

## Hypothesis/Commentary

## Cancer Epidemiology in the 21st Century

Transforming Epidemiology for 21st Century Medicine and Public Health 

Muin J. Khoury<sup>1,5</sup>, Tram Kim Lam<sup>1</sup>, John P.A. Ioannidis<sup>6</sup>, Patricia Hartge<sup>2</sup>, Margaret R. Spitz<sup>7</sup>, Julie E. Buring<sup>8</sup>, Stephen J. Chanock<sup>2</sup>, Robert T. Croyle<sup>1</sup>, Katrina A. Goddard<sup>12</sup>, Geoffrey S. Ginsburg<sup>13</sup>, Zdenko Herceg<sup>14</sup>, Robert A. Hiatt<sup>15</sup>, Robert N. Hoover<sup>2</sup>, David J. Hunter<sup>10</sup>, Barnet S. Kramer<sup>3</sup>, Michael S. Lauer<sup>4</sup>, Jeffrey A. Meyerhardt<sup>9</sup>, Olufunmilayo I. Olopade<sup>16</sup>, Julie R. Palmer<sup>11</sup>, Thomas A. Sellers<sup>17</sup>, Daniela Seminara<sup>1</sup>, David F. Ransohoff<sup>18</sup>, Timothy R. Rebbeck<sup>19</sup>, Georgia Tourassi<sup>20</sup>, Deborah M. Winn<sup>1</sup>, Ann Zauber<sup>21</sup>, and Sheri D. Schully<sup>1</sup>

## Abstract

In 2012, the National Cancer Institute (NCI) engaged the scientific community to provide a vision for cancer epidemiology in the 21st century. Eight overarching thematic recommendations, with proposed corresponding actions for consideration by funding agencies, professional societies, and the research community emerged from the collective intellectual discourse. The themes are (i) extending the reach of epidemiology beyond discovery and etiologic research to include multilevel analysis, intervention evaluation, implementation, and outcomes research; (ii) transforming the practice of epidemiology by moving toward more access and sharing of protocols, data, metadata, and specimens to foster collaboration, to ensure reproducibility and replication, and accelerate translation; (iii) expanding cohort studies to collect exposure, clinical, and other information across the life course and examining multiple health-related endpoints; (iv) developing and validating reliable methods and technologies to quantify exposures and outcomes on a massive scale, and to assess concomitantly the role of multiple factors in complex diseases; (v) integrating "big data" science into the practice of epidemiology; (vi) expanding knowledge integration to drive research, policy, and practice; (vii) transforming training of 21st century epidemiologists to address interdisciplinary and translational research; and (viii) optimizing the use of resources and infrastructure for epidemiologic studies. These recommendations can transform cancer epidemiology and the field of epidemiology, in general, by enhancing transparency, interdisciplinary collaboration, and strategic applications of new technologies. They should lay a strong scientific foundation for accelerated translation of scientific discoveries into individual and population health benefits. *Cancer Epidemiol Biomarkers Prev*; 22(4); 508–16. ©2013 AACR.

For decades, epidemiology has provided a scientific foundation for public health and disease prevention (1). Epidemiology has contributed to major scientific discoveries such as the relationship between cigarette smoking

and common diseases (2). Yet, the observational nature of much of epidemiologic research has attracted criticism including "excess expense, repudiated findings, studies that offer small incremental knowledge, inability to

**Authors' Affiliations:** <sup>1</sup>Division of Cancer Control and Population Sciences, <sup>2</sup>Division of Cancer Epidemiology and Genetics, and <sup>3</sup>Division of Cancer Prevention, National Cancer Institute, and <sup>4</sup>Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, NIH, Bethesda, Maryland; <sup>5</sup>Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia; <sup>6</sup>Stanford Prevention Research Center, Department of Medicine, Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, California; <sup>7</sup>Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas; <sup>8</sup>Division of Medicine, Brigham and Women's Hospital; <sup>9</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School; <sup>10</sup>Department of Epidemiology and Nutrition, Harvard School of Public Health; <sup>11</sup>Slone Epidemiology Center, Boston University, Boston, Massachusetts; <sup>12</sup>Center for Health Research, Kaiser Permanente, Portland, Oregon; <sup>13</sup>Center for Genomic Medicine, Institute for Genome Sciences and Policy, Duke University School of Medicine, Durham, North Carolina; <sup>14</sup>International Agency for Research on Cancer, Lyon, France; <sup>15</sup>Department of Epidemiology and Biostatistics, University of California at San Francisco, San Francisco, California; <sup>16</sup>Center for Clinical Cancer Genetics and

Global Health, Department of Medicine, University of Chicago Medicine, Chicago, Illinois; <sup>17</sup>Office of the Director, Moffitt Cancer Center, Tampa, Florida; <sup>18</sup>Departments of Epidemiology and Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; <sup>19</sup>Department of Biostatistics and Epidemiology and Abramson Cancer Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania; <sup>20</sup>Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee; and <sup>21</sup>Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York

**Note from the Editor-in-Chief:** This is one in a series of commentaries that have arisen from an initiative of the National Cancer Institute to advance epidemiological science in the 21st century.

**Corresponding Author:** Muin J. Khoury, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333. Phone: 404-498-0001; Fax: 404-498-0140; E-mail: muk1@CDC.GOV

doi: 10.1158/1055-9965.EPI-13-0146

©2013 American Association for Cancer Research.

