Health Research Nonformula Grants - State Fiscal Year 2014-15

Health research nonformula grants totaling $14,201,440 were awarded to three organizations in response to the Request for Application (RFA) # 67-32 for big data in health research. All research projects addressed the following research priority, which was established by the Department in conjunction with the Health Research Advisory Committee:

> For the purpose of priority setting, the Health Research Advisory Committee recommends combining the two nonformula funding categories of clinical and health services research and other research. At least 50 percent of the funds must be spent on clinical research or health services research or both clinical research and health services research. The research priority for nonformula-funded research is:
>
> Big Data in Health Research
>
> The priority is research to develop methods, software, and other technologies designed to analyze vast data sets at the level of molecules, proteins, organelles, cells, tissues, organs, physiological systems, organisms, populations, health care systems, and ecosystems.
>
> With the rapid expansion of high-throughput laboratory technologies and electronically integrated health care systems, biomedical researchers have access to more and more complex data than ever before. Vast data sets exist at the level of molecule, protein, organelle, cell, tissue, organ, physiological system, organism, population, health care system, and ecosystem. Health care systems now electronically record an ever-increasing volume and variety of variables from patient monitoring systems, imaging, and "omics" technologies as well as data in electronic health records (EHRs). The major challenge now is to manage these large and growing data sets and discover within them insights that can guide future research, education, and clinical care.
>
> Pennsylvania institutions are currently well represented in major national health research initiatives focused on big data, including the National Institutes of Health (NIH) Big Data to Knowledge (BD2K) and the Patient-Centered Outcomes Research Institute (PCORI) Clinical Data Research Networks (CDRN). The Commonwealth is well positioned to take a leadership role in advancing the field and exploiting its expertise and resources for significant advances in biomedical research and health care. The goal moving forward is to embrace the size and complexity of available data through the transfer, merging, storage, visualization, and processing of disparate data and the use of new algorithms and software tools to computationally discover predictive and causal relationships.
>
> Research activities should lead to improved design, implementation, and utilization of data systems, both research and clinical, to enhance the exploitation of data from the molecular to the population level to improve the health and well-being of Pennsylvanians.
>
> Funding for the big data priority must be spent on biomedical research or clinical research or health services research or any combination of these types of research as defined in Act 2001-77. Activities that are not biomedical, clinical, or health services research as defined by Act 2001-77 will not be considered.

Research may include, but is not limited to, the following areas:

- Research to develop algorithms and software/application programming interfaces to discover causality in big data from multiple sources. Some examples of driving biomedical projects on which to focus these efforts include cancer metastases, neurodegenerative disorders, and autoimmune diseases.
- Research that integrates some or all of the following data: uni- or multi-modal, multiplatform, or multiscale data across diverse data types and sources (for example, omics, imaging, laboratory, clinical, socioeconomic, environmental, social media, wearable or mobile devices, company loyalty programs, self-reported data) to create single or multiscale models of molecular, cellular, organ, system, organism, or population processes or behaviors. Some examples of driving biomedical projects on which to focus these efforts include obesity, chronic obstructive pulmonary disease, non-alcoholic fatty liver disease, and infectious disease outbreaks/epidemics.
- Research focused on the integration and mining of the EHR systems of multiple health care providers and systems across the Commonwealth for comparative effectiveness research or modeling. Some examples of driving clinical projects on which to focus these efforts include atrial fibrillation, diabetes mellitus and otitis media.
- Research to develop and test in community-based health care systems analytics software to identify linkages at the patient level (for example, clinical flags raised through comparison with hundreds to millions of other patients in the database with similar data) and at the system level (for example, quality assurance flags raised when trends are detected). Some examples of driving clinical projects include post-surgical complications, preventive screening, rare diseases, and pharmacogenomic testing indicators.
- Clinical research or health services research or both types of research that integrate multiple data sources through cloud computing in a way that addresses issues related to security, confidentiality, and consent.
- Research to model statewide health behaviors, trends, and needs prediction through the merging of multiple large data sets, including the integration of statewide datasets (for example, Pennsylvania Health Care Cost Containment Council [PHC4], Pennsylvania Cancer Registry, Epidemiologic Query and Mapping System [EpiQMS], Behavioral Risk Factor Surveillance System [BRFSS], Pennsylvania Statewide Immunization Information System [PA-SIIS]).

