Strategic Plan

Fiscal Years 2021–2025

Office of Portfolio Analysis

National Institutes of Health
Office of Portfolio Analysis

Table of Contents

Introduction

Message from Director, Office of Portfolio Analysis

George Santangelo, Ph.D.
Director

The biomedical research enterprise managed by Institutes, Centers, and Offices (ICOs) of the National Institutes of Health (NIH) spans a complex landscape of topics as diverse as epidemiology, social psychology, tissue engineering, and physical chemistry. Optimizing management of the more than 38,000 NIH-funded research project grants each year (totaling more than $21.5B in FY 2019) requires a deep understanding of that inherent complexity. Toward that end, the Office of Portfolio Analysis (OPA) will continue to support data-driven decision making by developing and disseminating validated, carefully normalized approaches that can analyze past decision making, measure the resulting impact, and forecast the productivity of NIH portfolios, including their potential for successful bench-to-bedside translation. We will also continue to pursue and develop discoveries in metascience that help to promote rigorous, innovative, and transformative research. These components are central to our ongoing commitment to support good stewardship of America's investments in the effort to obtain fundamental knowledge about the nature and behavior of living systems and apply that knowledge to enhance health, lengthen life, and reduce illness and disability.

Over the past several years, OPA has developed artificial intelligence and machine learning (AI/ML) approaches that use a broad spectrum of data sources to create metrics that standardize quantitation of the overall productivity of research investments, including their clinical and technological impact. These data sources include a variety of document types (e.g., grant applications and awards, publications, patents, and clinical trials); associated metadata (e.g., affiliations, authorship, and citations); natural language processing (NLP)-derived elements; libraries of drugs and lead compounds; electronic health records; and economic data. Our validation process ensures that OPA metrics are clearly defined, carefully formulated, scientifically rigorous, and fully transparent through public access to all raw data and peer review of the underlying science. Current OPA tools—which include *iCite*, *iSearch*, and most recently, the COVID-19 Portfolio—also apply AI/ML, NLP, and other analytical schema that provide decision makers with sophisticated ways to search documents, associated metrics and metadata, and other linked datasets. Although these metrics and tools have been shown to improve decision making, they are designed to augment, not substitute for, human judgment.

Our other metascience research projects, which leverage major advances in supercomputing architecture and software design, are rapidly expanding our ability to improve data quality and provide valuable insights. For example, OPA is optimizing our AI/ML disambiguation algorithms to create the comprehensive, interoperable databases needed to fully capture the impact of biomedical research and the scientists who create it. We are also using AI/ML to achieve a degree of dimensionality reduction that allows comparison of the pre-decisional portfolios of private and public funders, facilitating identification of potentially overlapping investments while honoring intellectual property rights and all other data-sharing restrictions. Finally, our efforts to build AI/ML that detects transformative emerging areas of research leverage citation analytics, graph theory, and NLP in a novel, consilient approach.

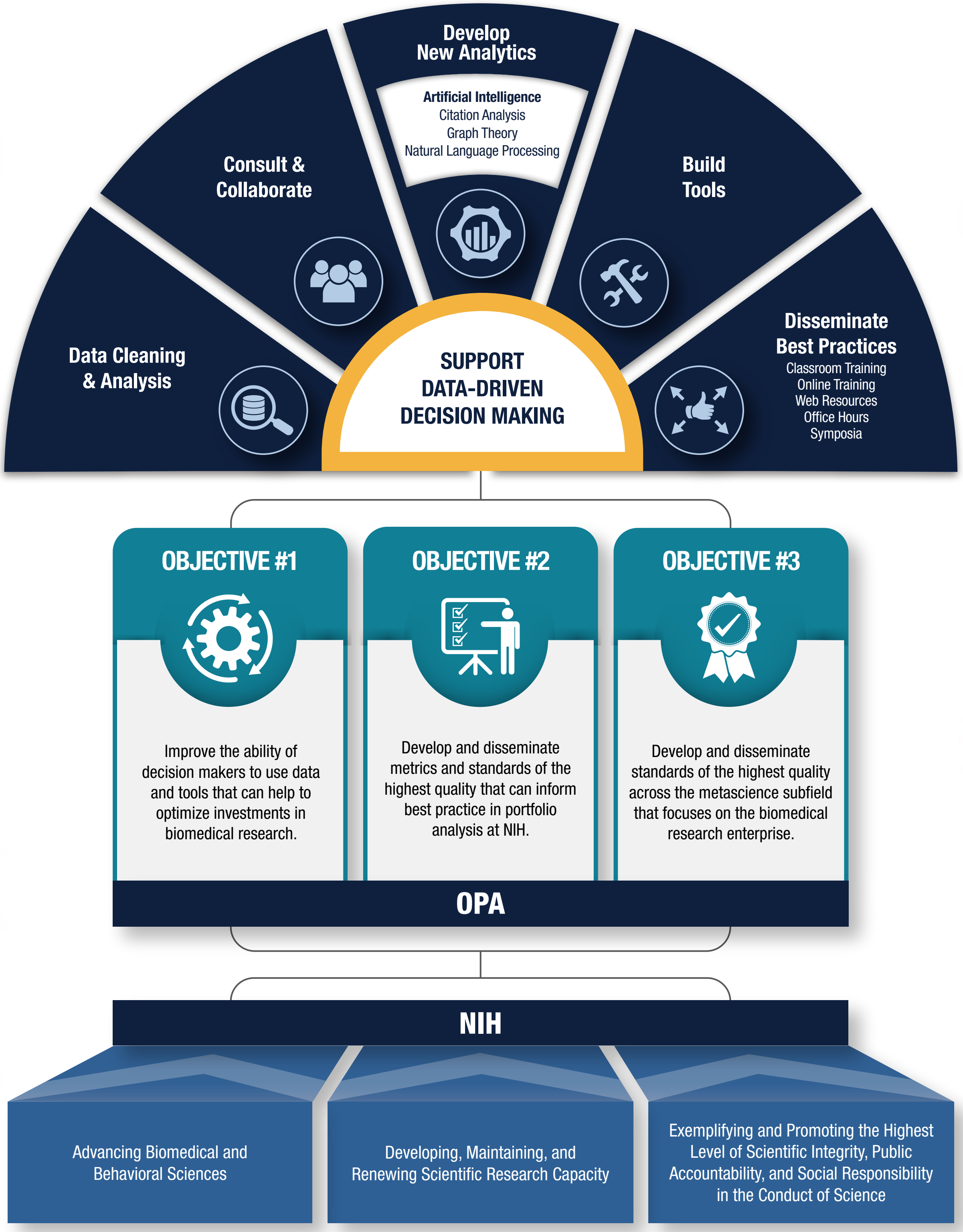
The three objectives in this plan focus on the OPA mission to continue the development and refinement of analytics and tools that can help manage portfolios, identify high-performing areas of research, and detect exciting new scientific opportunities. We will also maintain our robust training, service, and coordination efforts to improve data driven decision making across NIH and beyond, for example, as other federal agencies continue to seek our assistance. I look forward to this next phase of development of OPA as we work together with colleagues inside and outside of NIH in pursuit of these important goals.

