OMB No. 0925-0001 and 0925-0002 (Rev. 09/17 Approved Through 03/31/2020)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Gregory F. Cooper

eRA COMMONS USER NAME (credential, e.g., agency login): gfc@cbmi.pitt.edu

POSITION TITLE: Professor of Biomedical Informatics

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)

| INSTITUTION AND LOCATION | DEGREE(if applicable) | Completion DateMM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
| M.I.T., Cambridge, MassachusettsStanford University, Palo Alto, CaliforniaStanford University, Palo Alto, California | B.S.Ph.D.M.D. | 197719851986 | Computer ScienceMedical Information SciencesMedicine |

**A. Personal Statement**

Dr. Gregory Cooper is Professor of Biomedical Informatics, of Computational and Systems Biology, and of Intelligent Systems at the University of Pittsburgh. He was one of the initial developers of causal Bayesian networks (CBNs). His doctoral dissertation in 1985 developed and evaluated a computer-based medical diagnostic aid that integrated causal and probabilistic knowledge. It introduced methods for representing disease knowledge using CBNs and for using that representation in performing and explaining diagnostic inference given a set of patient findings.

He was also a pioneer in the development of machine learning methods for constructing CBNs from data, including Bayesian methods for learning CBNs from combinations of observational data, experimental data, and prior knowledge. A paper he co-authored with Dr. Edward Herskovits on learning CBNs from data has been cited more than 5,000 times, according to Google Scholar. He also wrote early papers on Bayesian methods for learning CBNs when there are hidden variables, and when there is selection bias. He did early work on active learning of causal models, including learning gene expression networks from DNA microarrays. He co-edited a book with Dr. Clark Glymour on causal modeling and discovery that is titled “Computation, Causation, and Discovery.” He continues to actively pursue basic and translational research in graphical causal modeling and discovery.

In general, his current research involves the development and application of new machine learning, artificial intelligence, and Bayesian statistical methods to address biomedical informatics research problems. He is involved in projects on causal discovery from observational biomedical data, infectious disease outbreak detection and characterization, personalized cancer diagnosis and outcome prediction, clinical alerting based on machine learning from an electronic-medical-record (EMR) archive, and an EMR system that learns to highlight the most useful patient information. He has over 180 peer-reviewed research papers related to these areas of study.

**B. Positions and Honors**

Professional Experience

|  |  |
| --- | --- |
| 1986 - 1987 | Postdoctoral Fellow, Medical Informatics, Stanford University |
| 1987 - 1989 | Research Associate, Medical Informatics, Stanford University |
| 1989 - 1990 | Senior Research Associate, Medical Informatics, Stanford University |
| 1990 - 1996 | Assistant Professor of Medicine, University of Pittsburgh |
| 1990 - | Secondary faculty appointment in the Intelligent Systems Program, University of Pittsburgh |
| 1997 - 2006 | Associate Professor of Medicine, University of Pittsburgh |
| 2006 - 20122006 - | Associate Professor of Biomedical Informatics, University of PittsburghSecondary faculty appointment in Computational and Systems Biology, Un. of Pittsburgh |
| 2006 -  | Vice Chair, Department of Biomedical Informatics, University of Pittsburgh |
| 2012 -2019 - | Professor of Biomedical Informatics (primary appointment), University of PittsburghSecondary faculty appointment in Professor of Pharmaceutical Sciences |
| 2019 - | Distinguished Professor of Biomedical Informatics, University of Pittsburgh |

