Knowledge integration in cancer: current landscape and future prospects

John P.A. Ioannidis (1-2), Sheri D. Schully (1), Tram Kim Lam (1), Muin J. Khoury (1, 3)

(1) Knowledge Integration Team, Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, NCI, NIH, (2) Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA, and (3) Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia
Address correspondence to: John P.A. Ioannidis, MD, DSc, Stanford Prevention Research Center, 1265 Welch Rd, MSOB X306, Stanford University School of Medicine, Stanford, CA 94305, USA. E-mail: jioannid@stanford.edu

Conflicts of interest: none of the authors has any conflicts of interest to declare.

Word count: 3,500
ABSTRACT

Knowledge integration includes knowledge management, synthesis, and translation processes. It aims to maximize the utility of collected scientific information and accelerate translation of discoveries into individual and population health benefits.

Accumulated evidence in cancer epidemiology constitutes a large share of the 2.7 million papers on cancer in PubMed. We examine the landscape of knowledge integration in cancer epidemiology. Past approaches have mostly used retrospective efforts of knowledge management and traditional systematic reviews and meta-analyses.

Systematic searches identify 2,332 meta-analyses, about half of which are on genetics and epigenetics. Meta-analyses represent 1:86-1:1162 of published articles in various cancer sub-fields. Recently there are more collaborative meta-analyses with individual-level data, including those with prospective collection of measurements (e.g., genotypes in genome-wide association studies); this may help increase the reliability of inferences in the field. However, most meta-analyses are still done retrospectively with published information. There is also a flurry of candidate gene meta-analyses with spuriously prevalent “positive” results. Prospective design of large research agendas, registration
of datasets, and public availability of data and analyses may improve our ability to identify knowledge gaps, maximize and accelerate translational progress or - at a minimum - recognize dead ends in a more timely fashion.
Introduction

Given the rapid expansion of scientific information, there is a critical need to ensure that maximal use is made of the collected data in a most efficient, and unbiased way. “Knowledge integration” describes the processes that aim at effective use of information from many sources for accelerating translation of scientific discoveries into clinical applications, evidence-based recommendations, use in practice, and eventually health benefits for individuals and populations (Figure 1). The term has been used with different definitions and perspectives (1-3), but here we adopt the definition that includes knowledge management (KM), knowledge synthesis (KS), and knowledge translation (KT) (1). These 3 processes can inform the continuum of translational research from discovery to population health impact (1). Knowledge integration engine can drive research progress, especially in data-intensive fields.

Previously, knowledge integration has depended mostly on retrospective efforts of horizon scanning, traditional systematic reviews and meta-analyses, and non-systematic knowledge brokering between stakeholders. These efforts may be inadequate in the current era of rapid accumulation of multi-level information - from molecular to the
macro level of environmental exposures and health system attributes. Revamping the

current knowledge integration processes may help drive the future of cancer
epidemiology across the translational research continuum. Here, we overview the

accumulated experience on knowledge integration with emphasis on cancer
epidemiology. We discuss what methods have been used over time, their strengths and
limitations, and what alternative approaches might lead to more efficient integration of

emerging information.

Published information on cancer

PubMed (search August 27, 2012) lists 2,673,926 articles with the search on
“cancer”, of which over three-quarters (n=2,146,156) are tagged with “human”. Table

1 shows selective subsets that present an overall picture of published evidence for
different types of designs relating to cancer epidemiology. Over 50,000 articles are
identified with the search word “cohort”, and slightly more with “case control”, while
the term “risk” retrieves over 200,000 articles, and “biomarker” just as many. At the
same time there are also over 100,000 clinical trial publications, many of which are
randomized trials. Efforts to synthesize this information are represented by over 6,000
meta-analyses, almost 30,000 systematic reviews, and well over 300,000 non-systematic reviews. Thus most efforts at evidence integration still use subjective opinion and non-systematic methods for data overview and interpretation.

It is difficult to accurately separate the literature on associations from the literature on treatments and interventions—sometimes they intermingled in the same article. Subsets searches in Table 1 should be reviewed with caution given the non-perfect sensitivity and specificity of PubMed searches. However, clearly articles with relevance to treatment far outnumber articles with relevance to prevention across all subsets. Table 1 also provides search results with the string “NOT (trial* OR treatment) to further exclude papers on clinical trials and treatment-related research (which can be either randomized or observational, e.g. predictive tools for treatment response). As shown, studies with “cohort” or “case control” are split between the literature with and without trials/treatment implications.