George M. Santangelo, Ph.D.
Director, Office of Portfolio Analysis

OFFICE OF PORTFOLIO ANALYSIS

STRATEGIC PLAN, FISCAL YEARS 2021–2025

OVERARCHING GOAL To accelerate biomedical research by providing access to improved methods of data-driven decision making



Data Cleaning & Analysis

Consult & Collaborate

Develop New Analytics

Artificial Intelligence
Citation Analysis
Graph Theory
Natural Language Processing

Build Tools

Disseminate Best Practices

Classroom Training
Online Training
Web Resources
Office Hours
Symposia

SUPPORT DATA-DRIVEN DECISION MAKING

OBJECTIVE #1

Improve the ability of decision makers to use data and tools that can help to optimize investments in biomedical research.

OBJECTIVE #2

Develop and disseminate metrics and standards of the highest quality that can inform best practice in portfolio analysis at NIH.

OBJECTIVE #3

Develop and disseminate standards of the highest quality across the metascience subfield that focuses on the biomedical research enterprise.

OPA

NIH

Advancing Biomedical and Behavioral Sciences

Developing, Maintaining, and Renewing Scientific Research Capacity

Exemplifying and Promoting the Highest Level of Scientific Integrity, Public Accountability, and Social Responsibility in the Conduct of Science

Mission and Organization

OPA was established in 2011 to support the NIH mission as part of the Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI) within the NIH Office of the Director (OD).

The National Institutes of Health Reform Act of 2006 (H.R. 6164-8, Sec. 102) assigns the NIH Director authority to "assemble accurate data to be used to assess research priorities, including information to better evaluate scientific opportunity, public health burdens, and progress in reducing health disparities." The same act also tasks DPCPSI with identifying "research that represents important areas of emerging scientific opportunities, rising public health challenges, or knowledge gaps that deserve special emphasis and would benefit from conducting or supporting additional research that involves collaboration between two or more national research institutes or national centers, or would otherwise benefit from strategic coordination and planning."

In keeping with this statutory authority, the original charge of OPA, equally relevant today, was to develop and disseminate validated new methods, computational tools, and best practices for portfolio analysis; generate high-quality, fully interoperable databases; use those resources to answer important questions about the NIH portfolio; and train NIH staff to do the same. Core features of the OPA mission have focused on recurring questions asked by NIH leadership and ICOs; specifically, how to—

- map topics across portfolios and measure their support by ICOs and other funders;
- quantify the relationship between past investments and the influence and impact of research outputs;
- track the expertise of the biomedical research community and the kinetics of the training pipeline; and
- identify new investments that can address research needs and catalyze transformative discoveries.

Regularly collected feedback about these and the other analytical needs of NIH decision makers has guided several major, long-standing OPA research and development (R&D) initiatives to create AI/ML that can—

- analyze scientific topics at scale;
- measure productivity and impact;
- characterize the scientific workforce;
- detect emerging areas and predict transformative discoveries; and
- create webtools to deliver—
  - access to fully interoperable databases;
  - easy-to-use, powerful analytics; and
  - compelling, informative, and interactive visualizations.

In pursuit of its mission, OPA coordinates extensively with other ICOs, including close collaboration with the National Library of Medicine (NLM) and the following OD offices: the Office of Extramural Research (OER); the Office of Intramural Research (OIR); the Office of Science Policy (OSP); the Office of Evaluation, Performance, and Reporting (OEPR); and the Office of Data Science Strategy (ODSS).

The groundbreaking work of OPA analysts, data scientists, and software engineers has generated substantial interest across NIH, the federal government, international funders, and the broader scientific community. As a result of this increased recognition, OPA is now a central resource for portfolio analysis and provides methodological and analytical support for data-driven decision making by NIH senior leadership, NIH ICOs, and extramural stakeholders. OPA regularly publishes advances in metascience research in peer-reviewed literature (see Figures 1–3) and has established collaborations with administrators, analysts, data scientists, and researchers in government agencies, academia, and the private sector, both in the U.S. and abroad. The major positive impact of OPA achievements on improved data-driven decision making thus extends well beyond NIH.

Figure 1

Using AI/ML to characterize topics described in grant applications and publications

Active management of the NIH portfolio requires effective separation of grants and publications into topically related clusters, including how they are distributed relative to ICO investments. OPA has developed automated AI/ML approaches to analyze trends across the scientific landscape over time, identify overlap between and among portfolios, and characterize emerging areas of research. See Appendix 1 for details.

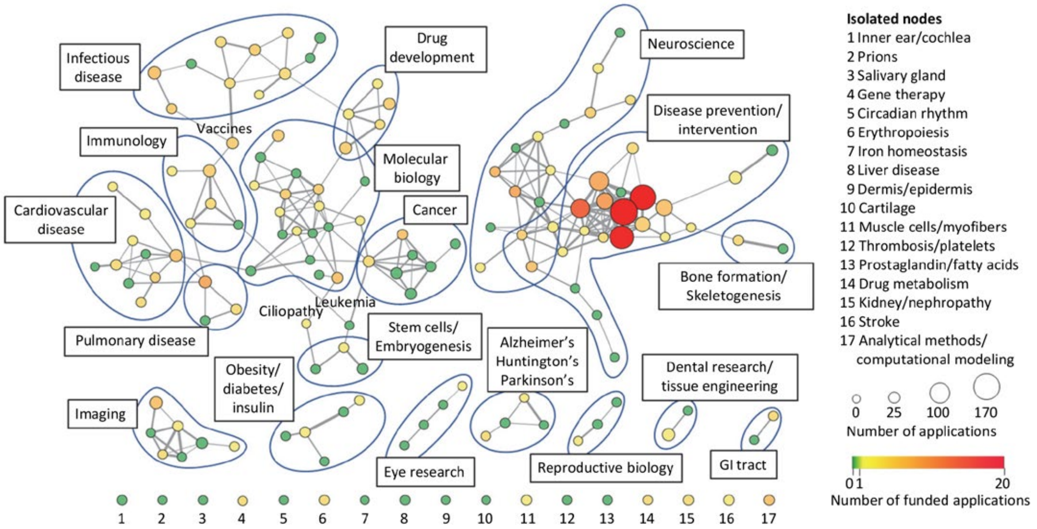


Image source: Hoppe TA, et al. *Sci Adv* 2019;5:eaaw7238. PMID: 31633016.

Figure 2

Using AI/ML to accelerate the discovery of human therapeutics

OPA researchers developed an AI/ML model to predict whether a scientific publication will have an impact on clinical research. This article-level metric, the Approximate Potential to Translate, delivers an accuracy of 84%, and can be used by decision makers as early as 2 years after publication to identify research with a high likelihood of contributing to clinical discoveries that can improve human health. It is freely available to the public in the Translation module of *iCite*. See Appendix 2 for details.