**Table 1.** Broad recommendations and proposed actions to transform epidemiology for 21st century medicine and public health

<b>Recommendation</b>	<b>Proposed actions</b>
<p><u>Extend the reach of epidemiology</u></p> <p>Balance the epidemiology research portfolio beyond traditional emphasis on discovery and etiologic research to encompass development and evaluation of clinical and population interventions, implementation, dissemination, and outcomes research</p>	<ul style="list-style-type: none"> <li>■ Create incentives to balance discovery and translational research</li> <li>■ Foster integration of observational epidemiologic studies with intervention trials</li> <li>■ Encourage academic and research institutions to promote career advancement that rewards collaborative, interdisciplinary, and translational research</li> </ul>
<p><u>Transform the practice of epidemiology</u></p> <p>Move toward greater access to data, metadata, and specimens to foster collaboration, to ensure reproducibility, replication, and to accelerate translation into population health impact</p>	<ul style="list-style-type: none"> <li>■ Support the harmonization of existing epidemiologic data (including cohorts and consortia) and the creation of study repositories</li> <li>■ Support processes for registration of new studies, data access and sharing, and collaborative analyses</li> <li>■ Work with scientific journals and academic institutions to create more incentives for data sharing, reproducibility, and replication</li> </ul>
<p><u>Expand cohort studies across the lifespan including multiple health outcomes</u></p> <p>Maximize the output and productivity from existing cohorts and assess the need for new cohorts of etiology and outcomes including multiple health-related outcomes and intermediate biomarkers</p>	<ul style="list-style-type: none"> <li>■ Map and register existing cohort studies worldwide</li> <li>■ Expand current studies to include multiple outcomes and to incorporate early life events and pre and postdiagnostic information</li> <li>■ Engage with stakeholders and field leaders to discuss the concept of a national (centralized or synthetic) cohort for multiple health-related outcomes</li> </ul>
<p><u>Develop, evaluate, and use novel technologies appropriately</u></p> <p>Develop and validate reliable methods and technologies to quantify exposures and outcomes in massive scale and to assess concomitantly multiple factors in complex diseases</p>	<ul style="list-style-type: none"> <li>■ Support pilot studies that leverage existing resources to validate new and emerging technologies for epidemiologic studies</li> <li>■ Support methodologic work for measuring and modeling concomitantly multiple risk factors and outcomes</li> </ul>
<p><u>Integrate "big data" science into the practice of epidemiology</u></p> <p>Develop systematic approaches to manage, analyze, display and interpret large complex datasets</p>	<ul style="list-style-type: none"> <li>■ Support the development and maintenance of scalable and sustainable bioinformatics and data storing infrastructures that can handle large, complex, and diverse data sets</li> <li>■ Promote cross-study best practices for managing complex datasets and develop novel analytic strategies</li> </ul>
<p><u>Expand knowledge integration to drive research, policy, and practice</u></p> <p>Support knowledge integration and meta research (systematic reviews, modeling, decision analysis, etc.) to identify gaps, inform funding, and to integrate epidemiologic knowledge into decision making</p>	<ul style="list-style-type: none"> <li>■ Develop and apply new methods for knowledge integration across basic, clinical, and population sciences</li> <li>■ Make knowledge integration activities integral to decision making by various sectors of society (e.g., medicine, public health, law, urban development etc.)</li> <li>■ Develop metrics of evaluation of success and impact of epidemiologic research</li> </ul>
<p><u>Transform training of 21st century epidemiologists</u></p> <p>Train 21st century epidemiologists with an increasing emphasis on collaboration, multilevel analyses, knowledge integration and translation</p>	<ul style="list-style-type: none"> <li>■ Modify training curricula to adapt a interdisciplinary approach to education by equipping future epidemiologists with practical skills to meet the needs of modern epidemiologic research (collaboration, translation, and multilevel)</li> <li>■ Foster collaborations and shared knowledge between Schools of Public Health and Schools of Medicine</li> <li>■ Train more epidemiologists in implementation and dissemination research</li> </ul>

*(Continued on the following page)*

**Table 1.** Broad recommendations and proposed actions to transform epidemiology for 21st century medicine and public health (Cont'd)

<b>Recommendation</b>	<b>Proposed actions</b>
<u>Optimize the use of resources for epidemiologic studies</u> Develop and design rational cost-effective epidemiologic studies and resources to optimize funding, accelerate translation, and maximize health impact	<ul style="list-style-type: none"> <li>■ Encourage the leveraging of existing resources instead of the creation of new ones</li> <li>■ Integrate information from different settings (e.g., RCTs, HMOs, and cancer registries) to spur new research and validate findings</li> <li>■ Develop initiation and sunset criteria for research studies to maximize return on investment</li> <li>■ Establish novel funding mechanisms that encourage multidisciplinary collaboration and translational research</li> <li>■ Leverage disease-specific funding resources across funding agencies to build basic cross-cutting epidemiologic capacity</li> </ul>

Abbreviations: RCTs, randomized controlled trials; HMO, health maintenance organizations.

innovate at reasonable cost, and failure to identify research questions with the greatest merit" (3).

