Automated Methods to Link Federally Funded Research Projects to Biomarker Development and FDA-Approved Products

Duane E. Williams*†, Sandeep Patel*, Elizabeth Hsu**, James Corrigan**, Joshua D. Schnell*
*Thomson Reuters, Rockville, MD
**Office of Science Planning and Assessment, National Cancer Institute
National Institutes of Health, Bethesda, MD

†Corresponding author's address: 1455 Research Blvd., 2nd Floor, Rockville, MD 20850
Email: duane.williams@thomsonreuters.com

Using public and proprietary databases, we developed and implemented fully automated processes to identify links from biomarkers and drugs Food and Drug Administration (FDA) approved drugs to research projects funded by the National Cancer Institute (NCI) at the National Institutes of Health (NIH). Several data sources were combined to obtain metadata for projects and other outputs. These data were then used in novel ways to characterize the contribution of federal funding to the various outputs. Using portfolios of NCI funded projects, we demonstrate that these methods enable a rapid and objective assessment of the role of research funding in the development of downstream outputs. Specifically, we present (1) a pilot study with a novel approach that traces funding from NCI through various stages of breast cancer biomarker development, and (2) the identification and characterization of links between FDA-approved drugs and NCI-funded grants through patent citations of research literature.

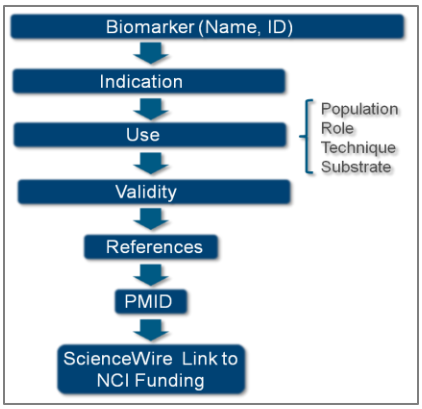

Figure 1. Biomarker Linking to NCI Funding

Biomarkers. This study leverages proprietary Thomson Reuters Integrity℠ data (linking publications to biomarkers) and the Thomson Reuters' ScienceWire® publication catalog (a database correlating MEDLINE® and Web of Science® publication records with federal grants) to link breast cancer biomarkers to NCI grants available in the NIH database, IMPACII. The subset of Thomson Reuters Integrity℠ data used here were obtained by manual review of scientific literature to identify biomarker research publications. These publications were used to identify metadata such as (1) use (e.g., populations studied, substrate, and type of biomarker), (2) indication (e.g., disease condition), and (3) validity (i.e., Thomson Reuters Integrity℠-defined stage of research, ranging from emerging to approved or recommended treatment).

Breast cancer biomarkers were selected for the manageable size of the data set and NCI's interest in breast cancer research. Figure 1 illustrates the data structure and how Thomson Reuters Integrity℠ data were used to link NCI funding through MEDLINE® unique publication identification numbers, PMIDs. PMIDs were used to link publication records to NCI funded grants using the Thomson Reuters ScienceWire® publication

catalog. This pilot study was restricted to publications indexed in MEDLINE® and links to funding were obtained from the acknowledgements sections within each publication.

Our findings show that in this subset of Thomson Reuters Integrity℠ data, 17% of the publications referencing breast cancer biomarkers acknowledged NCI support. The greatest proportion of these publications (24%) have a validity status of "late studies in humans" (Figure 2A). Although 342 of the 890 biomarkers found were not linked to publications acknowledging NCI support, for 220 of the biomarkers we found that all linked publications acknowledged support by NCI. (Figure 2B).

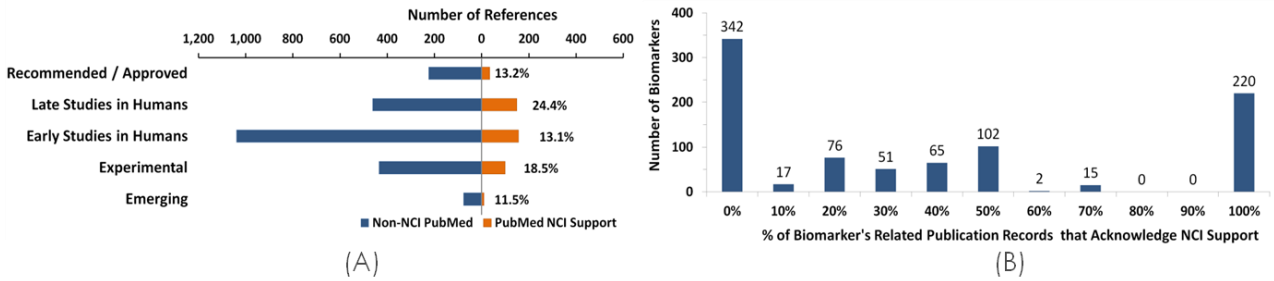

Figure 2. (A) The distribution of NCI-funded and "other" MEDLINE® publications identified as references for breast cancer biomarkers are shown by current validity status (as of March 2011). References are attributed to the highest validity status of a biomarker's use, which may not reflect the validity status of the publication that acknowledges NCI support. **(B)** Histogram of breast cancer biomarkers showing NCI support identified through MEDLINE® publications. Bins of 10 percentile points were created based on NCI-supported publications against the total number of MEDLINE® references attributed to a particular biomarker.

FDA-Approved Drugs. Drug links to NCI funding were made through publicly available patent data in the FDA's Orange Book, also known as the Approved Drug Products with Therapeutic Equivalence Evaluations. These patent records were traced to government funding via their government interest sections, as well as their references to other patents through Thomson Reuters' ScienceWire® patent catalog and other non-patent references. Non-patent reference information was determined by combining data obtained from The Patent Board with the publication catalog (Figure 3).

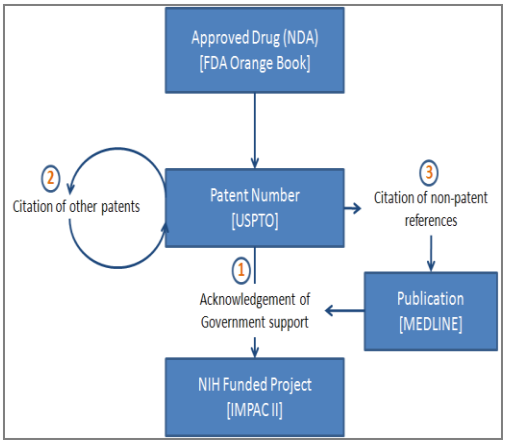

Figure 3. Schematic of Drugs Linked to NCI Funding

Approximately 22% of the 1,102 drugs examined were linked to NCI funding using publication acknowledgments. These links were then classified by the persistence of personnel, organizations and ideas from the grants through drug patents. Persistence of ideas was measured using an automated text similarity scoring algorithm. Our preliminary results suggest that the combination of data sources used here is a useful tool to distinguish the strength of the links between certain drugs and government funding through generational citation searches. This approach identified several strong links that were not found though direct acknowledgement of NCI support in the government interest section of the NDA-listed patent.