Research in the following areas will not be considered:

- Focus on enhancing computing infrastructure
- Focus on data collection/generation or the development of technology to generate data
- Secondary statistical or epidemiological analysis of single large data sets (or multiple comparable data sets of the same type of data)
- Narrow focus on a single disease without demonstration of generalizability of the selected approach to health research or clinical care more broadly
- Genome-wide association studies or other research focused on identifying disease risk or cause rather than on developing methods for using large omic or other data sets to identify disease risk or cause
- Design and development of registries, tissue banks, and other health data systems

The research should hold the potential for addressing the health needs of underserved segments of the population, including rural, urban, racial/ethnic minorities, older adults, or other high-risk constituencies in the Commonwealth. To foster cross-institutional collaborative research among organizations across the Commonwealth, an applicant must conduct research in collaboration with other research institutions and organizations. Collaboration is encouraged between academic institutions, health care systems, health care insurers, public health agencies or businesses or any combination of these organizations. To the extent possible, organizations that are not academic medical centers, such as smaller colleges and universities and local public health agencies, should be included in addition to major research institutions. Collaboration with a minority-serving academic institution or a minority-serving community-based organization in Pennsylvania is strongly encouraged and should include the mentoring and training of students. All research collaborators must play a substantive and meaningful role in multiple aspects of the proposed research. Research proposals must include clear objectives and targeted outcomes. No more than 50% of the funds may be used for research infrastructure as defined in the Act, as amended (for example, equipment, supplies, nonprofessional personnel, and laboratory or building construction or renovation).

The following list of grant awards provides the lead and collaborating institutions, title of the research project, amount of the grant award, grant award period, contact principal investigator, co-principal investigators, project purpose, project overview and expected research benefits and outcomes.

Big Data Research Projects

- The Geisinger Clinic, Pennsylvania State University and University of Pennsylvania - *Integrating Big Data for Biomedical Discovery: Methods, Tools, and Applications*, $4,385,863 for a 48-month project (June 1, 2015 — May 31, 2019)

  Contact Principal Investigator:
  Marylyn D. Ritchie, PhD
  Director, Biomedical and Translational Informatics
  Geisinger Clinic
  100 North Academy Avenue
  Danville, PA 17822
  Telephone: 570-214-7579
  Email: mdritchie@geisinger.edu

  Other Key Researchers:
  Geisinger Clinic - David J. Carey, PhD; Gregory J. Moore, MD, PhD; H. Lester Kirchner, PhD; Nicholas Marko, MD; Sarah A. Pendergrass, PhD; Joseph Leader, BA; Andrew Michael, PhD; Christopher Still, DO; Radhika Gogoi, MD; Vishal Mehra, MD; Brian Schwartz, MD; Brandon Fornwalt, MD
  The Pennsylvania State University - Kateryna Makova, PhD; Francesca Chiaromonte, PhD
  University of Pennsylvania - Jason H. Moore, PhD

  Type of Research: Clinical and Health Services

  Project Purpose: The purpose of this research is (1) to develop and validate advanced methodologies for elucidating features and patterns in clinical data for integration with genomic and environmental data to explore the genetic architecture of complex traits; (2) to develop a strategy for Phenome-Wide Association Studies using low frequency genetic variants and the full spectrum of available clinical, environmental, and behavioral data; and (3) to develop and apply data integration methods for maximizing signal to noise in complex analyses and to facilitate interpretation of results.

Project Overview: We propose a multidisciplinary project to develop a series of advanced algorithms, methodologies, and software for integrating and analyzing multiple types of biomedical Big Data and to apply these innovative approaches for better understanding and treatment of obesity and obesity-related comorbidities.

Specific Aim 1 - Develop and validate advanced methodologies for elucidating features and patterns in clinical data for integration with genomic and environmental data to explore the genetic architecture of complex traits. Using machine learning to mine the rich phenotypic data of the EHR, we will create homogeneous subsets with more consistent underlying genetics and other contributing factors affecting individual phenotypic data to uncover important biological insights and identify useful biomarkers.