In the past few years, translational research (4) has sought to accelerate the movement of scientific discoveries into practice and improved health outcomes. However, the main focus of translational research remains, by and large, on basic science to clinical applications (bench to bedside). Epidemiology and other population sciences can be integrated into a full translational framework that spans scientific discoveries through improved population health (4). Within this framework, Lam and colleagues have identified 4 drivers that are increasingly shaping the field of epidemiology: interdisciplinary collaboration, multilevel analysis, emergence of innovative technologies, and knowledge integration from basic, clinical, and population sciences (5). Epidemiology can be a key translational discipline for addressing questions of current great societal importance, such as the economics of health services, the aging of our population, the growing burden of common chronic diseases, the persistence of health disparities, and global health. The translational impact of epidemiology similarly must be achieved in an era of greater consumer awareness, open access to health, and other types of information and enhanced communications, via the web, mobile technologies, and social media.

In 2012, the National Cancer Institute (NCI) initiated a conversation aiming to shape the future of cancer epidemiology and to establish priorities for action (6). Web-based blog posts, several commentaries (5–11), online dialogue using social media (@NCIEpi #trendsinepi on Twitter), and an interdisciplinary workshop (12) informed the proposals presented herein. Table 1 outlines 8 broad recommendations with proposed actions targeted to funding agencies, professional societies, and the research community. Many of these actions already feature prominently in epidemiologic research but a more

systematic approach will be needed to increase the impact of epidemiology in the 21st century. Although the recommendations presented here are focused on cancer epidemiology, we believe they apply to the whole field of epidemiology.

### Recommendation 1: Extend the Reach of Epidemiology

The imperatives of the 21st century require epidemiology to extend its reach beyond the historical perspective on etiology to embrace the continuum of early detection, treatment, prognosis through survivorship, and to become more effective in translating scientific discoveries into individual and population health impact (13). Epidemiology in academic institutions has traditionally focused on advancing discoveries, whereas epidemiology in public health and healthcare settings focuses on disease control and program implementation and evaluation. In cancer, cohort studies increasingly try to assess factors that impact natural history, response to interventions, and long-term survivorship (14). Along the full translational continuum (4), most epidemiologic research, however, still focuses on etiology and replication/characterization of the findings (4). Funding agencies and research institutions need a more balanced epidemiologic portfolio including evaluation of interventions to develop evidence-based policies and guidelines, implementation strategies of applications in healthcare decisions and population health policy, and evaluation of impact, including benefits and harms of interventions in the "real world." For example, as epidemiology has uncovered strong associations between tobacco and mortality from various diseases (15), it should increasingly focus on developing, implementing, and evaluating pharmacologic, behavioral, policy, and environmental interventions.

Moving from observation and discovery to the development and evaluation of interventions will require a better integration of clinical and community trials with large scale epidemiologic studies (3). As randomized clinical trials face increasing challenges due to expense, complexity, and nonrepresentativeness, it would be cost-effective and efficient to embed trials into preexisting epidemiologic registries such as large scale cohort studies. These trials can relatively easily enroll large numbers of subjects at relatively low cost. In Scandinavia, there are examples of trials that have already been successfully integrated into preexisting registries or administrative databases, often at low marginal cost (3). Moreover, there will be an increasing need to integrate observational epidemiologic studies into the NCI clinical trials infrastructure. Finally, epidemiologic cohort studies can be cultivated for translational evaluation research, especially in the development and validation of biomarkers (10).

To extend the impact of epidemiology on translational efforts, epidemiologists need to become even more effective in team science (16) and translational research collaboration (17) as well as address multilevel determinants of diseases ranging from social and environmental determinants to biologic and molecular pathways and their interactions (18). Critical to this success is an enhanced effort by funding agencies and the research community to reward interdisciplinary and translational research. As such, the real value of epidemiology resides in informing both discovery research and translational research and embodying a broad perspective on the multilevel origins of disease and an appreciation for the need to apply incremental knowledge to advance population health (19).

### **Recommendation 2: Transform the Practice of Epidemiology**

Epidemiology has traditionally involved single teams with proprietary control of their data and specimens, which they use effectively to publish and garner additional funding. The inner workings of protocols and analyses are typically invisible to outsiders and raw data rarely became available. This practice can adversely impact reproducibility, accountability, and efficiency (20). Peer review usually depends on limited information communicated in a short scientific paper. Fragmentation of information and selective reporting are prominent, and published information is difficult to integrate with other studies after the fact. These practices have led to the kind of criticisms mentioned earlier including repudiated or inconsistent findings and studies that offer small incremental knowledge gains (3). The advent of genome-wide association studies has not only shown that reproducible results can be achieved with large enough sample sizes, but that new models of collaboration and data sharing can be developed (21). The time is right to ensure greater credibility of all epidemiologic studies by adopting a reproducibility culture through

greater sharing of data, protocols, and analyses (22–24). Funding agencies can catalyze this transformation, as they are responsible for shaping the incentive system for science. One possibility is that funding can be based in part on the extent to which investigators adopt sharing of data and specimens (25). Scientific journals can contribute to this transformation by making availability of protocols, raw data, and analyses a prerequisite to publication (26). Concurrently, the scientific community can assist by adopting a culture of data sharing and collaboration. Such a culture shift can acquire value in the academic coinage for appointments, promotion, and awards, and is required to propel the field forward into a more consistent realm of scientific credibility.

This transformation has to address potential obstacles, such as legal, ethical, or pragmatic limitations that may not allow full transparency and availability of information in public view. Issues of informed consent restraints, privacy of participants, and the extra effort and resources needed to make data, protocols, and analyses available widely in sufficiently high quality and accessibility should be anticipated (22, 27). These issues are more prominent for studies that were designed in the past and continue data collection and/or analyses, but should be more straightforward to tackle in new studies. Nevertheless, other considerations must still be addressed including potential impact on participation rates and on the quality and types of data participants will be willing to provide.