Knowledge management: studies, data, analyses

KM efforts can take many different forms, depending on whether one is tapping into published information; retrieving unpublished information; developing databases
with raw data; or allowing a live stream of all collected data and analyses (Table 2).

Published information is only a fraction of the total raw data that have been collected for or repurposed for research purposes, and of the analyses that have been formally performed, probed, or contemplated (4).

Efforts of KM for published data concentrate usually at search optimization, curation, cleaning, and harmonization. Published data may often be a selected, even distorted, subset of the whole information chosen based on statistical significance and/or other selection filters (5). Although some journals have begun efforts to publish “null” results (6, 7), these remain under-represented in the published literature. If so, KM targeting past published data may yield largely misleading results. Empirical meta-research evaluations on the credibility of research findings in different research fields and with different methods and technologies may help anchor some credibility estimates. They may help decide whether a field is severely biased that it is a wasteful effort to collect, clean, and use published information. Conversely, the KM process may indicate that a systematic synthesis can indeed yield reliable results.
Unpublished data and analyses results are notoriously difficult to unearth. Some investigators may claim that unpublished data can be ignored, since they have not passed peer-review. However, the seal of peer-review is not a perfect discriminant. Registration of protocols and analyses would have helped to understand the depth of the problem. Nevertheless, publication of these documents is limited. For example, it is well documented that many clinical trials protocols and/or analyses are never published. Of those that are published, half or more of the originally considered outcomes are not reported, while many others are reported with analyses and results that deviate from the originally intended analyses (8-10). For observational epidemiology, these problems are probably more common (11). However, a priori registration of protocols is difficult for such studies, given their exploratory and iterative nature. Instead, it has been argued that registration should focus on study datasets (4, 12), i.e. information on what variables have been collected and measured. This allows an understanding and assessment how many registered datasets could have undergone specific analyses.

Public deposition of raw datasets has attracted increasing attention over time, with several successful efforts for laboratory research, e.g. genomic sequencing databases
and functional databases (13). For microarray, macromolecular, and protein data, most high-impact journals have policies that require delineating some plan of making raw data, protocols, and analysis codes publicly available, routinely or after request to the authors; public deposition of such information is often a prerequisite for publication, but these policies are not necessarily enforced (14,15).

There are several challenges related to deposition of raw datasets (Table 2). Some datasets may be deposited with poor documentation that hinders their usage by an outsider, or may lead to erroneous data readings and misleading inferences. Additionally, some public databases have minimum requirements when depositing data. Even if investigators are required to adhere to data-sharing policies (either from funding agencies or journal requirements), they often enter the minimal amount of required information. There are different modes of data access: open-to-all; cursory approval-based; and access to select investigators passing stringent standards of recognized expertise. Striking a balance between credit and independence is also challenging. Original investigators could (or should) be credited for analyses performed on their data. However, it may also be advisable to keep further analyses separate from them:
subsequent investigators who then use these published data should feel free to repeat
and challenge the original analyses.

Finally, the live stream information model suggests that data, protocols, and
analyses are readily available and visible to a wider circle, even the full public, as they
accumulate, change, and evolve. This practice has been piloted in experiments trying to
replicate the finding of bacteria with arsenic-containing DNA (16). Other fields,
including cancer epidemiology, may also learn from it.

Knowledge synthesis of same-level information

There is a wide variety of KS methods (Table 2), but systematic reviews and
meta-analyses are the most common. The majority of such reviews still depend on
published information. Meta-analyses of published data are popular in many disciplines,
especially those where unadjusted estimates and plain 2x2 tables are convenient. Some
efforts may also be made to retrieve and include unpublished information, but success in
this endeavor varies. For fields where there are many meta-analyses, field synopses
have emerged (17) with simultaneous compilation of tens to millions of meta-analyses on
the same field, as in several examples of applications in human genome epidemiology, e.g. AlzGene, SzGene, and PDGene (18-20).