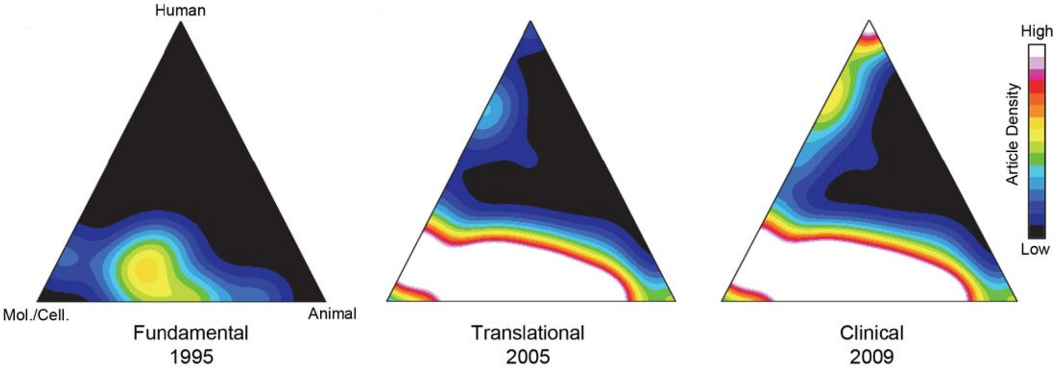


Image shows triangle of biomedicine depiction of the development of cancer immunotherapy. See movie here.

Image source: Hutchins BI, Davis MT, Meseroll RA, Santangelo GM. *PLOS Biol* 2019;17(10): e3000416.

Figure 3

Measuring article influence and disseminating AI/ML-enhanced, high-quality input data

OPA developed the Relative Citation Ratio (RCR), an article-level metric of scientific influence, as an alternative to mathematically flawed measurements of citation activity (e.g., the journal-level impact factor). OPA used AI/ML to identify and extract citation metadata from full-text scientific documents; this work was instrumental in using unrestricted data sources to create the NIH Open Citation Collection (NIH-OCC), a citation database available to the public through the Citations module of *iCite*. See Appendix 3 for details.

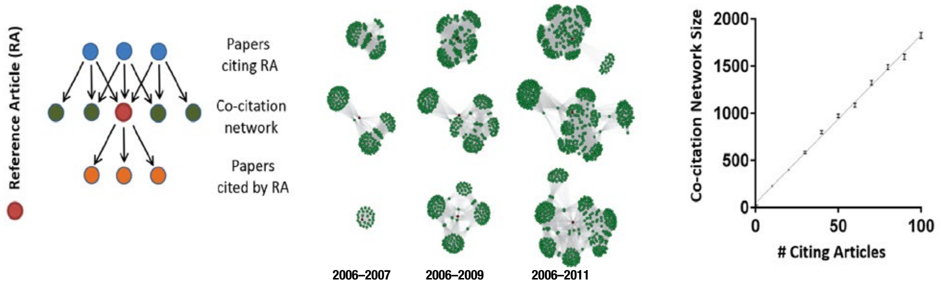


Image source: Hutchins BI, Yuan X, Anderson JM, Santangelo GM. *PLOS Biol* 2016;14(9): e1002541.

Areas of Focus

The following descriptions of current OPA areas of focus provide the background for understanding the goals and strategies presented in this Strategic Plan.

OPA is regularly asked to analyze the entire NIH landscape or particular areas within it. Such analyses include—

- tracking technology development for single-cell analysis and investigating the degree to which there are synergies between Common Fund and ICO technology development initiatives;
- identifying data and code repositories used in NIH research (e.g., bioinformatics, analytics, AI/ML), as well as the ICOs and mechanisms funding these projects;
- characterizing different approaches to COVID-19 diagnostics related to the NIH Rapid Acceleration of Diagnostics (RADx^{SM}) program;
- providing the evidence base for selecting certain therapeutic approaches in the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) program;
- analyzing the current status of NIH funding for surgeon-scientists in the U.S. (published in the *Journal of the American College of Surgeons*); and
- use of AI to generate an NIH Nutrition portfolio, analyze NIH funding for nutrition research across subtopics, ICs and mechanisms, evaluate the productivity of the portfolio, and identify subtopics in which research is progressing rapidly. This analysis was done in support of the Office of Nutrition Research and its transition to the NIH OD.

These analytical activities often highlight areas that require further R&D to improve methodology, data quality, or both. They also build upon previously developed OPA AI/ML algorithms and R&D efforts (see Figures 1–3) that have informed data-driven decision making by—

- improving data quality and database interoperability;
  - disambiguation and fractionation
- refining analysis of the biomedical research landscape;
  - counterfactual analysis
  - tracking and parameterizing interdisciplinary and team science
  - tracking COVID-19 grant applications and publications
- accelerating the acquisition of knowledge needed to improve human health
  - prediction of transformative breakthroughs in biomedical research.

Disambiguation. Disambiguation refers to accurate identification of a person or entity that may go by several names. Effective disambiguation solutions are important for many reasons, perhaps most of all because research contributions cannot be measured accurately without them. Data elements that need to be disambiguated include author names,[1] institution names presented in different forms, and drugs referred to differently as a lead compound versus its trade name. These are the three highest priorities for OPA development of AI/ML disambiguation algorithms.

Fractionation. Because many publications have more than one funded contributor and many publications are supported by more than one research award, OPA is developing fractionation methods that eliminate multiple-counting errors by accurately calculating fractionated grant support and research productivity.

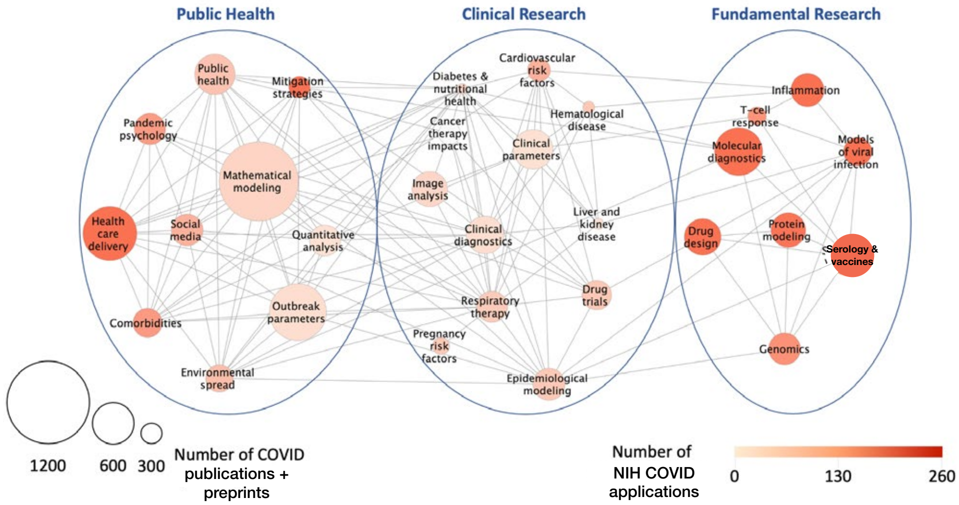
Counterfactual analysis. This type of analysis is made possible by the improved data quality afforded by OPA AI/ML disambiguation and fractionation. Identifying promising applicants who failed in attempts to obtain an NIH research award, but have nevertheless succeeded as biomedical researchers, has the potential to improve NIH decision making by identifying missed opportunities.