One can consider multiple levels of transparency in access to information and decide what would be maximally attainable for each study (as suggested in Table 2). At a minimum, registration of datasets should be achievable for all epidemiologic studies, past and future (28). Funding agencies can support pilot studies and expert panels to assess the feasibility, advantages and disadvantages, and ways to optimize reporting. Some efforts would require creating and expanding existing repositories for information, and there is already substantial experience from some scientific fields, for example, microarray experiments. Making data and protocols more accessible will accelerate harmonization of existing datasets, as in the case of collaborative efforts involving consortia, cohort studies, and biobanks (29, 30). Expansion of open access repositories of data and biologic specimens will require partnership among funding agencies, academic institutions, and scientific journals to create more incentives for data sharing, reproducibility, and replication.

### **Recommendation 3: Expand Cohort Studies Across the Life Course Including Multiple Health Outcomes**

Case-control studies, the traditional workhorse of epidemiology, will continue to make strong contributions to the field in the next decade. In particular, these studies can contribute to indepth examinations of patients with specific (and especially rare) cancers.

**Table 2.** Potential registration levels for epidemiologic research

Level	Registration	Comments
1	No registration	Current predominant paradigm; may continue to be common, but novel published results from such studies should be seen primarily as exploratory analyses requiring confirmation
2	Dataset registration	Should be feasible to achieve in large-scale; each dataset registers the variables that it has collected and their definitions; this would allow knowing how many studies with how many participants who have measured variables or markers of interest, instead of guessing what data are available on that marker beyond what has been published
3	Availability of detailed data	Individual-level (raw) data are made available; this practice may be subject to policy/consent/privacy constraints for past studies and their data; easier to anticipate and encourage in the design of future studies
4	Availability of data, protocols, and analyses codes	Optimal ability to evaluate the reproducibility of analyses, to maximize the integration of information across diverse studies, and to allow improvements on future studies based on exact knowledge of what was done in previous studies
5	Live streaming of analyses	Investigators not only post all their data and protocols online, but analyses are done and shown in realtime to the wider community as they happen. Live streaming can be coupled with crowdsourcing of analyses across large communities of analysts

Nevertheless, with increasing interest in early antecedents of disease and prediagnostic risk factors and biomarkers, large scale prospective cohort studies for disease etiology and outcomes will become increasingly important (and will undoubtedly include nested case-control components). Such studies should be conducted in informative populations, apply validated methods to measure genetic and environmental influences, and include prediagnostic data and biologic samples. In cancer risk cohort studies, organ-specific incidence remains a main outcome of interest for discovering etiology, but other outcomes can be studied as well. First, with advances in molecular tumor classification, we are distinguishing among cancer subtypes by means other than histopathology. Second, the expanding list of recognized precursors (e.g., colon polyps and Barrett esophagus) can provide insight because they occur years or decades before the development of cancer and progression to cancer is highly variable. Third, many etiologic studies are expanding to include treatment and outcome information to allow the evaluation of response to interventions and long-term survival. These efforts complement new and ongoing cancer patient cohorts designed to collect epidemiologic, clinical, genomic, and detailed treatment information after a cancer diagnosis (14).

Ideally, the cohort study should collect information using a life course approach with documented medical histories and exposure information and appropriate biologic tissue collection. Assembling a cohort with these key features is expensive and difficult within the United States health care system (31). In response, NCI and other

research organizations have created approximations to a singular cohort by developing a consortium of multiple cohorts of more than a million people followed for many years (32). In addition, efforts are underway to build cohorts within existing medical care delivery systems by linking epidemiologic data with electronic health records. Cohort studies can be conducted as consortia at multiple sites, combinations of existing ongoing studies, a single large site system, or centralized approach, such as the one used by the United Kingdom Biobank, which completed recruitment of more than half a million participants between 2007 and 2010 (33). Given the existence of many ongoing cohort studies, serious considerations need to be given to mapping and registering all existing prospective cohorts worldwide, harmonizing efforts in data collection and analyses, and expanding current disease-specific studies to include multiple outcomes and to incorporate early life exposures and prediagnostic information. Critical issues for success include collaboration and sharing, modern recruitment structures that facilitate outcome determination, using comprehensive and flexible information technology, automated biologic specimen processing, and broad stakeholder engagement (31). Better coordination and collaboration in funding by disease-specific research agencies will be needed.

Cancer epidemiology is unusual because of the opportunity to work with two genomes, the germline genome that can be used to understand susceptibility to specific cancers and the somatic genome of the cancers can sometimes be used to understand the exposures that gave rise to the cancer by using mutational fingerprints of exposures, mutational determinants of tumor

progression and recurrence, as well as drug sensitivity and resistance. Flagship projects such as the Cancer Genome Atlas have been mostly conducted on anonymized tumor samples (34, 35). Completing the life course approach by using tumor samples from cases of cancer that arise within cohort studies offer the opportunity to study the prediagnostic predictors of both cancer incidence and survival.