**Honors**

|  |  |
| --- | --- |
| 1975 | Tau Beta Pi, M.I.T. |
| 1977 - 1984 | Medical Scientist Training Program Trainee, Stanford University, Stanford, CA |
| 1985 | Martin Epstein award for best paper in the student paper competition at the Ninth Annual Symposium on Computer Applications in Medical Care |
| 1991 | Elected as a Fellow of the American College of Medical Informatics |
| 1999 | General Chair, Conference on Uncertainty in Artificial Intelligence |
| 2005 | Distinguished paper award, Symposium of the American Medical Informatics Association |
| 2006 | Elected as a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI) |
| 2007 - 2008 | Executive board committee member, American College of Medical Informatics |
| 2010 | Homer R. Warner research award at the 2010 Annual Symposium of the American Medical Informatics Association for a co-authored paper  |
| 2011 | Marco Ramoni distinguished paper award at the 2011 AMIA Summit on Translational Bioinformatics for a co-authored paper |
| 2009 – 2013 | Biomedical Library and Informatics Review Committee (BLIRC study section), National Library of Medicine, National Institutes of Health (chair 2011-2013) |
| 2018 | UPMC Endowed Chair, University of Pittsburgh |

**C. Contributions to Science**

1. **Bayesian network modeling, inference, learning, and explanation**. Bayesian networks (BNs) are now used commonly in biomedical informatics and many other fields. Dr. Cooper was one of the initial developers of BNs. His doctoral dissertation introduced a version of BN modeling and inference and applied it to develop a medical diagnostic system (1a). He analyzed the computational complexity of BN inference and showed it to be NP-hard (1b). He also helped develop several early exact and approximate BN-inference algorithms. He co-developed a Bayesian method for learning BNs from data and prior knowledge (1c), which has been widely cited and applied during the past 25 years. He also helped develop and evaluate a method for explaining BN inference to users (1d).
2. Cooper GF. *NESTOR: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge* (doctoral dissertation, Stanford University, 1984). [cited over 270 times, according to Google Scholar]
3. Cooper GF. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42 (1990) 393–405. [cited over 2700 times]
4. Cooper GF, Herskovits EH. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 (1992) 309–347. [cited over 5000 times]
5. Suermondt HJ, Cooper GF. An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research* 26 (1993) 242-254. [cited over 50 times]
6. **Causal modeling and discovery**. Much of science consists of discovering and modeling causal relationships in nature. Dr. Cooper was one of the early developers of machine learning methods for learning causal BNs from data, including Bayesian methods for learning CBNs from combinations of observational data, experimental data, and prior knowledge (2a). He also wrote early papers on Bayesian methods for learning CBNs when there are hidden variables and when there is selection bias. In recent years, he has helped develop and apply machine learning methods that use genomics data to discover the causal effects of tumor-specific genomic alterations (2b) and to predict patient response to cancer therapies (2c). He is the Director of the Center for Causal Discovery (2d) which is making state-of-the-art causal discovery algorithms readily available and easy to use by biomedical scientists.
7. Glymour C, Cooper GF. (Eds.). *Computation, Causation, and Discovery* (MIT Press, Cambridge, MA, 1999). Note: This is an edited book and was not peer reviewed. [cited over 440 times]
8. Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, Chen X, Xue Y, Lee AV, Clark N, Chen V, Lu S, Chen L, Yu L, Hochheiser HS, Jiang X, Wang QJ, Lu X. Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLOS Computational Biology,* 15 (2019). PMID: 31276486, PMCID: PMC6650088
9. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*, 16 (2018) 269-278. PMID: 29133589, PMCID: PMC5821274.