Other KS efforts involve investigators who control existing primary data. Such collaborative meta-analyses use a central secretariat to collect, query, clean, and synthesize individual-level data or statistics derived from individual-level data analyses procured by the original investigators of each included study. The advantages (potential standardization or harmonization of data and analyses, consistency of adjustments, multivariable models, interactions and other complex models) and disadvantages (cost, effort, inability to fully standardize post-hoc, selective availability of information, political difficulties) of this approach versus meta-analyses of the published literature have been extensively discussed previously (21, 22).

There is increasing interest in collaborative meta-analyses that use prospectively collected measurements from existing studies. This is the dominant paradigm in meta-analyses of genome-wide association studies (23, 24) that have led to a massive increase in the number of discovered genetic variants with strong statistical support (25). These meta-analyses may avoid the potential for selective reporting bias that threatens
collaborative meta-analyses of previously collected data. Consortia working in this framework are common in human genome epidemiology, but less common in other fields. Finally, one can envision prospective meta-analyses, where not only specific measurements but even the primary studies are designed prospectively, with the plan to eventually combine them. Such examples currently exist mostly from randomized trials (26). Nevertheless, the concept can conceivably be applied to future designed case-control studies, cohort studies, and biobanks (27), with prospective standardization of their designs and data collection and analyses procedures (28).

**Landscape of KS methods used in cancer epidemiology**

Table 3 shows the landscape of practiced KS methods for the PubMed subset of “Cancer NOT (trial* OR treatment)”. The large majority of systematic reviews (85.5%) and all but 13 meta-analyses addressed human data. More effort is needed to systematically appraise evidence from animal studies (29, 30) which can be informative and influential for judging biological plausibility and for other preclinical inferences.

The fields of genetics or epigenetics dominate almost a third of the literature. Correspondingly, the same distribution applies to systematic reviews, while these two
disciplines account for about half of the existing meta-analyses. The literatures on biomarkers, hormones, and infectious agents are extensive, but have relatively fewer meta-analyses (N=51-109 in each). Conversely, other concentrations with smaller shares of the total literature instead have as many or more published meta-analyses, in particular smoking, occupational, and nutritional fields.

The number of systematic reviews is 2-3 fold larger than the number of meta-analyses in most areas, except for biomarkers, immune/allergy/asthma, and social/socioeconomic factors, where the ratio is even larger. This may reflect the difficulty of performing quantitative syntheses (e.g. for social and socioeconomic factors with extreme heterogeneity of definitions and measurements) or less established traditions for performing meta-analyses. The ratio of all published articles per published meta-analysis is 556 overall - despite their emerging popularity, meta-analyses are still a small portion of the literature. Moreover, there is large variability in this ratio across different fields. It is smaller (N(all)/N(MA)=89-133) for smoking, occupational, nutritional, and lifestyle areas, modestly high (N(all)/N(MA)=215-350) for alcohol,
social, genetics, carcinogens, and radiation, and very high \( \frac{N(\text{all})}{N(\text{MA})} = 728-1162 \) in
epigenetics, biomarkers, immune factors, hormones, and infectious agents.

A more detailed examination of a sample of meta-analyses published in 1992,
2002, and 2012 shows the evolution of the application of these methods over time.

“Cancer NOT (trial* OR treatment) AND meta-analysis [type]” yields 25 items in 1992,
49 in 2002, and 232 in the first 8 months of 2012 alone. On closer examination, 20 of
the 25 meta-analysis-tagged articles published in 1992 are indeed meta-analyses related
to cancer epidemiology, and the same applies to 44 of 49 tagged articles in 2002 and 50
of the 53 latest indexed articles in 2012.

Besides the geometric increase in the number of published meta-analysis articles
each year, the areas represented have changed over time. The advent of meta-analyses
on genetics and epigenetics is impressive. In 1992, there was only one quantitative
review on leukemia cytogenetics. In 2002, of the 44 meta-analyses, 8 (18%) assessed
genetic variants, 1 (2%) genomic hybridization, and 1 (2%) microarrays. In 2012, of the
50 most recently indexed published meta-analyses, 25 (50%) were on genetic variants, 2
(4%) on epigenetics, and another one on gene-menopause interaction. No other field in
2012 had such staggering rise in the number of meta-analyses (smoking n=3, alcohol n=3, biomarkers n=2, infectious agents n=2, dietary n=2, social n=1, occupational n=1, diagnostic tests n=3, other n=3 among 50 meta-analyses examined).