#### **Recommendation 4: Develop, Evaluate, and Use Novel Technologies Appropriately**

New technologies and platforms of biomarker measurement continuously become available for incorporation into epidemiologic studies. Examples include genomic, proteomic, metabolomic, noncoding RNA, epigenomic markers, mitochondrial DNA, telomerase platforms, infectious agent markers and microbiota, and immune marker profiles. Similarly, a wide array of environmental measurements using increasingly sophisticated sensor technologies may be measured in blood or other tissues as well as incorporated into portable devices and mobile phones (9, 36, 37). Exploring the potential of the "exposome" may provide a way for assessing the impact of multiple exposures on key internal metabolic processes also using new lab-based technologies (38). It is premature to predict how these approaches will evolve in practice, but techniques for inexpensively sampling a wide array of exposures offer great conceptual appeal. Likewise, we cannot anticipate what new platforms will be available and ready for prime time even in a few years from now, but measurement capacity is likely to continue expanding at a rapid pace. What should be anticipated, however, is the need for careful attention to the proper collection, sampling, processing, and storage of biologic specimens to be interrogated with these evolving technologies, and the development of principles for their optimal use in epidemiologic studies of all types. This need is particularly acute for cohort studies that collect biologic samples today, but may assay these samples many years in the future using measurement platforms that were unknown at the time of sample collection.

Analytic methods for these platforms need to evolve and may need to account for platform-specific peculiarities as well as study design issues. An even greater challenge is how to integrate multiple platforms within the same analysis. These platforms are likely to offer complementary information, but may also have redundancies that need to be avoided. A series of carefully designed studies can move from proof-of-concept to wide-scale validation and successful application of these new technologies. As the possibilities for false leads and dead ends increase exponentially with each new measurement platform, methodologic work is essential in evaluating any technology's analytic performance, reproducibility, replication, disease associations, ethical and legal issues, and clinical use (26).

#### **Recommendation 5: Integrate Big Data Into the Practice of Epidemiology**

The unquestionable reality of 21st century epidemiology is the tsunami of data spanning the spectrum of genomic, molecular, clinical, epidemiologic, environmental, and digital information. The amalgamation of data from these disparate sources has the potential to alter medical and public health decision making. Nevertheless, we currently do not have a firm grasp on how to systematically and efficiently tackle the data deluge. In 2012, the U.S. government unveiled the "Big Data" Initiative with \$200 million committed to research across several agencies (39). Epidemiologists have traditionally been involved in the collection and analysis of large data sets, and therefore should play a central role in directing the use of financial resources and institutional/organizational investment to build infrastructures for the storage and analysis of massive datasets. Critical to the implementation of big data science is the need for high-quality biomedical informatics, bioinformatics, and mathematics and biostatistics expertise.

The development of systematic approaches to robustly manage, integrate, analyze, and interpret large complex data sets is crucial. Overcoming the challenges of developing the architectural framework for data storage and management may benefit from the lessons learned and the knowledge gained from other disciplines (40). Adaptation of technological advancements like cloud-computing platforms, already in use by private industries (e.g., Amazon Cloud Drive and Apple iCloud), can further facilitate this virtual infrastructure and transform biomedical research and health care (41). The tasking challenges for integration of multiscale data to promote progress in research lies more in the realm of bioinformatics and in the unwieldy and politically charged details related to data sharing (e.g., data sovereignty, buy-ins from stakeholders, see Recommendation #2) and to adopt standards and metrics that can cross studies and disciplines. As we write this commentary, the National Institute of Standards and Technology (NIST) is sponsoring the "Cloud Computing and Big Data Workshop" precisely to deliberate on some of these pressing challenges (42). For data acquired from disparate sources, harmonization of definitions can be a challenge. The epidemiology community and funding agencies can integrate the insights gained from this NIST workshop toward better integration of big data science in future epidemiologic studies.

#### **Recommendation 6: Expand Knowledge Integration to Drive Research, Policy, and Practice**

With data-intensive 21st century epidemiology, there is a need for a systematic approach to manage and synthesize large amounts of information (43). Knowledge integration is the process of combining information or data from many sources (and disciplines) in a systematic way to accelerate translation of discoveries into population

health benefits. Knowledge integration also seeks to achieve the effective incorporation of new knowledge in the decisions, practices, and policies of organizations and systems (13). As illustrated by Ioannidis and colleagues in this issue, knowledge integration involves 3 interconnected components (8). First, knowledge management is a continuous process of identifying, selecting, storing, curating, and tracking relevant information across disciplines. Second, knowledge synthesis is a process of applying tools and methods for systematic review of published and unpublished data using *a priori* rules of evidence, including systematic reviews and meta-analysis. In addition, decision analysis and modeling can provide valuable additional synthesis tools to guide policy actions and clinical practice, even with disparate observational and randomized controlled trial (RCT) data (8). Third, knowledge translation uses synthesized information in stakeholder engagement and in influencing policy, guideline development, practice, and research. Moreover, conducting meta-research (or research on research) analyses can aid in understanding evidence across research fields and can reveal patterns of study design, reporting, and biases (20).

A current limitation of knowledge integration is that researchers rely heavily on published literature, which tends to overly report positive associations due to selective reporting and other biases (44). Furthermore, raw data are rarely available to incorporate with the existing published results to uncover true associations. Ioannidis and colleagues (8) outline future suggestions for knowledge integration that may diminish these biases. In knowledge management, there is a need for improved methods for mining published and unpublished data; registration of studies, datasets and protocols; availability of raw data and analysis codes; and facilitation of repeatability and reproducibility checks. With regard to knowledge synthesis, consortia that run analyses prospectively should optimize collaboration and communication. Prospective stakeholder engagement at the outset of a study is essential for knowledge translation (8).