Tracking and parameterizing interdisciplinary and collaborative team science. OPA is using AI/ML-based NLP, disambiguation, and graph theory to characterize interdisciplinarity and team science—also known as collaborative science—at scale. We now have an excellent opportunity to follow up on a seminal paper from two of our collaborators that explores the strengths of large and small teams in pursuing important goals in scientific research.[2]

Figure 4

The COVID-19 Portfolio: OPA adaptability to urgent decision making needs

The *iSearch* COVID-19 Portfolio tool was developed as a comprehensive collection of publications and preprints related to the biomedical science of COVID-19 or the novel coronavirus SARS-CoV-2. It is the only resource for literature on the pandemic that both includes preprints and is manually curated by subject-matter experts to eliminate irrelevant search results. It is updated daily with the latest available data, enabling users to analyze the rapidly growing set of advances, as they accumulate in real-time, using the sophisticated *iSearch* platform. The recently launched NIH Preprint Pilot relies upon the selection of preprints in our COVID-19 Portfolio tool. See Appendix 4 for details.



[1] Torvik VI, Smalheiser NR. *Ann Rev Info Sci Tech* 2009; 43:1-43. PMID: 20072710.

[2] Wu L, et al. *Nature* 2019; 556:378-82. PMID: 30760923.

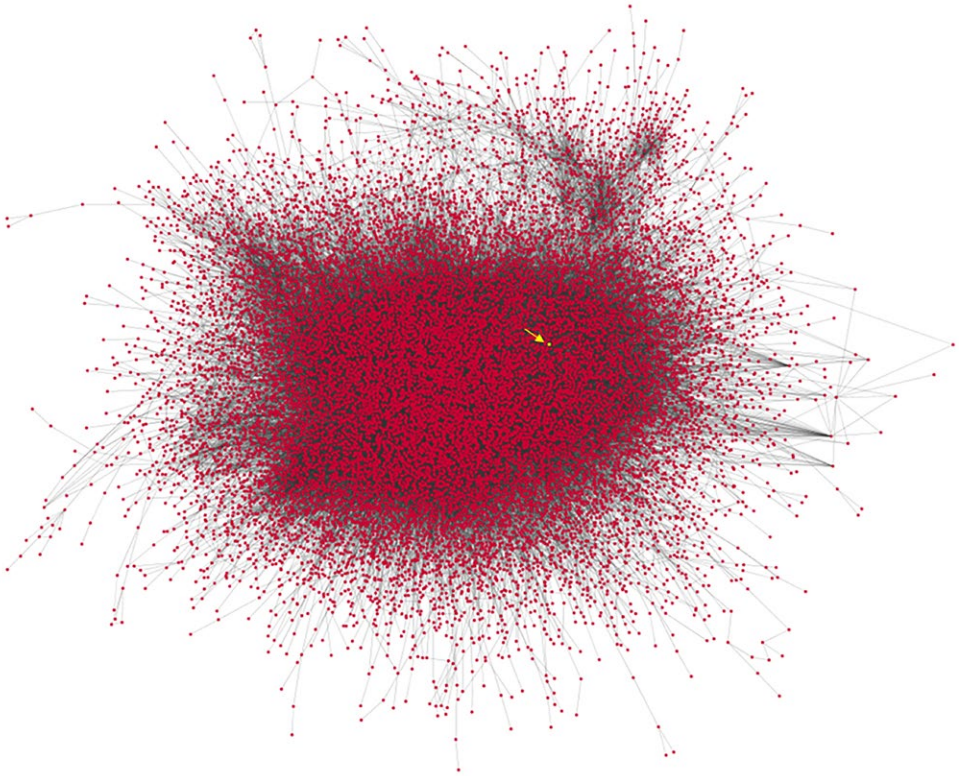
Tracking COVID-19 grant applications and publications. As part of the NIH response to the SARS-CoV-2 pandemic, OPA launched the *iSearch* COVID-19 Portfolio tool on April 12, 2020. See Figure 4 and Appendix 4 for details.

Prediction of transformative breakthroughs in biomedical research. One of the rate-limiting steps in advancing scientific knowledge is the speed with which transformative breakthroughs are recognized. Given the exponential growth of biomedical literature collected in the PubMed database over the past six decades—more than a million papers are now added each year (see Figure 5)—human curation is an increasingly inefficient way to track trailblazing discoveries. Human curation at scale, however, is a daunting challenge due to the rarity of major breakthroughs and the very large database in which they occur. OPA is developing an innovative consilient approach—AI/ML that takes advantage of citation analytics, NLP, and graph theory—to interrogate scientific breakthroughs across PubMed over the last 35 years. Common features of the run-up to these seminal discoveries can be used to predict which areas of science are likely to emerge in the future; these predictions can, in turn, be used to prioritize investments that increase the rate of progress in these emerging areas.

Figure 5

Network of the entire biomedical research literature

This network comprises the 17.2 million publications in PubMed (1981–2017) that received at least one citation. Each node represents a group of topically coherent publications, and they are connected by edges representing co-citations among publications in closely related nodes. The yellow node contains the publications describing the breakthrough discoveries in super-resolution fluorescence microscopy that eventually received a Nobel prize.



The Office of Portfolio Analysis Strategy

Objectives

This Strategic Plan presents three overarching objectives that will continue to provide data and methods that can be used to supplement subject-matter expertise in prioritizing investments and seeking new opportunities to advance science that can improve human health: (1) Expand NIH capabilities in data science and analytics that provide valuable inputs to data-driven decision making; (2) disseminate the highest-quality data and resources to all stakeholders, as well as a comprehensive training program that serves the analytical needs of NIH staff; and (3) play a globally recognized leading role in promoting the most impactful metascience R&D focused on the biomedical research enterprise.

With this overarching goal of accelerating biomedical research by providing access to improved methods of data-driven decision making, OPA will focus on the following FY 2021–2025 objectives. OPA will assess the progress on these objectives regularly and adopt new approaches needed to optimize support for NIH decision makers. Each of the objectives and strategies outlined below provides measurable deliverables via new methods, tools, publications, or training and consultation statistics.

OBJECTIVE 1:

Improve the ability of decision makers to use data and tools that can help to optimize investments in biomedical research

OPA supports the NIH mission by providing data, tools, analyses, and training necessary for NIH leadership, ICO staff, and other stakeholders to make informed decisions about biomedical research investments. Portfolio analysis can reveal the outcomes of past decisions and inform current and future choices. Reliable, well-formulated data can help decision makers improve the return on investment in biomedical research by accelerating the advancement of scientific knowledge. The benefits of a data-driven approach include the potential to inform strategic planning, evaluate the use of different funding approaches, optimize peer review, improve portfolio management, and avoid unnecessary duplication of effort while maintaining a diverse and broad distribution of investments in biomedical research. Data can also help decision makers capitalize on scientific opportunities on the horizon, support transformative emerging areas, shift resources in response to scientific breakthroughs, and address sudden or emerging public health needs. Workforce analysis has also proven to be challenging for many fields; a dynamic model of the biomedical research workforce can help define and establish the optimal biomedical research training pipeline. Of course, new methods and tools capable of addressing these challenges must be thoroughly validated and tested prior to dissemination.

