of which requires profound knowledge of bibliometrics and mathematics. Although quantitative metrics are not a panacea, they do provide valuable information regarding the effectiveness of particular research grants. When these metrics are utilized hand-in-hand with qualitative analysis, they can produce an invaluable evaluation tool.

iii. Recommendation 3: Establish An External Validation Process

The third recommendation is to develop an external validation process for research grants. This process would allow for solid qualitative assessment of each research grant to accompany the quantitative analysis. Conducting these validations approximately three to five years after the grant's conclusion will provide time for the results to begin to take some affect in the field. The objective is not to penalize those that did not achieve much, but to utilize the information to improve the process of selecting quality grants in the future. If trends can be established then it will allow the NLM to avoid certain types of grants which appear more apt to fail. Conducting the review years later may provide enough time for some research results to become apparent. The experts who conduct these reviews would be individuals in the same capacity as those who review research applicants, provided the reviewer is not tasked with assessing a grant which he/she reviewed at the application stage. The above-mentioned process is an invaluable assessment, which would improve the grant process dramatically.

iv. Recommendation 4: Continue In-depth Application Review with Benchmarks

The final recommendation is to include benchmarks within the application review process. Priorities for the NLM and medical research in general, change overtime. Flexibility should be built into the application process to allow for these dynamic priorities. Benchmarks should be established before the application cycle to account for the shift in priorities. An example is the need to sustain the corps of researchers by providing opportunities to new, first time researchers. If the NLM decides that fostering new researchers is a priority, then a benchmark could be set to where at least twenty percent of the recipients for a certain year are selected to become new researchers. In addition, during the application review process for the given year, new applicants would be awarded more points (or less as is the numbering system for NLM grants). This allows the program to ensure diversity in their grants. This method of prioritized process is common amongst college admissions, in job hiring and even military promotion boards. It allows the NLM to have some degree of control in guiding the overall direction of research.

References

About NIH: Mission. (2008). NIH Website. http://www.nih.gov/about/#mission

Bergstrom, C. (2007). Frequently Asked Questions. *Eigenfactor.org.* http://www.eigenfactor.org/faq.htm

Bergstrom, C. (2007). Why Eigenfactor. *Eigenfactor.org.* http://www.eigenfactor.org/whyeigenfactor.htm

Bibliometrics. Wikipedia Org. http://en.wikipedia.org/wiki/Bibliometrics

Butler, L. (2008). Using a Balanced Approach to Bibliometrics: Quantitative Performance Measures In the Australian Research Quality Framework. *Ethics in Science And Environmental Politics, Vol. 8*. doi: 10.3354/esep00077

Delaney, P. (2002). Support for People with Head and Neck Cancer. *FDA*. http://www.spohnc.org/teleconferences/10-02-telecon-slides2.swf

Fact Sheet: The National Library of Medicine. (2007). NIH Website. http://www.nlm.nih.gov/pubs/factsheets/nlm.html

Falagas, M., Papastamataki, P., & Bliziotis, I. (2006). A Bibliometric Analysis of Research Productivity in Parasitology by Different World Regions during a 9-Year Period (1995-2003). *BMC Infectious Diseases*, *6*(1), 56. doi: 10.1186/1471-2334-6-56.

Grants and Funding: Extramural Programs. (2008). NIH Website. http://www.nlm.nih.gov/ep/Grants.html

Grant, J., Cottrell, R., Cluzeau, F., & Fawcett, G. (2000). Evaluating "Payback" On Biomedical Research From Papers Cited In Clinical Guidelines: Applied Bibliometric Study. *BMJ*, *320*(7242), 1107-1111.

Gueorguieva, V. A., Accius, J., Apaza, C., Bennett, L., Brownley, C., Cronin, S., et al. (2008). The Program Assessment Rating Tool and the Government Performance and Results Act: Evaluating Conflicts and Disconnections. *The American Review of Public Administration*, doi: 0275074008319218.

Haeffner-Cavaillon, N., Graillot-Gak, C. & Bréchot, C. (2005). Automated Grading Of Research Performance Clearly Fails To Measure Up. *Nature.* 2005 December 1; 438(7068): 559.

Hirsch, J. E. (2005). An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-16572.

Impact Factor. Wikipedia Org. http://en.wikipedia.org/wiki/Impact_Factor

Leibnitz, R. (1999). Dr. K. Frank Austen Receives Award for Asthma Research. *NIH News*. http://www3.niaid.nih.gov/news/newsreleases/1999/frank.htm

Lewison, G. (2002). Researchers' and Users' Perceptions of the Relative Standing of Biomedical Papers in Different Journals. *Scientometrics, 53(2),* 229-240.

Lewison, G., Lipworth, S., Rippon, I., Roe, P., & Cottrell, R. (2003). Geographical Equity between Outputs of Biomedical Research Grants and Research Capability as an Indicator of the Peer-Review Process for Grant Applications. *Research Evaluation, 12(3),* 225-230.

Lewison, G., Thornicroft, G., Szmukler, G., & Tansella, M. (2007). Fair Assessment of the Merits of Psychiatric Research. *The British Journal of Psychiatry*, *190*(4), 314-318.

Lin, B.W., and Chen, J. (2005). Corporate Technology Portfolios and R&D Performance Measures: A Study of Technology Intensive Firms. *R&D Management*, *35*(2), 157-170.

McCarthy, M. (2007). US scientists press Congress to boost NIH funding. *The Lancet*, *369*(9567), 1071. doi: DOI: 10.1016/S0140-6736(07)60509-1.

Meho, L.I., and Yang, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58(13)*, 2105-2125.

Murphy, K. M., and Topel, R. H. (2003). Measuring the Gains from Medical Research: An Economic Approach, ed. K. M. Murphy and R. H. Topel. Chicago: University of Chicago Press.

NIH News Release. (2001). NIH Website. http://www.nih.gov/news/pr/oct2001/nigms-08.htm

NSF Merit Review Process. (2008). NSF Website. http://www.nsf.gov/bfa/dias/policy/meritreview/

PART Website. http://www.expectmore.gov/

PubMed Website. http://www.ncbi.nlm.nih.gov/sites/entrez/

Scopus Website. http://www.info.scopus.com/journalanalyzer/

The PLoS Medicine Editors. (2006). The Impact Factor Game. *PLoS Medicine, 3*(6), e291. doi: 10.1371/journal/pmed.0030291

Thomson Reuters Website. (2008). Jounal Citation Reports. http://www.thomsonreuters.com/products_services/scientific/Journal_Citation_Reports

Two Australian Universities Select Scopus Database. (2007). *Worldwide Databases*. http://goliath.ecnext.com/coms2/summary_0199-6541253_ITM