Funding agencies and journals can also help knowledge integration efforts. They can facilitate the development and use of online tools and databases to capture published and unpublished data, datasets, studies, and protocols from funded epidemiologic studies. Journals can promote the publication of relevant "null results" to minimize publication bias, as the *Cancer Epidemiology, Biomarkers & Prevention* already does. The NIH and other funding agencies can also capitalize on the process of knowledge integration to systematically track existing research and resources to identify gaps and redundancies to guide future funding.

#### **Recommendation 7: Transform Training of 21st Century Epidemiologists**

Academic training in modern epidemiology requires a problem-solving, action-oriented approach. Traditional-

ly, epidemiologic investigations tend to end with the discovery of risk factors, and leave the translation of that research to others (45). There is a need to shift from epidemiologic research that is etiologic to that which is applied with a focus on innovation and translation (46). Ness (47) has further outlined a toolbox of evidence-based creativity programs to be incorporated into every epidemiology curriculum.

Core training of the next generation of epidemiologists should offer skills in integrating biology and epidemiology into studies of etiology and outcomes, mastering sufficient quantitative skills, understanding new quantitative methods, and integrating rapidly evolving measurement platforms (48). The epidemiologist of the 21st century will need deeper immersion in informatics and emerging technologies, as such skills are critical to appropriately leverage and interpret increasingly dense biologic, clinical, and environmental data across multiple sources and platforms.

At the same time, there is a need to reorient the training of practicing epidemiologists toward implementation and dissemination research. The training curriculum must be modified to adapt an interdisciplinary approach to graduate and postdoctoral education by equipping future epidemiologists with practical skills to meet the needs of modern epidemiologic research in collaboration, translation, and multilevel analysis (17). Training must incorporate concepts of knowledge integration to promote the most effective use of information from many sources to further accelerate translation of scientific discoveries into clinical and public health applications. Likewise, there is a need for integration of epidemiologic concepts into training curricula for clinical and public health practitioners to meet the increasing challenge of translating scientific discoveries into population health benefits (4). Medical schools and schools of public health are beginning to work more closely to create a climate of collaboration and shared knowledge across disciplines that nurtures and rewards team efforts. This could include more encouragement for medical students and clinicians to get training in public health (e.g., Master of Public Health) and for epidemiology students and practitioners to get more exposure to basic and clinical sciences.

#### **Recommendation 8: Optimize the Use of Resources for Epidemiologic Studies**

In an environment of funding limitations and rapid technology advances, funding agencies and the epidemiology community need to optimize their strategies for the most efficient use of data, biosamples, and other research resources. First, we should practice the art of bricolage, a critical attribute of resourcefulness, which refers to the novel use of available resources to construct new forms or ideas, creativity under constraints. Second, there needs to be a fair and transparent process to critically examine the criteria needed to discontinue, extend, or expand existing studies and to permit the funding of new cutting-edge

studies. Some benefits can be achieved by extending existing cohorts to integrate data on multiple health-related endpoints. Optimization of resources include, leveraging biospecimens from existing biobanks, harnessing data gathered from various sources [e.g., health maintenance organizations (HMO), Medicare/Medicaid, and cancer registries], linking and mining information from electronic health records, randomized clinical trial networks, as well as other databases (e.g., census bureau) to conduct research, test novel hypotheses, and discover novel exposures. For example, to characterize the natural history of human papilloma virus-associated carcinogenesis, molecular epidemiologists can capitalize on the samples stored in cervical cytology biobanks (49). Patient-provided data and health information can be collected and delivered, respectively, within an existing health care system (3). The Moffitt Cancer Center's *MyMoffitt* Patient portal, for example, represents one archetype of this future approach (50). Current collaborations with the HMO Research Network can be encouraged, enhanced, and incentivized to conduct population-based research on a multitude of health-related outcomes (51). Investigators may expand their interest across the boundaries of different disease-specific endpoints and diverse biologic/genomic exposures (e.g., to include, stress and social variables) while keeping in mind the translational value of the research question (4, 5). As outlined in Recommendation #6, a robust knowledge integration process can be used to determine how best to allocate resources.

Optimizing resources for epidemiologic research will require the direct involvement of funding agencies to serve as active liaison with researchers to improve efficiency in the research process, communication, and management. The overarching push for epidemiology to more collaborative, interdisciplinary, and translational research also requires novel funding mechanisms and enlightened study review teams. Alternative avenues need to be explored to provide investigators with the incentives to abandon nonyielding research courses without causing disruption to their academic career and funding situation.