Metrics developed by OPA to track scientific influence and predict the likelihood that research will translate from bench to bedside represent major advances toward a multifaceted approach to research assessment (read this peer-reviewed commentary authored by the OPA Director in 2017). OPA provides various webtools in support of data-driven decision making, including the *iSearch* next-generation analytical suite and the publicly available *iCite* resource. OPA has begun development of a public version of *iSearch*, which will provide external decision makers with a powerful tool to analyze publications, patents, clinical trials, drugs, and awards funded by NIH and other agencies. The launch of public *iSearch* will represent an important advance in the OPA mission to support scientific stewardship across the biomedical enterprise.

Analytical innovation is another important OPA role; one example is improved assessment of productivity by resolving the complex contributions to a single achievement made by multiple investigators or grant awards. Analytical innovation will also focus on developing tools to measure aspects of scientific rigor, transparent reporting and data and resource sharing, enabling NIH to assess the reproducibility of its funded research, guide decision making, and maximize return on investment. These metrics, tools, and innovative AI/ML, complemented by the judgment of subject-matter experts, can help ensure that the NIH portfolio is productive, robust, well prioritized, and balanced across the broad spectrum of NIH investments.

Strategies

- **Determine past productivity by quantifying the advancement of scientific knowledge**
  Accurately measuring productivity requires the synthesis of many different data sources and analytical methods that fully consider the broad scope of what it means for research to be productive. OPA is developing the following approaches to capture the productivity of NIH investments:
  - Use a multifaceted approach to assessing scientific contributions by following the IQRST framework that considers **I**nfluence, **Q**uality, **R**igor/Reproducibility, data/resource **S**haring, and **T**ranslation/Tech transfer (see Figure 6 and strategies outlined below).
  - Improve methods to fractionate input, output, outcome, and impact measurements (i.e., eliminate multiple-counting errors).
  - Adjust for the unique properties and contributions of different areas of research, as well as different types of investigators (e.g., fundamental versus clinical), by using statistical and/or econometric methods (e.g., multivariate regression, propensity matching) when assessing impact.
- **Develop methods that measure scientific rigor, reproducibility, and transparency**
  One of the four stated goals of NIH is to "exemplify and promote the highest level of scientific integrity, public accountability, and social responsibility in the conduct of science." These components are central to the NIH commitment to support good stewardship of America's investments in biomedical research. In 2014, NIH renewed its focus on the importance of rigor, reproducibility, and transparency in biomedical research, implementing new requirements within the grant application process and developing new training modules focused on experimental design. Together, these steps help ensure that NIH-funded biomedical research demonstrates robust statistical analysis, reporting of study design and methodology, and data and material sharing, and that due consideration is given during the grant review process to considerations of scientific premise, rigor, biological variables, and authentication of key biological and/or chemical resources. In this way, NIH joins a worldwide effort focused on ensuring maximum value of the scientific enterprise, optimal return on investment, and increased rates of translational success of preclinical biomedical research. Although OPA does not establish or communicate policy or monitor compliance, it can assist NIH decision makers with those responsibilities by developing and disseminating carefully validated metrics and tools:
  - *Metrics* and *tools*: The term *reproducibility* encompasses multiple tiers (e.g., Methods Reproducibility, Results Reproducibility, and Inferential Reproducibility[3,4]). In developing metrics and tools, OPA will take these concepts of reproducibility into

[3] https://www.acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf.

[4] Pulverer, B. *Embo J* 2015;34(22):2721-4. PMID: 26538323.

account, alongside other quantifiable examples of best research practice (e.g., pre-registration of relevant animal research). The process of generating the required yardstick(s) will require, as always, input from both intramural and extramural stakeholders. OPA will thereby facilitate the NIH commitment to ensuring that its research investment supports the most rigorous science, to evaluating the effects of existing interventions, and to supporting a more comprehensive model of data-based decision making.

- *Dissemination*: By disseminating valid metrics and tools, OPA can aid the efforts of NIH decision makers in promoting best research practices and generating a research culture change.[5] NIH Director Dr. Francis Collins has commented on the current academic incentive system, including overemphasis on publishing in high-profile journals and the consequences for scientific fidelity. In this spirit, OPA will carefully consider these issues in the multifaceted attempt to quantify productivity, return on investment, and impact.[6]

- **Develop methods that measure data/resource sharing**

Scientific advances are facilitated not only by the dissemination and refinement of higher-level ideas but also by the sharing of underlying, fundamental datasets from which either those ideas or unique knowledge can be independently derived. Accordingly, NIH has long-recognized research data to be valuable products of each biomedical investigation and has established policies that encourage or require data/resource sharing by their funding recipients. Data/resource sharing can enhance the efficiency of biomedical research investments by increasing return on investment and avoiding unnecessary duplication of effort via reuse of extant data. Measuring the sharing and reuse of experimental datasets produced by NIH-funded research projects therefore has the potential to substantially improve management of the NIH portfolio.

- *Metrics* and *tools*: The commitment of NIH, as well as many scientific journals, to promoting data/resource sharing has ushered in an increased deposition of research datasets into online repositories that are accessible by the broader scientific community. OPA will enhance existing resources and develop new metrics and tools that quantify data and resource sharing. In part, this requires effective linkage of data resources to both research awards and publications, which would have the further advantage of capturing the productivity and impact associated with such practices. These efforts will be conducted in collaboration with NIH ICOs and outside entities working toward similar goals and will adhere to privacy rules that prevent the misuse of personally identifiable health information. Developing such metrics and tools is a necessary step toward establishing an infrastructure whereby data/resource sharing behaviors can be accurately evaluated and appropriately accounted for in the optimization of investments in biomedical research.

- *Dissemination*: By disseminating ways to quantify data/resource sharing, OPA can assist in the effort to incentivize investigators to share their data and to motivate data stewards (e.g., repositories) to make data more Findable, Accessible, Interoperable, and Reusable (FAIR).

[5] Koroshetz WJ, et al. *Elife* 2020;9:e55915. PMID: 32127131.

[6] Moher D, et al. *PLOS Biol* 2020;18(7) e3000737. PMID: 32673304.