10. Cooper GF, Bahar I, Becich MJ, Benos PV, Berg J, Espino JU, Jacobson RC, Kienholz M, Lee AV, Lu X, Scheines R, Center for Causal Discovery team. The Center for Causal Discovery of biomedical knowledge from big data. *Journal of the American Medical Informatics Association,* 22 (2015) 1132-1136. PMID: 26138794, PMCID: [PMC5009908](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009908/).
11. **Clinical outcome prediction**. Accurate prediction of outcomes is an important component of clinical care. The increasing availability of clinical data provides greater opportunities for using machine learning to develop accurate prediction methods. Dr. Cooper and his colleagues have made several contributions to clinical prediction research. He has led efforts to compare the clinical predictive performance of a wide range of machine-learning methods (3a); a significant outcome of that work is the demonstration that even modest improvements in predictive performance are projected to sometimes have a major impact on the quality and cost of healthcare for diseases that are common. Dr. Cooper has also investigated novel approaches to machine-learning-based outcome prediction, including highly efficient Bayesian model averaging (3b), methods for calibrating probabilistic predictions (3c), and a framework for learning personalized clinical models (3d).
12. Cooper GF, Abraham V, Aliferis CF, Aronis JM, Buchanan BG, Caruana R, Fine MJ, Janosky JE, Livingston G, Mitchell T, Monti S, Spirtes P. Predicting dire outcomes of patients with community acquired pneumonia. *Journal of Biomedical Informatics* 38(2005) 347-366. PMID: 16198995
13. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer’s disease from genome-wide data. *Journal of the American Medical Informatics Association* 18 (2011) 370-375. PMID: 21672907 PMCID: PMC3128400.
14. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. In: *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence* (2015) 2901-2907. PMID: 25927013 PMC4410090.
15. Visweswaran S, Ferreira A, Cooper GF. Personalized modeling for prediction with decision-path models. *PLoS One* 10 (June 22, 2015) e0131022. PMID: 26098570 PMCID: PMC4476684
16. **Biosurveillance.** Early and accurate disease outbreak detection and characterization are important public health activities. During the past 16 years, Dr. Cooper has contributed to several innovative approaches to biosurveillance. One approach works by building a mathematical bridge between epidemiological modeling and patient diagnosis (4a, 4d). The probability distribution over individual patient diseases influences belief about the disease outbreaks in the population. Conversely, the probability distribution over the disease outbreaks in the population influence the probabilities of individual patient diseases. He also helped develop methods for detecting outbreaks due to aerosol release of disease agents, such as anthrax spores (4b), and outbreak-detection methods that augment human disease detection (4c).
17. Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM. Bayesian biosurveillance of disease outbreaks. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (2004) 94-103.
18. Hogan W, Cooper GF, Wallstrom G, Wagner MW, Depinay JM. The Bayesian aerosol release detector: An algorithm for detecting and characterizing outbreaks caused by an atmospheric release of *Bacillus* *Anthracis*. *Statistics in Medicine* 26 (2007) 5225-5252. PMID: 17948918
19. Shen Y, Adamou C, Dowling JN, Cooper GF. Estimating the joint disease outbreak-detection time when an automated biosurveillance system is augmenting traditional clinical case finding. *Journal of Biomedical Informatics* 41 (2008) 224-231. PMID: 18194876. Publication not directly supported by NIH.
20. Cooper GF, Villamarin R, Tsui FC, Millett N, Espino J, Wagner MM. A method for detecting and characterizing outbreaks of infectious disease from clinical reports. *Journal of Biomedical Informatics* (2014). pii: S1532-0464(14)00192-0. doi: 10.1016/j.jbi.2014.08.011. [Epub ahead of print] PMID:25181466 PMC4441330