## References

1. Koplans JP, Thacker SB, Lezin NA. Epidemiology in the 21<sup>st</sup> century: calculation, communication, and intervention. *Am J Public Health* 1999;89:1153-5.
2. Greenwald P, Dunn BK. Landmarks in the history of cancer epidemiology. *Cancer Res* 2009;69:2151-62.
3. Lauer MS. Time for a creative transformation of epidemiology in the United States. *JAMA* 2012;308:1804-5.
4. Khoury MJ, Gwinn M, Ioannidis JP. The emergence of translational epidemiology: from scientific discovery to population health impact. *Am J Epidemiol* 2010;172:517-24.
5. Lam TK, Spitz M, Schully SD, Khoury MJ. "Drivers" of translational cancer epidemiology in the 21st century: needs and opportunities. *Cancer Epidemiol Biomarkers Prev* 2013;22:181-8.
6. Khoury MJ, Freedman AN, Gillanders EM, Harvey CE, Kaefer C, Reid BC, et al. Frontiers in cancer epidemiology: a challenge to the research

## Concluding Remarks

The 8 broad recommendations and corresponding proposed actions presented here are intended to transform cancer epidemiology by enhancing transparency, multidisciplinary collaboration, and strategic applications of new technologies. The recommendations apply more broadly to the field of epidemiology, and should lay a strong scientific foundation for accelerated translation of scientific discoveries into individual and population health benefits. Clearly, more details are needed to address the opportunities and challenges that permeate each of these recommendations requiring further deliberation by the scientific and consumer communities. We invite ongoing conversation on how to strengthen the future of epidemiology using our cancer epidemiology matters blog (7).

## Disclosure of Potential Conflicts of Interest

G.S. Ginsburg has a commercial research grant from Novartis. No potential conflicts of interest were disclosed by the other authors.

## Authors' Contributions

**Conception and design:** M.J. Khoury, T.K. Lam, J.P.A. Ioannidis, J.E. Buring, S.J. Chanock, R.N. Hoover, M.S. Lauer, D. Seminara, T.R. Rebbeck, S.D. Schully

**Development of methodology:** M.J. Khoury, T.K. Lam, D.J. Hunter  
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** M.J. Khoury, T.K. Lam, R.A. Hiatt, S.D. Schully  
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** M.J. Khoury, T.K. Lam, J.P.A. Ioannidis, P. Hartge, S.J. Chanock, R.A. Hiatt, R.N. Hoover, B.S. Kramer, T.R. Rebbeck  
**Writing, review, and/or revision of the manuscript:** M.J. Khoury, T.K. Lam, J.P.A. Ioannidis, P. Hartge, M.R. Spitz, J.E. Buring, S.J. Chanock, R.T. Croyle, K.A. Goddard, G.S. Ginsburg, Z. Herceg, R.A. Hiatt, Robert N. Hoover, D.J. Hunter, B.S. Kramer, M.S. Lauer, J.A. Meyerhardt, O.I. Olopade, J.R. Palmer, T.A. Sellers, D. Seminara, D.F. Ransohoff, T.R. Rebbeck, G. Tourassi, D.M. Winn, A. Zaubler, S.D. Schully  
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** T.K. Lam, S.D. Schully

## Acknowledgments

The authors thank Martin L. Brown for contributing to the NCI's "Trends in 21st Century Cancer Epidemiology" workshop (<http://epi.grants.cancer.gov/workshops/century-trends/>) and are grateful for the comments received on the NCI's "Cancer Epidemiology Matters Blog" (<http://blog-epi.grants.cancer.gov/>).

Received February 7, 2013; accepted February 17, 2013; published OnlineFirst March 5, 2013.

community from the Epidemiology and Genomics Research Program at the National Cancer Institute. *Cancer Epidemiol Biomarkers Prev* 2012;21:999-1001.

7. National Cancer Institute. Cancer Epidemiology Matters Blog. [cited March 9, 2013]. Available from: (<http://blog-epi.grants.cancer.gov/>).
8. Ioannidis JP, Schully SD, Lam TK, Khoury MJ. Knowledge integration in cancer: current landscape and future prospects. *Cancer Epidemiol Biomarkers Prev* 2013;22:3-10.
9. Verma M, Khoury MJ, Ioannidis JP. Opportunities and challenges for selected emerging technologies in cancer epidemiology: mitochondrial, epigenomic, metabolomic, and telomerase profiling. *Cancer Epidemiol Biomarkers Prev* 2013;22:189-200.
10. Ransohoff DF. Cultivating cohort studies for observational translational research. *Cancer Epidemiol Biomarkers Prev* 2013; 22:481-4.