Moreover, there have been an increasing number of genes and genetic variants examined in meta-analyses over time. All genetic meta-analyses in 2002 focused on specific genes. Conversely, meta-analyses in 2012 included also genome-wide association meta-analyses, consortium analyses examining a number of variants, and field synopses. There is also a discernible change in the types of meta-analyses performed over time, in particular regarding the use of consortium approaches and use of individual-level data (31). In the examined samples, there were 2 meta-analyses using individual-level data in 1992 (among 20), 5 in 2002 (among 44), and 9 among the most recent 50 meta-analyses in 2012. All the meta-analyses with individual-level data in 1992 and 2002 combined information that had been already collected in existing studies. One meta-analysis combined data from publicly available data on microarray experiments, while all the other meta-analyses created collaborative structures where investigators contributed their data and participated in the final manuscripts. These analyses
pertained to nutritional factors (n=4), hormones (n=1), or smoking and alcohol (n=1).

Conversely, in 2012, the 9 meta-analyses using individual-level data targeted a very different set of risk factors: genetic factors (n=6), gene expression data (n=1), biomarkers (n=1) and endometriosis (n=1). The 6 meta-analyses of genetic factors were done by consortia performing with genotype data generated prospectively for the project. The gene-expression meta-analysis used data from publicly available databases, and the other two meta-analyses were done by investigators contributing previously collected data.

Caveats in current meta-analyses in cancer epidemiology

Despite the increase in the number and proportion of meta-analyses with individual-level data over time, these still represent the minority. Most currently published meta-analyses in cancer epidemiology continue to depend on published summary data. Many of these meta-analyses focus on genetic variants, often targeting a single or a few candidate genes and variants thereof. Interestingly, among the 50 published meta-analyses from 2012, twelve were done in China focusing on specific candidate genes from the era preceding genome-wide association studies (GWAS). With
one exception (32), all of these Chinese meta-analyses concluded that the examined
candidate genes are significantly associated with the phenotypes of interest, although
the $p$-values were always very modest. Based on previous experience on candidate gene
associations (33, 34), the credibility of these associations is very low. Another 3 meta-
analyses from China addressed genetic variants previously highlighted from GWAS and
also included primary data that the authors generated in their own sample and claimed
replication of the genetic effects. Including also other fields beyond genetics, overall, 19
of the 50 (38%) meta-analyses in 2012 were from China, and 17 of these 19 concluded
with significant, favorable results. Previous empirical evaluations suggest that studies
from China in different fields have frequent or even ubiquitous “positive” findings (35,
36).

The very high prevalence of “positive” meta-analyses, at the face of what should
be mostly null associations, is worrisome. Apparently automation has allowed the
massive production of potentially unreliable meta-analyses. The problem seems to be
most acute for genetic epidemiology, which carries a lion’s share in currently published
meta-analyses, but may extend also in other disciplines.
Knowledge synthesis: multiple-level information

Besides KS involving the same type of information combined across different studies, KS may also try to synthesize multiple levels of information and/or simulated rather than real data. Cross-design synthesis approaches can combine data from different types of designs and umbrella reviews try to compile information on different aspects of questions of interest, e.g. incidence, prevalence, associations, predictive performance, and clinical treatment effects, if pertinent (37). The IARC monographs combine basic and epidemiological data to arrive at a systematic approach to classification of carcinogens (38). The HuGENet Venice criteria attempt to do this for genetic associations (39) and Boffetta et al. recently proposed a merging of IARC monograph and Venice methods appraising evidence on gene-environment interactions (40).

As knowledge progresses from discovery to health related applications, mixed methods become increasingly used to examine the evidence of validity and utility of the information. Examples of knowledge synthesis using a mixture of methods for more advanced translation steps include the US Preventive Services Task Force documents
for clinical preventive services (41) such as prostate and breast cancer screening (T2 translation stage, “does it work”), the CDC Community Guide for Preventive Services (42) (T3 translation stage, “how does it work in community settings”), and CISNet, a NCI funded consortia that evaluate using modeling and empirical data the impact of different interventions on real-world population outcomes (T4 translation stage), e.g. as in a recent modeling paper on contribution of screening and survival differences to racial disparities in colorectal cancer rates (43).