Specific Aim 2 - Develop a strategy for Phenome-Wide Association Studies (PheWAS) using low frequency genetic variants and the full spectrum of available clinical, environmental, and behavioral data. We will develop comprehensive algorithms to apply to low-frequency genetic variant data, in addition to the common variant data, and to integrate the genetic data with environmental, behavioral, and EHR data for PheWAS analyses that will enable new discovery of relationships between genetic architecture and outcome for obesity and related comorbidities.

Specific Aim 3 - Develop and apply data integration methods for maximizing signal to noise in complex analyses and to facilitate interpretation of results. By integrating public domain data and exploring unusual features of the genome and phenome we will maximize our ability to detect important genetic, clinical, behavioral, and environmental effects. We will integrate existing and novel approaches to maximize important factors for complex trait susceptibility.

Training Aim - Provide internship opportunities in bioinformatics for 12 minority students and develop a two-year didactic-research opportunity for post-baccalaureate minority students. The interdisciplinary nature of research in Big Data analysis for biomedical discovery has created enormous educational opportunities. We will work with our collaborative partners to foster an increase in minority student training in this emerging field.

Expected Research Outcomes and Benefits:
(1) New algorithms, methodologies, and software for integrating and analyzing multiple types of biomedical Big Data.
(2) Improved understanding and treatment of obesity and related comorbidities.
(3) Availability to the larger research community of the innovative and novel methodological, algorithmic, and software tools developed by this project to study other diseases and conditions and thereby improve the understanding and treatment of those diseases and conditions.
(4) New approaches to assess the risk and progression of obesity associated non-alcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH) that combine clinical, imaging, laboratory, genomic and other data into predictive algorithms and improve diagnosis and therapies. NAFLD affects 30% of US adults and increases risk for type 2 diabetes and NASH is a leading cause of cirrhosis.

(5) Better understanding of the relationship of obesity and/or weight loss to degenerative joint disease through integrated analyses of genomic, environmental, behavioral, and clinical data.
(6) New insights into the mechanisms underlying the effects of obesity on endometrial cancer progression, recurrence, and overall survival. Obesity is the most significant risk factor for endometrial cancer, and understanding these mechanisms can lead to new treatment options.
(7) Improved understanding of the relationship between obesity and coronary heart disease grounding new methods of diagnosis and treatment
(8) Increased number of minority students trained in Big Data analysis for biomedical discovery.

- The University of Pennsylvania, Carnegie Mellon University and Temple University - *Smarter Big Data for a Healthy Pennsylvania: Changing the Paradigm of Healthcare*, $4,772,786 for a 48-month project (June 1, 2015 — May 31, 2019)

  Contact Principal Investigator:
  Daniel Polsky, PhD
  Executive Director, Leonard Davis Institute of Health Economics, Professor of Medicine and Health Care Management
  University of Pennsylvania
  Colonial Penn Center
  3641 Locust Walk
  Philadelphia, PA 19104-6218
  Telephone: 215-573-5752
  Email: polsky@wharton.upenn.edu

  Other Key Researchers:
  University of Pennsylvania – Kevin Volpp, MD, PhD; Charles Branas, PhD; Michael Draugelis, BS; Peter Groeneveld, MD, MS; C. William Hanson, MD; John Holmes, PhD; Raina Merchant, MD, MSHP; Katherine Milkman, PhD; Amol Navathe, MD, PhD; Mitesh Patel, MD, MBA, MS; Dylan Small, PhD; Lyle Ungar, PhD
  Carnegie Mellon University - George Loewenstein, PhD
  Stony Brook University (SUNY) – H. Andrew Schwartz, PhD
  Temple University - Aaron A. Sorenson, MA; Mark Weiner, MD

  Type of Research: Health services

  Project Purpose: The overarching goal of this project is to improve the health of Pennsylvanians at an individual, community, and population level by changing the paradigm of medical care and health care delivery to predicting and preventing onset, exacerbation, and advancement of disease rather than principally reacting to clinical events as they happen. Using medical record data with linkages to administrative claims, wearable monitor data, and social media data, we will develop algorithms to better predict clinical events in the hospital, at home, and in the community. The proposed project will expand an established and highly successful minority health services research training program to provide opportunities for training in big-data research to support the career development of under-represented minorities within Pennsylvania.