11. Lam TK, Schully SD, Rogers SD, Benkeser R, Reid BB, Khoury MJ. Provocative questions in cancer epidemiology in a time of scientific innovation and budgetary constraints. *Cancer Epidemiol Biomarkers Prev* 2013;22:496–500.
12. National Cancer Institute Workshop. Trends in 21<sup>st</sup> century epidemiology: from scientific discoveries to population health impact; 2012 [cited 2012 Dec 12–13]. Available from: (<http://epi.grants.cancer.gov/workshops/century-trends/>).
13. Best A, Hiatt RA, Norman CD. Knowledge integration: conceptualizing communications in cancer control systems. *Patient Educ Couns* 2008;71:319–27.
14. Elena JW, Travis LB, Simonds NI, Ambrosone CB, Ballard-Barbash R, Bhatia S, et al. Leveraging epidemiology and clinical studies of cancer outcomes: recommendations and opportunities for translational research. *J Natl Cancer Inst* 2013;105:85–94.
15. Jha P, Ramasundarahettige C, Landsman V, Rostron B, Thun M, Anderson RN, et al. 21<sup>st</sup> -century hazards of smoking and benefits of cessation in the United States. *N Engl J Med* 2013;368:341–50.
16. Hall KL, Feng AX, Moser RP, Stokols D, Taylor BK. Moving the science of team science forward: collaboration and creativity. *Am J Prev Med* 2008;35:S243–9.
17. Hiatt RA. Epidemiology: key to translational, team, and transdisciplinary science. *Ann Epidemiol* 2008;18:859–61.
18. Lynch S, Rebbeck TR. Bridging the gap between biological, individual, and macro-environmental factors in cancer: a multi-level approach. *Cancer Epidemiol Biomarkers Prev* 2013;22:485–95.
19. Zoghbi HY. The basics of translation. *Science* 2013;339:250.
20. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
21. Hunter DJ. Lessons from genome-wide association studies for epidemiology. *Epidemiology* 2012;23:363–7.
22. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: practices and perceptions. *PLoS ONE* 2011;6:e21101.
23. Guttmacher AE, Nabel EG, Collins FS. Why data-sharing policies matter. *Proc Natl Acad Sci U S A* 2009;106:16894.
24. Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, et al. Prepublication data sharing. *Nature* 2009;461:168–70.
25. Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. *Science* 2011;334:1230–2.
26. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. *PLoS ONE* 2011;6:e24357.
27. Kaye J. The tension between data sharing and the protection of privacy in genomics research. *Annu Rev Genomics Hum Genet* 2012;13:415–31.
28. Ioannidis JP. The importance of potential studies that have not existed and registration of observational data sets. *JAMA* 2012;308:575–6.
29. Fortier I, Burton PR, Robson PJ, Ferretti V, Little J, L'Heureux F, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol* 2010;39:1383–93.
30. Harris JR, Burton P, Knoppers BM, Lindpaintner K, Bledsoe M, Brookes AJ, et al. Toward a roadmap in global biobanking for health. *Eur J Hum Genet* 2012;20:1105–11.
31. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K, Kramer BS, et al. New models for large prospective studies: is there a better way? *Am J Epidemiol* 2012;175:859–66.
32. National Cancer Institute. Epidemiology and Genomics Research Program. Cohort Consortium. Available from: <http://epi.grants.cancer.gov/Consortia/cohort.html>.
33. Collins R. What makes UK Biobank special? *Lancet* 2012;379:1173–4.
34. Cancer Genome Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61–70.
35. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
36. Haring R, Wallaschofski H. Diving through the "-omics": the case for deep phenotyping and systems epidemiology. *OMICS* 2012;16:231–4.
37. Paules RS, Aubrecht J, Corvi R, Garthoff B, Kleinjans JC. Moving forward in human cancer risk assessment. *Environ Health Perspect* 2011;119:739–43.
38. Rappaport SM, Smith MT. Epidemiology. environment and disease risks. *Science* 2010;330:460–1.
39. Mervis J. U.S. science policy. Agencies rally to tackle big data. *Science* 2012;336:22.
40. Birney E. The making of ENCODE: lessons for big-data projects. *Nature* 2012;489:49–51.
41. Pechette JM. Transforming health care through cloud computing. *Health Care Law Mon* 2012;2012:2–12.
42. National Institute of Standards and Technology Workshop: cloud computing and big data; 2013 [accessed March 9, 2013]. Available from: <http://www.nist.gov/itl/math/cloud-112912.cfm>.
43. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol* 2010;39:97–106.
44. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 2008;3:e3081.
45. Armenian HK. Epidemiology: a problem-solving journey. *Am J Epidemiol* 2009;169:127–31.
46. Kuller LH. Point: is there a future for innovative epidemiology? *Am J Epidemiol* 2013;177:279–80.
47. Ness RB. Tools for innovative thinking in epidemiology. *Am J Epidemiol* 2012;175:733–8.
48. Spitz MR, Caporaso NE, Sellers TA. Integrative cancer epidemiology—the next generation. *Cancer Discov* 2012;2:1087–90.
49. Arbyn M, Andersson K, Bergeron C, Bogers JP, von Knebel-Doebertitz M, Dillner J. Cervical cytology biobanks as a resource for molecular epidemiology. *Methods Mol Biol* 2011;675:279–98.
50. Fenstermacher DA, Wenham RM, Rollison DE, Dalton WS. Implementing personalized medicine in a cancer center. *Cancer J* 2011;17:528–36.
51. Lieu TA, Hinrichsen VL, Moreira A, Platt R. Collaborations in population-based health research: the 17th annual HMO Research Network Conference, March 23–25, 2011, Boston, Massachusetts, USA. *Clin Med Res* 2011;9:137–40.

# Cancer Epidemiology, Biomarkers & Prevention

## Transforming Epidemiology for 21st Century Medicine and Public Health

Muin J. Khoury, Tram Kim Lam, John P.A. Ioannidis, et al.

*Cancer Epidemiol Biomarkers Prev* 2013;22:508-516. Published OnlineFirst March 5, 2013.

**Updated version** Access the most recent version of this article at:  
doi:[10.1158/1055-9965.EPI-13-0146](https://doi.org/10.1158/1055-9965.EPI-13-0146)

**Cited articles** This article cites 47 articles, 15 of which you can access for free at:  
<http://cebp.aacrjournals.org/content/22/4/508.full#ref-list-1>

**Citing articles** This article has been cited by 17 HighWire-hosted articles. Access the articles at:  
<http://cebp.aacrjournals.org/content/22/4/508.full#related-urls>

**E-mail alerts** [Sign up to receive free email-alerts](#) related to this article or journal.

**Reprints and Subscriptions** To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at [pubs@aacr.org](mailto:pubs@aacr.org).

**Permissions** To request permission to re-use all or part of this article, use this link  
<http://cebp.aacrjournals.org/content/22/4/508>.  
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.