Knowledge synthesis: meta-research

Meta-research (research on research) may allow obtaining wide views on evidence concerning multiple research questions across one or more fields. It may help understand general patterns of study design, reporting, and biases. For example, meta-research evaluations have documented the problems of selective reporting, and excess significance biases in cancer epidemiology studies and their meta-analyses (44-50). One may list here also efforts to reproduce published results. Such efforts, ranging from “forensic bioinformatics” to “reproducibility checks”, have demonstrated major reproducibility
problems in several research fields such as -omics signatures (51) or preclinical data on drug targets (52).

Knowledge translation: using science to influence research, policy and practice

KM and KS are not sufficient to move promising applications and interventions into practice. KT is a proactive process that involves communicating and disseminating synthesized information to influence policy, guideline development, practice and research across the translation continuum. This is the most “messy” component of knowledge integration; it requires the “buy-in” from stakeholders with different perspectives, e.g. see the recently discussed dissemination and implementation agenda for NIH (53).

Many forces affect the diffusion, adoption and implementation of evidence-based recommendations into policy and practice and often operate independently from KS. They include public and private investments in research and development, policy and legal frameworks, oversight and regulation, product marketing, coverage and reimbursements, consumer advocacy, provider awareness, access, and health services development and implementation. Deverka et al. (54) demonstrated that for cancer genomic applications, different stakeholders hold disparate views of the synthesized
knowledge presented to them. For example, payers generally require a higher level of evidence of clinical utility than genomic researchers or test developers. Issues around differential access and implementation may contribute to the “lost in translation” phenomenon (55).

One aspect of KT may involve convening stakeholders around KS to address differences in evidentiary thresholds that drive decision making. This convening function of “knowledge brokering” links researchers and policy makers to facilitate interactions and forge partnerships to use evidence from existing knowledge and define areas for future research. In fields with rapidly changing landscape such as cancer genomics, KT knowledge brokering may need more robust proactive stakeholder engagement earlier in the decision-making process rather than later (1).

Conclusions and future prospects

The landscape of knowledge integration in cancer epidemiology has changed substantially over time and it continues to change. New methods are used more widely for managing, synthesizing, and translating information. Table 4 summarizes some possibilities that may enhance knowledge integration efforts in the future.
KM may benefit from more proactive steps rather than waiting to handle selective reported, fragmented published data. Registration of observational datasets (4,56), more systematic availability of raw data and analysis codes (57,58), facilitation of repeatability and reproducibility checks (59), building a replication culture (60), and even consideration of live streaming of information may accelerate science, allow prompt recognition of false positives and dead ends, and facilitate translation of interesting observations that can be repeated and validated. The optimal way and implementation mode to achieve these changes needs carefully study.

In knowledge synthesis, the paradigm of large-scale international collaboration with prospective data collection has become dominant in human genome epidemiology (29), but it can also permeate more broadly many other fields. New epidemiologic studies and biobanks (25) may also be designed with the outlook that they will form part of a larger prospective network, rather than isolated proprietary experiments. For many diseases and sub-fields, it is possible that there several consortia with overlapping purposes may continue to co-exist. This may not necessary be a drawback, as this may promote competition and independent replication across consortia. Regardless, it is
important to have wide views of the information that is or could become available. This would help avoiding having to fund yet another study when there are hundreds available that can easily address the same question (4); or to prioritize studies and data collection in fields where the wider map of global evidence seems to have a dearth of data.

Finally, knowledge translation could benefit by wider spread and brokering of sound evidence from more reliable KM and KS efforts. It may be easier to set upfront goals, expectations, and rules of engagement and make all stakeholders aware of them, rather than wait for debates to be settled post-hoc. A fascinating aspect of science is that not everything can be anticipated, but this does not mean that we cannot try to have some upfront planning and more transparency in protocols, analysis plans, and results.
Figure legend

Figure 1: The central role of knowledge integration in driving translational research.