  Project Overview: With the recent explosion in the volume and types of data collected in medicine, there is the opportunity for a paradigm shift from a health care system that reacts to disease exacerbations as they happen to one that is proactive in targeting disease and clinical events before they occur. Our goal is to make new data assets smart and actionable by predicting and preventing clinical events in hospitals, in homes, and in communities. The aims proposed in this application will impact the health of Pennsylvanians through smarter uses of big data to predict clinical events earlier than ever before. The aims may have their greatest impact on addressing health disparities because we will move prediction beyond the walls of the hospital and into homes and communities where the most vulnerable face the greatest challenges. Through harnessing big data for more precise and timely prediction on actionable events, this application fills a critical gap in our health care system in its efforts to improve quality and to ameliorate health disparities.

AIM 1 – In-hospital: Algorithm development for the dynamic and timely prediction of in-hospital post-surgical complications using multi-modal clinical data.
AIM 2 – At home: Algorithm development for the dynamic and timely prediction of out-of-hospital risk for 30-day and 90-day readmission using a multimodal, integrated dataset from insurer and pharmacy claims, electronic health records, and mobile devices.
AIM 3 – In Community: Integrate social media data with statewide data and use these data to build and validate a tool for monitoring and predicting the health of Pennsylvanians with respect to high morbidity health conditions and real-time dynamic health events.
AIM 4 - Training minority students in research: Expand established and highly successful minority health services research training program to provide opportunities for training in big-data research and to expand program to include undergraduate scholars from Lincoln University.

To achieve our Aims we have created a unique set of collaborations between the University of Pennsylvania and Penn Medicine, Temple University, Carnegie Mellon University, Independence Blue Cross, and CVS Health that will also involve Apple and Intel and data from several other private insurers. We will use state-of-the-art methods to integrate large static datasets (claims, census, surveys), structured and streaming clinical data (EHR, Operating Room system data, Intensive Care Unit monitoring data), unstructured clinical data, data from mobile devices, and social media data to create models to improve health and reduce disparities that could be used throughout Pennsylvania.

Expected Research Outcomes and Benefits: The research outcomes will be prediction algorithms that allow for real-time identification of changes in health for patients at highest risk in the hospital, at home, and in the community. In-hospital prediction algorithms will focus on complications from common surgeries using data streams, primarily from electronic health record data including lab and radiology data and the ongoing monitoring of vital signs during the hospitalization. These predictions will be modeled within the Penn Medicine health system and tested in the Temple health system to examine the generalizability of the approach. To reduce the risk of rehospitalization, we will develop at home prediction algorithms using a combination of clinical data from the index hospitalization, clinical and claims data from both before and after the index hospitalization, data on social determinants of health, and wearable device data to identify changes in the risk of rehospitalization. Community-level prediction algorithms will be developed using social media data from Twitter that can recognize in real time dynamic shifts in the level of risk of, for example, infectious diseases and violence, within a community. We will also train up to 17 minority undergraduates in big data research. The health benefits for Pennsylvanians could be significant as our algorithms are focused on predicting common and high risk medical events, many of which are most prevalent among Pennsylvania's most vulnerable populations. By creating tools to predict changes in the risk of medical events in real time, we aim to help facilitate transformation of health care delivery from a paradigm that reacts to clinical events to one in which clinical events can be anticipated and thereby avoided.