Figure 6

The IQRST Framework for evaluating productivity

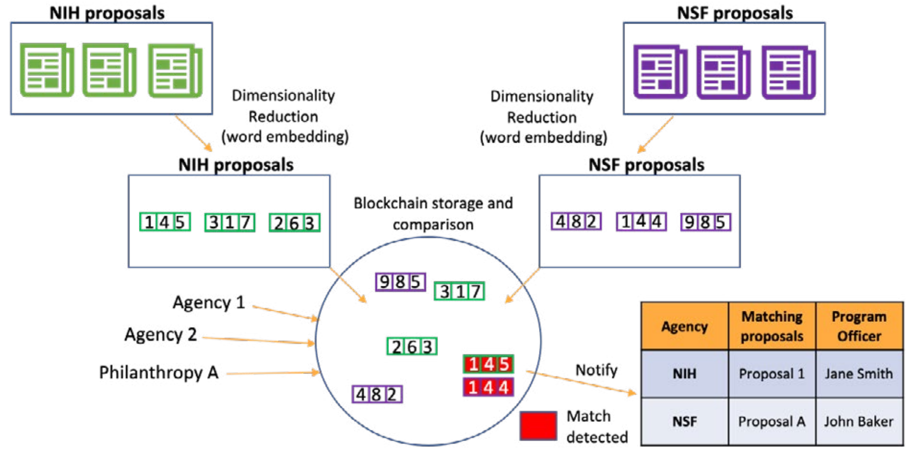
I = **I**nfluence (weighted Relative Citation Ratio [RCR])
Q = **Q**ualitative human judgment
R = **R**igor/Reproducibility of research
S = **S**haring of scientific data/resources
T = **T**ranslation/Tech transfer (aggregated data on clinical trials, patents, drugs, and devices and/or Approximate Potential to Translate score)

Further detail on the value of using diverse, validated metrics to assess scientific output is described in a peer-reviewed commentary authored by the OPA Director in 2017. The *R* was previously discussed in this article as *Reproducibility* but has been updated to include the fundamental and, perhaps more important, principles of rigor and thorough reporting that underlie reproducibility.

- **Facilitate increased productivity by using AI/ML to detect and/or predict the impact of policy and funding decisions in real time**
  OPA has successfully deployed AI/ML to measure scientific influence and predict clinical impact at the level of individual articles. OPA will continue to build on this momentum, developing new algorithms that harness the power of AI/ML to inform decision making. OPA AI/ML projects that are either planned or currently underway include the following:
  - Detect emerging areas and predict their potential to produce transformative science
  - Determine the age and rate of progress for each research topic
  - Find gaps in the NIH portfolio and identify new scientific opportunities
  - Identify missed opportunities to fund excellent science with counterfactual analyses
  - Inform improved data-driven management of portfolios with AI/ML NLP and other methodologies
  - Investigate leading-edge indicators—such as meeting abstracts, social media posts, and news articles—as data sources that have the potential to improve detection or prediction of impact
  - Optimize methods that can identify overlap among NIH proposals and those in other agencies in real time (see Figure 7)
  - Support nimble decision making in response to public health crises and other emerging challenges (see Figure 4)
- **Inform effective management of the training pipeline and scientific workforce**
  OPA is also improving on strategies we began developing over the past few years to analyze trainee populations, training mechanisms, and workforce dynamics:
  - Study the distribution of expertise, career age, co-authorship, and mentor/mentee relationships among both prospective and current NIH principal investigators (PIs), using AI/ML to extract and structure the biosketch data of NIH applicants
  - Model demographic changes in the scientific workforce with stock flow analyses
  - Identify factors that play an important role in successful career progression (e.g., the transition from postdoctoral trainee to NIH PI)
  - Characterize the diversity of topics studied by NIH trainees
  - Define successful collaborations and determine the most effective way(s) to initiate them and maintain support
  - Track and parameterize team science and interdisciplinarity to assess the relative contributions of various approaches to building successful research teams

Figure 7 Pre-decisional management of overlapping proposals at funding organizations

Coordinating R&D investments across federal agencies and other public and private funders is not happening at scale. OPA data scientists and analysts are collaborating with a team from the National Science Foundation (NSF) to develop AI/ML that flags pre-decisional applications with overlapping aims, without the need to share private information or intellectual property. This approach has the potential to improve stewardship of research investments across funding entities by reducing or eliminating unnecessary duplication, managing overlapping research, and encouraging collaboration between and among researchers with similar interests.



Agency	Matching proposals	Program Officer
NIH	Proposal 1	Jane Smith
NSF	Proposal A	John Baker

Figure 8 Request For Information (RFI) Tool

The federal government issues hundreds of RFIs every year, resulting in thousands of responses. Depending on the complexity and format of the responses, the review, analysis, and summarization of responses to each of these RFIs is often laborious and burdensome. The OPA RFI tool has simplified and streamlined the coding and analysis of RFI responses or other text collections. This tool allows the normalization of organization names, creation of curatable fields and custom vocabularies, and user-friendly drill-down on coded results. It is widely used across NIH and within the U.S. Department of Health and Human Services (HHS), our parent department.

Objective 1 aligns with the IC strategic planning requirement of the 21st Century Cures Act (42 USC 289a-2(a)(3)), which requires the inclusion of women and minorities and a focus on reducing health disparities. By developing and disseminating tools that evaluate compliance with policies governing data sharing (e.g., genomics data and clinical trial reporting), Objective 1 activities can also enhance risk-based approaches to NIH oversight and help NIH staff monitor compliance with NIH policies and initiatives.

OBJECTIVE 2:

Develop and disseminate metrics and standards of the highest quality that can inform best practice in portfolio analysis at NIH

From its inception, one of the critical imperatives of OPA R&D has been to share data, methods, and tools with the NIH community and the public. OPA R&D—including the creation of novel data pipelines, high-quality interoperable databases, and AI/ML—continues to improve analytical capabilities across NIH and the entire biomedical community. Widespread dissemination and validation of OPA data, methods, and tools, for example, by publication in internationally recognized, peer-reviewed journals, is an essential means of promoting the highest level of scientific integrity, reproducibility, public accountability, and good stewardship of OPA resources. Because new methods cannot have impact without establishing a broad consensus, these activities have the added benefit of promoting adoption. They also encourage the valuable feedback needed to improve accuracy and performance and enhance the quality and reproducibility of analyses generated by NIH staff and extramural stakeholders alike.

Internal dissemination is augmented by the OPA training curriculum and one-on-one consultations. Additional OPA activities to coordinate portfolio analysis across NIH and enhance collaboration among all portfolio stakeholders include sponsoring and hosting poster sessions, workshops, and symposia.

OPA R&D and training activities will also continue to focus on empowering NIH staff to produce high-quality analyses that inform the public about the accomplishments, impact, and data-driven prioritization of NIH-funded research. Priority setting of OPA activities is important and is determined by the trans-NIH nature of the project when allocating OPA time and resources. Finally, to ensure that OPA resources are clearly differentiated from those maintained by other ICOs, we will continue to work closely with OER, OEPR, ODSS, and other ICOs to facilitate harmonization of all resources and training used by NIH staff and other stakeholders.