Table 1. Published papers in cancer literature

<table>
<thead>
<tr>
<th></th>
<th>PubMed</th>
<th>Treatment</th>
<th>Prevention</th>
<th>Not (trial or treatment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>2,673,926</td>
<td>1,360,697</td>
<td>208,187</td>
<td>1,295,958</td>
</tr>
<tr>
<td>+Animal</td>
<td>481,080</td>
<td>206,009</td>
<td>40,891</td>
<td>269,653</td>
</tr>
<tr>
<td>+Cell</td>
<td>1,118,600</td>
<td>510,286</td>
<td>66,935</td>
<td>604,187</td>
</tr>
<tr>
<td>+Cohort</td>
<td>53,567</td>
<td>29,808</td>
<td>8,683</td>
<td>22,819</td>
</tr>
<tr>
<td>+Case control</td>
<td>59,248</td>
<td>11,255</td>
<td>5,848</td>
<td>26,973</td>
</tr>
<tr>
<td>+Risk</td>
<td>267,490</td>
<td>158,529</td>
<td>58,002</td>
<td>105,276</td>
</tr>
<tr>
<td>+Biomarker</td>
<td>204,419</td>
<td>91,298</td>
<td>15,675</td>
<td>111,025</td>
</tr>
<tr>
<td>+Clinical trial, type</td>
<td>105,939</td>
<td>93,172</td>
<td>14,807</td>
<td>4,535</td>
</tr>
<tr>
<td>+RCT, type</td>
<td>34,449</td>
<td>31,862</td>
<td>8,047</td>
<td>0</td>
</tr>
<tr>
<td>+Meta-analysis, type</td>
<td>6,406</td>
<td>3,902</td>
<td>1,153</td>
<td>2,332</td>
</tr>
<tr>
<td>+Systematic review, type</td>
<td>28,922</td>
<td>21,398</td>
<td>5,755</td>
<td>6,763</td>
</tr>
</tbody>
</table>
Table 2. Different approaches to knowledge management and knowledge synthesis

Knowledge management

Published data: optimization of search engines, curation and cleaning, harmonization

Unpublished data: detection, registration, cleaning

Deposition of raw datasets in public: documentation, access control, ease of use, credit, independence

Live stream information

Knowledge synthesis

Same-level of information

Systematic reviews of published information

Meta-analyses of published information

Meta-analyses including also retrieved unpublished data

Field synopses of many meta-analyses

Collaborative meta-analyses of previously collected individual-level data
Collaborative meta-analyses of prospectively collected data from existing studies

Prospective consortia and meta-analyses thereof

Multiple levels of information

Cross-design synthesis and multi-level evidence appraisals

Modeling with real or simulated data

Meta-research (research on research)
Table 3. Systematic reviews and meta-analyses in different fields of cancer (excluding trials and treatment)

<table>
<thead>
<tr>
<th>Search terms</th>
<th>All articles</th>
<th>N(SR)</th>
<th>N(MA)</th>
<th>N(SR)/N(MA)</th>
<th>N(all)/N(MA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene/genome/genetic</td>
<td>268,597</td>
<td>1,999</td>
<td>920</td>
<td>2.2</td>
<td>291</td>
</tr>
<tr>
<td>Epigenetic/methylation/mutation</td>
<td>115,763</td>
<td>497</td>
<td>159</td>
<td>3.1</td>
<td>728</td>
</tr>
<tr>
<td>Immune/allergy/asthma</td>
<td>29,046</td>
<td>107</td>
<td>25</td>
<td>4.3</td>
<td>1162</td>
</tr>
<tr>
<td>Hormone</td>
<td>53,679</td>
<td>148</td>
<td>51</td>
<td>2.9</td>
<td>1032</td>
</tr>
<tr>
<td>Social/socioeconomic</td>
<td>11,531</td>
<td>224</td>
<td>50</td>
<td>4.5</td>
<td>231</td>
</tr>
<tr>
<td>Diet/dietary/nutrition/nutritional</td>
<td>19,549</td>
<td>289</td>
<td>147</td>
<td>2.0</td>
<td>133</td>
</tr>
<tr>
<td>Physical activity/exercise/obesity</td>
<td>8,919</td>
<td>192</td>
<td>74</td>
<td>2.6</td>
<td>121</td>
</tr>
<tr>
<td>Virus/bacteria/infection/infectious</td>
<td>88,881</td>
<td>331</td>
<td>109</td>
<td>3.0</td>
<td>815</td>
</tr>
<tr>
<td>Carcinogen</td>
<td>30,286</td>
<td>201</td>
<td>88</td>
<td>2.3</td>
<td>344</td>
</tr>
<tr>
<td>Radiation</td>
<td>30,124</td>
<td>229</td>
<td>86</td>
<td>2.6</td>
<td>350</td>
</tr>
<tr>
<td>Occupation/occupational</td>
<td>16,839</td>
<td>344</td>
<td>177</td>
<td>1.9</td>
<td>95</td>
</tr>
<tr>
<td>Smoking/smoke/tobacco</td>
<td>20,660</td>
<td>413</td>
<td>232</td>
<td>1.8</td>
<td>89</td>
</tr>
<tr>
<td>Alcohol</td>
<td>17,921</td>
<td>170</td>
<td>83</td>
<td>2.0</td>
<td>215</td>
</tr>
</tbody>
</table>
Based on PubMed searches conducted on August 29, 2012. N(all): total number of articles retrieved; N(SR): number of articles retrieved with type: systematic review; N(MA): number of articles retrieved with type: meta-analysis. The search resulted in 1,295,958 items overall, of which 6,763 were tagged by PubMed as systematic reviews and 2,332 as meta-analyses; when limited to studies tagged as human, there were 941,360 items overall, 5,780 systematic reviews and 2,319 meta-analyses. All searches use as a backbone the search strategy “Cancer NOT (trial* OR treatment)” so as to avoid capturing articles on clinical trials and treatments. Some articles from case-control and cohort studies where treatment or treatment-related questions are discussed would be excluded by this strategy, but this latter group accounts generally for few meta-analysis, e.g. a search with “Cancer AND (trial* OR treatment) AND (case-control or cohort) AND meta-analysis [type]” yields an additional 451 items (besides the 2,332 identified with “Cancer NOT (trial* OR treatment) AND meta-analysis [type]”); in a sample of 20 of the 451, only 10 are related to cancer epidemiology. Conversely, in detailed scrutiny of a sample of 127 items among those identified with the search “Cancer AND (trial* OR treatment) AND (case-control or cohort) AND meta-analysis [type]”, 114 are indeed meta-analyses of cancer epidemiology topics (positive predictive value 90%). Thus, the provided estimate of 2,332 meta-
analysis papers in cancer epidemiology seems to be quite accurate, with <250 falsely-identified papers and roughly equal number of falsely non-identified papers.

