The National Cancer Institute Cohort Consortium: an international pooling collaboration of 58 cohorts from 20 countries

Anthony J Swerdlow[1,2], Chinonye E Harvey[3], Roger L Milne[4,5], Camille A Pottinger[3], Celine M Vachon[6,7], Lynne R Wilkens[8], Susan M Gapstur[9], Mattias Johansson[10], Elisabete Weiderpass[11,12,13,14], Deborah M Winn[3]

[1] Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK

[2] Division of Breast Cancer Research, Institute of Cancer Research, London, UK

[3] Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, US

[4] Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia

[5] Centre for Epidemiology and Biostatistics, School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia

[6] Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester MN, US

[7] Mayo Clinic Cancer Center, Rochester, MN, US

[8] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, US

[9] Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA

[10]International Agency for Research on Cancer (IARC), 69372 Lyon CEDEX 08, France

[11] Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway

[12] Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway

[13] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[14] Genetic Epidemiology Group, Folkhälsan Research Center and Faculty of Medicine, University of Helsinki, Helsinki, Finland

Running title

National Cancer Institute Cohort Consortium

Keywords

Cancer, cohort, pooling, international, consortium.

Corresponding author: Professor Anthony Swerdlow, The Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG. Tel: +44 208 722 4012, Fax: +44 208 722 4019, email address: Anthony.swerdlow@icr.ac.uk

Conflict of interest disclosure: No authors have declared any potential conflicts of interest.

Word count: Abstract 189, main text 5,794

Page total: 27

Total Tables: 4

Abstract

Cohort studies have been central to the establishment of the known causes of cancer. To dissect cancer etiology in more detail, for instance for personalized risk prediction and prevention, assessment of risks of subtypes of cancer, and assessment of small elevations in risk, there is a need for analyses of far larger cohort datasets than available in individual existing studies. To address these challenges the US National Cancer Institute Cohort Consortium was founded in 2001. It brings together 58 cancer epidemiology cohorts from 20 countries to undertake large-scale pooling research. The cohorts in aggregate include over nine million study participants, with biospecimens available for about two million of these. Research in the Consortium is undertaken by >40 working groups focussed on specific cancer sites, exposures, or other research areas. More than 180 publications have resulted from the Consortium, mainly on genetic and other cancer epidemiology, with high citation rates. This paper describes the foundation of the Consortium, its structure, governance and methods of working, the participating cohorts, publications and opportunities.

The Consortium welcomes new members with cancer-oriented cohorts of 10,000 or more participants and an interest in collaborative research.

Introduction

For over 60 years, since Doll & Hill (1) and Hammond & Horn (2) demonstrated the ill-effects of smoking, and Case et al (3) demonstrated the hazards of dyestuff manufacture to the bladder, cohort studies have been the principal method to provide definitive evidence of the carcinogenicity to humans of noxious agents, behaviors, and other exposures. Almost every cause of cancer that is known was established by this means (4). Increasingly, however, the need to dissect cancer etiology in more detail, and to pursue risk factors with smaller effects, has meant that even the largest cohorts are proving to be too small – for instance, to assess risks from uncommon exposures, or of uncommon types or subtypes of cancer, or to examine interactions or risks in population subgroups. The advent of molecular genetics, often examining very small elevations or decreases in risk, and of personalized risk prediction, examining risks subdivided across multiple susceptibility strata, have exacerbated the problem.

Existing cohorts, however, have tended to be limited in size by financial and practical constraints. They have frequently been recruited from a restricted subset of the population, for instance teachers (5) or nurses (6, 7), and, with notable exceptions such as the Multiethnic Cohort Study (8), have often been limited in ethnic diversity. Similarly, cancer-focused cohorts, with some exceptions (e.g. (5-7, 9, 10)), have tended to recruit study participants who were at least 35 years of age at baseline, to increase the numbers of incident cancers early in follow-up. Furthermore, study investigators have focussed most of their energy on their cohort's areas of research strength. Consequently, questionnaire data and biospecimens from these cohorts have often remained unused (or if used, greatly underpowered) for analyses that require larger numbers.

One potential solution to this problem has been to assemble ever-larger new cohorts (11, 12), but there are practical and financial limitations to the size of these, and new cohorts take many years before they accrue sufficient follow-up and outcome events. They have also tended to be of limited diversity by age, sometimes by sex, and usually by ethnicity and country. Pooling data from existing cohorts internationally can relatively cheaply and quickly provide cohort data on a very large scale, with diversity of populations and exposures, capitalizing on the investments already made in such cohorts and their follow-up.

Foundation of the Cohort Consortium, and its initial objectives

To address these challenges, and to exploit new opportunities such as advances in methods for molecular genetics, the US National Cancer Institute (NCI), led by Drs. Robert Hoover and Robert Hiatt, convened a meeting in 2001 with the principal investigators of several cancer epidemiology cohorts. The meeting was embedded in a wider strategy that NCI developed in relation to genetic risk factors, cancer risks, and strategies for prevention.