For complete list of published, peer-reviewed papers, see: <http://www.dbmi.pitt.edu/pubs-with-pdfs/26>

**D. Additional Information: Research Support and/or Scholastic Performance**

ACTIVE

**R01LM012011 (Lu)**  04/01/2015 – 03/31/2024

NIH/NLM

Interpretable deep learning models for translational research

This project aims to develop advanced machine learning methods, referred to as deep learning models, to simulate cellular signaling systems, at both multiple cell and single cell levels. Success of these models will enable researchers to investigate cellular behaviors under physiological and pathological conditions, and such information can be used to guide therapy of cancer patients.

**IIS-1636786 (Patel)** 01/01/2017-12/31/2020

NSF

BD Spokes: SPOKE: NORTHEAST: Collaborative Research: Integration of Environmental Factors and Causal Reasoning Approaches for Large‐Scale Observational Research

The current Health Spoke proposes to: (1) integrate environmental data, such as air pollution, climate, sociodemographic information with OHDSI, and (2) incorporate state-of-the-art causal discovery methods into OHDSI. We hypothesize that our application will yield valuable hypotheses about the causal relationships between environmental factors and health outcomes.

COMPLETED WITHIN THE PREVIOUS THREE YEARS

**UPMC CCA (Wasan)** 04/01/2019 – 05/31/2020

UPMC Enterprises

Personalized Pain Treatment (PPT)

The Personalized Pain Treatment (PPT) algorithm will generate a user-friendly, personalized shared-decision making tool that is appealing to patients and providers. The report will display the probabilities of success of several possible treatments to guide appropriate therapy selection and reduce financial waste in the healthcare system.

**PA-18-02 (Cohen)** 11/01/2018-4/30/2020

DARPA

Curating Probabilistic Relational Agent-based Models

Although agent-based modeling (ABM) is promising and widely used, agent-based curation is surprisingly primitive. We will develop probabilistic models and algorithms for incremental, human-machine curation of ABMs and demonstrate curation in a disease outbreak problem and a long-term economic risk modeling problem.

**R01 GM088224 (Clermont; Cooper; Hauskrecht)** 08/15/2014-12/31/2019

NIH/NLM

Real-time Detection of Deviations in Clinical Care in ICU Data Streams

This project is applying advanced machine-learning-based methods to detect anomalous clinical decision making in acutely ill patients. The approach works by identifying patient-management patterns that are unusual with respect to patterns associated with comparable patients and by raising a patient-specific alert when such a patient is prospectively encountered.

**U54 HG008540 (Cooper)**  09/29/2014-08/31/2019

NIH

Center for Causal Modeling and Dscovery of Biomedical Knowledge from Big Data

This center of excellence is developing, implementing, and evaluating an integrated set of tools that support causal modeling and discovery (CMD) of biomedical knowledge from very large and complex biomedical data. This Center will make these methods widely available, highly efficient when applied to big datasets, and easy to use. The Center is in the process of providing a powerful set of concepts and tools that accelerate the discovery and sharing of causal knowledge derived from very large and complex biomedical datasets. The approaches and products emanating from this Center are likely to have a significant positive impact on our understanding of health and disease, and thereby on the improvement of human health.

**1R01LM012095 (Visweswaran)** 09/15/2015-06/30/2019

NIH

Development and Evaluation of a Learning Electronic Medical Record System

The goal of this project to develop and evaluate a learning electronic medical records (L-EMR) system that draws a physician’s attention to the right data, at the right time. It learns how to do so by analyzing patterns of patient data access of many physicians in many past cases in the EMR, and learning which EMR data to highlight that are relevant for making clinical decisions in a given patient.

**PA DOH (Cooper)** 06/01/2015-05/31/2019

Big Data for Better Health (BD4BH) in Pennsylvania

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 the project will develop methods for merging, managing, utilizing, processing, analyzing, and sharing large amounts of diverse types of data by developing automated methods for extracting and representing patient-level data, including clinical text, about the course of disease. We will also develop and apply machine learning methods to large complex datasets to predict cancer outcomes based on constructed features that are biologically meaningful. In addition, plans are to partner with another PA institution in order to train underrepresented minority students in the analysis of Big Data.

**U24 GM110707 (Wagner)** 08/05/2014-04/30/2019

NIH

MIDAS Informatics Services Group (ISG)

The broad goal of the project is to catalyze research in infectious disease epidemiology and to improve the related practice of disease control. The project is using the methods of service-oriented architectures and ontologies to build an informatics infrastructure that will enable MIDAS researchers to develop larger and more complex models and larger and more capable systems.

**1R01LM012011 (Lu)** 04/01/2015-03/31/2019

NIH

Deciphering Cellular Signaling System by Deep Mining a Comprehensive Genomic Compendium

In this project, we will compile a comprehensive compendium of human gene expression data and then employ modern deep-learning algorithms and supercomputers to mine the data. We aim to reveal major cellular signals that regulate gene expression under physiological and pathological conditions and to infer the organization of signals in human CSTS. Combined the identified signals with genomic alteration data and drug response data, we aim to further identify pathways underlying disease such as cancers, to use the genomic data to predict drug sensitivity of cancer cell lines, and to predict patient clinical outcomes, all in a pathway-centered manner.