All searches in the table, with the exception of Gene/Genome/Genetic and Epigenetic*/Methylation/Mutation, use the string “NOT (Gene OR Genome OR Genetic)”, so as to exclude items that deal primarily with genetics, e.g. immune response genes, or hormone-related genes.
Table 4. Suggestions for the future of knowledge integration

Knowledge management:

- Methods for mining published and unpublished data
- Registration of observational datasets and, when appropriate, protocols
- Availability of raw data and analysis codes
- Facilitation of repeatability and reproducibility checks, replication culture
- Consideration of live stream information

Knowledge synthesis:

- Facilitation of consortia with prospective measurements
- Optimization of multi-consortial space, competition and communication
- Prospective study networks

Knowledge translation:

- Anticipatory rather than post hoc brokering
References:


4. Ioannidis JP. The importance of potential studies that have not existed and registration of observational data sets. JAMA 2012;308:575-6.


18. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE.
Systematic meta-analyses of Alzheimer disease genetic association studies: the

19. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK,
Khoury MJ, et al. Systematic meta-analyses and field synopsis of genetic
association studies in schizophrenia: the SzGene database. Nat Genet
2008;40:827-34.

20. Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S,
Schjeide BM, et al. Comprehensive research synopsis and systematic meta-
analyses in Parkinson's disease genetics: The PDGene database. PLoS Genet
2012;8:e1002548.

21. Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR; International
Meta-analysis of HIV Host Genetics. Commentary: meta-analysis of individual


30. Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol 2010;8:e1000344.


37. Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009;181:488-93.


42. The guide to community preventive services.


53. Glasgow RE, Vinson C, Chambers D, Khoury MJ, Kaplan RM, Hunter C. National Institutes of Health approaches to dissemination and


60. The Reproducibility Initiative.  

Clinical Applications

Knowledge Integration

Discoveries

Evidence-based Recommendations

KM

KT

KS

Population Health Impact

Use in Practice

Figure 1
Cancer Epidemiology, Biomarkers & Prevention

Knowledge Integration in Cancer: Current Landscape and Future Prospects

John P.A. Ioannidis, Sheri Schully, Tram Kim Lam, et al.

Cancer Epidemiol Biomarkers Prev  Published OnlineFirst October 23, 2012.

Updated version  Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-12-1144

Author Manuscript  Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

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.
Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.