- The University of Pittsburgh and Carnegie Mellon University - *Big Data for Better Health (BD4BH) in Pennsylvania,* $5,042,791 for a 36-month project (June 1, 2015 — May 31, 2018)

  Contact Principal Investigator:
  Gregory F. Cooper, MD, PhD
  Vice Chair, Dept. of Biomedical Informatics
  University of Pittsburgh
  5607 Baum Blvd, Fifth Floor, Rm 524
  Pittsburgh, PA 15206
  Telephone: 412-624-3308
  Email: gfc@pitt.edu

  Other Key Researchers:
  University of Pittsburgh - Uma Chandran, PhD, MSIS; James G. Herman, MD; Rebecca S. Jacobson, MD, MS; Xia Jiang, PhD; Adrian V. Lee, PhD; Xinghua Lu, MD, MS, PhD; Steffi Oesterreich, PhD; Liza C. Villaruz, MD
  Carnegie Mellon University - Ziv Bar-Joseph, PhD; Clark Glymour, PhD; Seyoung Kim, PhD; Carleton Kingsford, PhD; Chris J. Langmead, PhD; Robert F. Murphy, PhD; Russell S. Schwartz, PhD; Eric P Xing, PhD; Nicholas A. Nystrom, PhD (Pittsburgh Supercomputing Center)

  Type of Research: Clinical

  Project Purpose: The purpose of BD4BH is to develop the methods and software needed to integrate large amounts of multiple types of molecular, clinical, and demographic data to construct features that are more informative than the data themselves and to use both data and features in machine learning models to predict clinically important outcomes. We will use as the platform for developing these tools the prediction of progression, recurrence, metastasis, and overall survival in patients with breast and lung cancer. We will also evaluate whether disparities in these outcomes exist due to age, race/ethnicity, and/or residence location in cancer patient cohorts from Pennsylvania.

  Project Overview: Rapidly increasing volumes of molecular and electronic health record data in health care hold great promise for predicting patient outcomes, personalizing care, reducing geo-demographic disparities, and improving health. To realize this potential, however, fundamental research is needed regarding how to wrangle complex clinical and molecular Big Data into a form that is ready for analysis; how to select the most informative features from each type of data; how to combine such datasets to construct informative network and pathway features; and how to apply machine-learning methods to constructed features and raw data to derive models that accurately predict clinical outcomes. Our specific aims are to:

  1. Develop methods for merging, managing, utilizing, processing, analyzing, and sharing large amounts of diverse types of data. For electronic health record data, we will develop automated methods for extracting and representing patient-level data, including clinical text, about the course of disease. For large-scale molecular data, including sequencing, expression, and epigenetic data, we will develop methods for fast and accurate

processing and alignment, mutation calling, quantification, and evolutionary reconstruction of molecular cancer progression.

2. Develop and apply machine learning methods to large complex datasets to predict cancer outcomes based on constructed features that are biologically meaningful.
2a. We will develop, extend, and modify machine learning methods for selecting and constructing complex features by discovering and modeling cancer mechanisms in tumor cells using molecular data.
2b. We will evaluate the extent to which such constructed features (2a), when combined with rich clinical data, improve the prediction of cancer recurrence, metastasis, rate of progression, and length of life for breast and lung cancer, which have higher incidence rates in Western Pennsylvania than the national average.
2c. We will further develop computational methods to investigate biological and geo-economic factors contributing to health disparities in disease risk and outcome related to age, race/ethnicity, and geography.

3. Train underrepresented minority students in the analysis of Big Data. We will partner with Lincoln University to engage their students and faculty in research training programs at Pitt and CMU.

Expected Research Outcomes and Benefits: Our software tools will allow both healthcare and research institutions to manage and exploit large, complex sets of molecular and electronic health record data for predicting patient outcomes, personalizing clinical care, reducing geo-demographic disparities, and improving health.

Our methods for extracting text-based and temporal data from electronic health records will be essential for our predictive modeling and will also useful as stand-alone tools for health services research and health system analytics. Other groups will be able to apply our methods for processing and analyzing molecular data (including RNA-Seq quantification, determining isoform expression, and unmixing) for analysis of data from other diseases. Finally, our feature selection, construction, and utilization methods (both for regression and classification), while applied specifically to breast and lung cancer in this project, are general and would be valuable tools for researchers studying other types of cancers and, more broadly, for investigators utilizing advanced genomic and text resources for translational research.

Our methods will be published, and all software developed will be open-source and therefore benefit the larger research and health care communities in Pennsylvania and beyond.