Strategies

- **Catalyze adoption of data-driven approaches to decision making across NIH and beyond**
  OPA will continue to provide critical resources for portfolio analysis to NIH staff and the wider community of stakeholders, including the public. Trans-NIH efforts will continue to be the highest priority for OPA engagement. OPA efforts in support of data-driven decision making include the following:
  - Maintain web resources
  - Develop new tools
  - Regularly collect feedback (e.g., surveys and focus groups)
  - Engage with NIH staff through consultations
  - Enhance the NIH training curriculum
  - Support and host workshops and other gatherings of the portfolio analysis community
  - Support and host symposia
- **Stewardship of OPA R&D efforts**
  OPA dissemination of R&D will ensure transparency, rigor, reproducibility, and data/resource sharing:
  - Publish all methodological advances in internationally recognized peer-reviewed journals

  - Share both raw and processed data (includes creating and maintaining open resources [e.g., the *iCite* tool], specifically the NIH-OCC module)
  - Share code (e.g., by depositing on GitHub)
- **Coordinate with other NIH ICOs—including OER, OEPR, and ODSS—to ensure that our data, methods, tools, and training meet the highest possible standards of quality and are well aligned to avoid unnecessary duplication or expense**
  - Ensure that all underlying raw and processed data, whether used in NIH analyses or as the databases underlying internal or public-facing NIH tools, attain the highest possible quality
  - Streamline and functionally differentiate the user interface of internal NIH tools to optimize the workflow of NIH staff in administrative, reporting, and analytic activities
  - Collectively provide resources and training opportunities that highlight the functional differentiation of internal NIH tools and clearly explain the purpose and appropriate uses of each tool
  - Streamline and functionally differentiate the user interface of external tools to optimize transparency and data quality, and provide external stakeholders with the most effective public resource for analytic inquiries

OBJECTIVE 3:
Develop and disseminate standards of the highest quality across the metascience subfield that focuses on the biomedical research enterprise

The emerging field called metascience uses interdisciplinary approaches to study how scientists are supported, do their work, and advance scientific knowledge. Metascience provides a quantitative approach to understanding, among other things, how discoveries arise, how funding priorities are established, and how science is pursued within and across disciplines and institutional structures. In the context of biomedical research, the aim is to understand the factors that advance scientific knowledge that improves human health. Metascience researchers have begun to address issues as diverse as the effects of team size and constitution, the role of bias in peer review, and the emergence of new scientific topics. The scope and salience of metascience research have grown as more structured, higher-quality datasets become available.

OPA contributes to the development of the metascience field by implementing improvements in data quality and availability and establishing best practices in portfolio analysis. OPA has an active presence in the metascience community and collaborates with other practitioners across the federal government, private sector, and academia to meet our shared goal of improving the effectiveness of the scientific enterprise. As data collection and analysis continue to evolve, OPA has a role to play in supporting further improvements to both the datasets and methodology, as well as dissemination of best practices. Both independently and in partnership with other funders, academics, and private sector entities, OPA is uniquely situated to contribute, in part due to the very large footprint and well-elaborated data structure of the biomedical research enterprise.

Strategies

- **Coordinate analytical efforts within NIH and share OPA resources and best practices with other funders**
  - Work with analysts and decision makers across NIH to establish best practices for portfolio analyses, methodologies, and tools focused on the biomedical research enterprise
  - Share best practices in portfolio analysis with other domestic and international funders, both public and private, to foster scientific stewardship across the global metascience landscape
- **Catalyze improvements in human health by linking NIH accomplishments to downstream clinical or commercialization activities**

  Several approaches developed by OPA may be applicable to decision making in the private sector:
  - Measure the clinical and technological impact of NIH investments
  - Disambiguate drug names
  - Predict clinical impact
  - Predict which emerging topics or transformative discoveries are likely to yield drugs or devices approved by the U.S. Food and Drug Administration (FDA)
- **Collaborate with leading metascience practitioners in the public and private sectors**

  OPA will continue to engage with metascience practitioners, including academics and other thought leaders, to establish the analysis of the biomedical research enterprise as a high-quality, well-defined, and widely recognized subdiscipline

Alignment With NIH-Wide Strategic Plan and Objectives

NIH Objective #1
Advancing Biomedical and Behavioral Sciences

OPA Objective 1 aligns directly with Objective #1 of the NIH-Wide Strategic Plan for Fiscal Years 2021–2025, including its following subsections: Building Data Resources to Enable Research Progress; Inventing Tools and Technologies to Catalyze Discovery; and Harnessing Technology to Inform Decision Making.

NIH Objective #2
Developing, Maintaining, and Renewing Scientific Research Capacity

OPA Objective 1 aligns directly with this NIH-wide Objective by helping to characterize, track and Enhance the Biomedical and Behavioral Research Workforce.

NIH Objective #3
Exemplifying and Promoting the Highest Level of Scientific Integrity, Public Accountability, and Social Responsibility in the Conduct of Science

All three OPA Objectives align directly with this NIH-wide Objective, including its following subsections: Monitoring Expenditures and Scientific Progress; Making Evidence-Informed Decisions; Assessing Programs, Processes, Outcomes, and Impact; Communicating Results; Enhancing Reproducibility Through Rigorous and Transparent Research (via developing methods to assess scientific rigor and reporting); developing tools to assess Transparency Through Data Access and Sharing; and Managing Risks to the Research Enterprise (via portfolio performance and counterfactual analysis).

Figure 9

Alignment with Crosscutting Themes of the NIH-Wide Strategic Plan

Promoting Collaborative Science

One crosscutting theme of the NIH-Wide Strategic Plan is promoting team-driven research that involves collaboration across multiple scientific fields. OPA goals—which include characterizing, tracking, and parameterizing interdisciplinary and collaborative team science—align directly with this theme.

Leveraging Data Science for Biomedical Discovery

OPA AI/ML efforts, which are focused on developing new analytic and data visualization tools, directly align with this NIH-wide theme of using innovative approaches to address opportunities and challenges in data science. Similarly, OPA training programs directly support the NIH goal of enhancing training in computational and data science fields.

Figure 10

Alignment with the NIH-Wide Strategic Plan for COVID-19 Research

OPA will continue to assist NIH leadership with large-scale analyses of therapeutic and vaccine candidates and their potential for translation. In April 2020, OPA also released the COVID-19 Portfolio, a comprehensive database of all COVID-19-related publications and preprints (see Figure 4 and Appendix 4). OPA curates and updates this dataset daily to provide a valuable resource for the latest information on COVID-19 research to both the NIH and the public. These activities align well with NIH Strategic Priorities 1 (Improve Fundamental Knowledge of SARS-CoV-2 and COVID-19) and 3 (Support Research to Advance Treatment).

Bold Prediction

Ambitious and aspirational goals are important for OPA as we strive to push boundaries in data analytics and methodologies that inform decision making by NIH communities and beyond. Our bold prediction for the next 5 years is therefore the following:

NIH will be the first funder of science, either in the government or in the private sector, to develop AI/ML that both serves as a leading indicator of emerging topics and predicts which topics will produce transformative breakthroughs.

OPA has developed an AI/ML-based method to identify research topics that will experience a transformative breakthrough and continues to develop methods to quantify the relationship between investments and the advancement of scientific knowledge. NIH decision makers will be able to use these methods to accelerate discovery in both mature and emerging areas of biomedical research.

Conclusion

This 5-year plan involves many action items that follow on, complement, or supplement existing activities. Other action items will be started anew, and still others will unfold in the coming years. The methods and tools derived from OPA R&D will continue to evolve, informed by metrics we have developed to monitor performance. OPA will revisit this plan periodically to assess our progress and remain responsive to the needs of NIH leadership, staff, and the scientific community. New developments in information science, computer architecture, or software engineering—or in biomedical research itself—might spark new ideas or suggest that existing strategies are misaligned with new opportunities. OPA will engage in an iterative process to ensure that we continue to stay on track in meeting the three objectives outlined here. As always, OPA will continue to seek new partners and opportunities for collaboration within and outside NIH on all our strategies, approaches, and objectives.

Description of the Strategic Planning Process

This OPA Strategic Plan was created with input from colleagues and stakeholders across NIH and the federal government, metascience researchers in academia, the broader biomedical research community, and the public. This input helped refine the team's priorities and outline its role in advancing data-driven decision making. OPA also regularly conducts surveys and collects feedback from the NIH community and the public; this feedback has informed the tool and methodological development objectives reflected in this Strategic Plan. Close cooperation with stakeholders will continue to be essential and may reveal new unanticipated directions for OPA for the future.

Representative Accomplishments

Appendix 1

Using AI/ML to characterize topics described in grant applications and publications

Active management of the NIH portfolio requires effective separation of grants and publications into topically related clusters, including how they are distributed relative to ICO investments. OPA has automated this process by further development of AI/ML approaches, such as word2vec,[a] which we have used to analyze trends across the scientific landscape over time, identify overlap between and among portfolios, and characterize emerging areas of research. OPA published the validation of this method in a paper describing the discovery that African American/Black scientists who apply for NIH R01 funding are more likely to study topics with lower funding rates.[b] In collaboration with numerous ICOs, OPA has also used AI/ML to solve portfolio classification problems that were previously intractable. One example is ongoing work to classify disease prevention grants with high precision and recall.[c] Several ICs have also requested assistance from OPA to analyze the content of their portfolios to support strategic planning or portfolio reorganization.

[a] Mikolov T, et al. arXiv preprint 2013; arXiv:1301.3781.

[b] Hoppe TA, et al. *Sci Adv* 2019;5:eaaw7238. PMID: 31633016.

[c] Villani J, et al. *Am J Prev Med* 2018;55:926-31. PMID: 30458951.

Appendix 2

Using AI/ML to accelerate the discovery of human therapeutics

Scientific advances can take decades to translate into improvements in human health. Shortening this interval could increase the rate at which scientific discoveries lead to successful treatment of human disease. OPA researchers developed an AI/ML model to predict whether a scientific publication will have an impact on clinical research.[a] As soon as 2 years after a paper appears, OPA scientists can determine the probability that it will eventually be cited by a clinical article (published clinical trial or guideline). This article-level metric—the Approximate Potential to Translate score, or APT—can be used by decision makers to identify research with a high likelihood to contribute to clinical discoveries that can improve human health. It is freely available to the public in the Translation module of *iCite*.

[a] Hutchins BI, et al. *PLOS Biol* 2019;17:e3000416. PMID: 31600189.

Appendix 3

Measuring article influence and disseminating AI/ML-enhanced, high-quality input data

Bibliometric analyses that can help to determine the impact of a portfolio of grants or publications must use accurate and transparent citation data. OPA developed the Relative Citation Ratio (RCR),[a] an article-level metric of scientific influence, as an alternative to mathematically flawed measurements of citation activity (e.g., the journal-level impact factor) that continue to be used inappropriately as a proxy for the quality or impact of research publications. Quantifying the influence of a paper must be done at the article level, normalized to its field of study, and be transparent and reproducible—which, in turn, requires open access to raw citation data. Historically, citation data have remained locked behind restrictive licensing agreements, hampering the ability of researchers to identify reference linkages between scientific articles. To address this barrier, OPA used AI/ML to identify and extract citation metadata from full-text scientific documents; this work was instrumental in using unrestricted data sources to create the NIH Open Citation Collection (NIH-OCC),[b] a public citation database available in the Citations module of *iCite* and as a bulk download from figshare. Citations from the NIH-OCC—which are a validated source of reproducible, trustworthy bibliometric data—underlie all calculations and metrics in *iCite*.

[a] Hutchins BI, et al. *PLOS Biol* 2016;14:e1002541. PMID: 27599104.

[b] Hutchins BI, et al. *PLOS Biol* 2019;17:e3000385. PMID: 31600197.

Appendix 4

The COVID-19 Portfolio: OPA adaptability to urgent decision making needs

The *iSearch* COVID-19 Portfolio tool was developed as a resource for probing a comprehensive collection of publications and preprints curated by subject-matter experts to focus entirely on the biomedical science of COVID-19 or the novel coronavirus SARS-CoV-2. Our COVID-19 Portfolio is the only resource for literature on the pandemic that both includes preprints and is manually curated by subject-matter experts to eliminate irrelevant search results. The tool also leverages the next-generation analytical capability of the *iSearch* platform and includes peer-reviewed articles from PubMed and preprints from arXiv, bioRxiv, ChemRxiv, medRxiv, Preprints.org, Research Square, and SSRN. *iSearch* is updated daily with the latest available data, enabling users to explore and analyze the rapidly growing set of advances in COVID-19/SARS-CoV-2 research as they accumulate in real time. We designed the COVID-19 Portfolio platform to provide maximum flexibility and ease of use. Users can access interactive visualizations for download or further queries; search the full text of articles or supplemental datasets, in addition to titles and abstracts; store and share URLs to save and periodically update the results of useful search strategies; and access a host of other features. Developing this COVID-19 Portfolio presented a quick-response technological tour de force in ingesting preprint literature. The recently launched NIH Preprint Pilot is relying on the selection of preprints we have included in our COVID-19 Portfolio search tool, demonstrating that our novel approach to tracking and inclusion of preprints has broad appeal.

Appendix 5

Acronym List

ACTIV	Accelerating COVID-19 Therapeutic Interventions and Vaccines
AI/ML	artificial intelligence and machine learning
APT	Approximate Potential to Translate
DPCPSI	Division of Program Coordination, Planning, and Strategic Initiatives
FAIR	findable, accessible, interoperable, and reusable
FDA	U.S. Food and Drug Administration
ICOs	Institutes, Centers, and Offices
IQRST	influence, quality, rigor, data/resource sharing, and translation/tech transfer
NIH	National Institutes of Health
NIH-OCC	NIH Open Citation Collection
NLM	National Library of Medicine
NLP	natural language processing
NSF	National Science Foundation
OD	Office of the Director
ODSS	Office of Data Science Strategy
OEPR	Office of Evaluation, Performance, and Reporting
OER	Office of Extramural Research
OPA	Office of Portfolio Analysis
OSP	Office of Science Policy
PI	principal investigator
R&D	research and development
RADx^{SM}	Rapid Acceleration of Diagnostics
RCR	Relative Citation Ratio
RFI	Request for Information

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of Portfolio Analysis