These investigators agreed on the value of pooling projects to enable analyses that no one cohort, nor even a few cohorts in alliance, could do alone, while still allowing the individual cohorts to meet their own scientific objectives. Another contributor to the ethos was the Pooling Project of Prospective Studies of Diet and Cancer (13), started in 1991, which includes many of the studies within the Consortium, and now forms a Working Group within it. The NCI therefore created the Cohort Consortium, initially fostering within it the Breast and Prostate Cancer Cohort Consortium (BPC3) (14), which examined risks of breast and prostate cancers, using a nested case-control design within each of nine cohorts. BPC3 initially assessed risks in relation to germline variants in more than 60 candidate genes related to steroid hormone

metabolism and the IGF pathway (15-17) with greater precision than previously available and then, taking advantage of the rapidly declining costs of genomic assays, conducted a series of genome-wide association studies (GWAS) of these cancers (18, 19). The infrastructure and collaborative trust built up over time were central to the success of these GWAS.

Structure and operations of the Consortium

The Consortium has since grown greatly in size and scope. It now comprises over 200 scientists from multiple institutions internationally, who have agreed to participate in collaborative research efforts and to pool data from their cohorts to address scientific questions that cannot otherwise be addressed through single institutions and cohorts.

Governance and leadership

The Consortium's bylaws (https://epi.grants.cancer.gov/Consortia/bylaws.html) describe its current overall governance structure and the roles and responsibilities of the steering committee, Consortium members, and NCI staff, and are designed to facilitate dynamic, collaborative research, for instance by frequent rotations of the steering committee membership, and by each cohort having one vote on matters related to the Consortium.

The Consortium activities are overseen by a steering committee elected by the members. The steering committee is responsible for policy development, management, and setting the scientific direction. It includes 6-9 principal investigators who reflect the institutional, geographic and gender diversity of the member cohorts, and three NCI ex-officio voting members representing the NCI Divisions of (i) Cancer Control and Population Sciences and (ii) Cancer Epidemiology and Genetics. The chair and chair-elect are elected by the steering

committee. Steering committee members are appointed for up to three three-year terms, and the chair for a one-year term. The overall day-to-day operations and technical support are managed by four NCI Executive staff.

The steering committee holds monthly teleconferences to monitor the progress of ongoing Consortium studies and projects, address problems that arise in those projects, assess proposals for new scientific research projects and working groups, and organize the annual meeting.

Membership

There are currently 58 epidemiological cohorts in the Consortium, representing populations from 20 countries across four continents (North America, Europe, Asia, and Australia). Membership of the Consortium is open, on application, to any cohort study with a minimum of 10,000 participants, in which cancer incidence is accurately assessed and some risk factor data are available. Membership also requires a general commitment to scientific collaboration through contribution of data for pooling research, but for each specific pooling project, individual cohorts decide freely whether or not they wish to participate. Membership is granted after a review of the application and a vote of the steering committee.

Annual meetings

The Consortium members meet annually in person to review progress, gain updates on ongoing and new projects, discuss new research ideas, share study results, and address methodological challenges. The content of the annual meetings is decided by the steering committee and the practical organisation is by the NCI Leadership Staff. The meetings last 2-3 days, and include: talks by some individual working groups, reporting on their progress; talks on scientific methodology or on areas, e.g.

metabolomics, that have potential for use in cohort epidemiology; and interactive sessions about the working of the Consortium. In addition the annual meetings provide a forum for individual Consortium working groups to meet to progress their pooling research. Such working group meetings may be open to all attending the Consortium meeting, or limited to cohorts participating in the working group, for instance because confidential new results are to be shown, including cohort-specific results from cohorts that have not yet published their own data separately.

Working groups and new research projects

The collaborative research in the Consortium is conducted by investigator-led working groups, of which there are currently more than 40. The steering committee assesses proposals monthly for new projects and hence for new working groups. Proposals from non-members are evaluated on the same basis as those from members of the Consortium. The proposals are made on a standard form that captures the rationale, design and proposed funding sources for the intended research, and practical details such as the minimum number of incident cancers that will be required for a cohort to join the particular pooling study. Each proposal is evaluated by the Consortium steering committee, based on the need for prospectively collected data/samples pooled from multiple cohorts, whether there is overlap or duplication of efforts by existing working groups, and the project's potential to make a novel contribution to scientific research and public health. Initial appraisal frequently leads to a request to the proposer for clarification or further information, with the intention to improve the clarity of the proposal to individual cohort members of the consortium and hence improve the chance that these cohorts will take part. The great majority of proposals are either then accepted by the steering committee to be disseminated to members as new working groups, or

directed toward joining an existing working group as a sub-project. Although approval by the steering committee is required for a new working group to be formed, decisions on subsequent research or spin off projects within an existing working group are the responsibility of the group itself.

Once a working group is formed, participation of cohorts in the group is solicited by an email from the steering committee to members, and then direct communication by the lead investigator. The great majority of working groups go on to conduct and publish research successfully, but occasionally one does not, because the research proves to be infeasible (e.g. there are fewer cases than had been anticipated), or too few investigators choose to join, or the researchers have not been able to obtain funding.

Communications

A growing, large scale international consortium of this kindrequires effective and efficient bi-directional communication to maximize collaborations and productivity. A news letter, including information on new projects, is sent to Consortium members monthly. Webinars are used to host virtual meetings of working groups, as well as in conjunction with the in-person annual meetings to include members who cannot attend. Working groups provide regular progress updates on their projects to the steering committee in conference calls, to discuss accomplishments, challenges, lessons learned and suggestions to improve the working of the consortium. They also share their study results through oral and poster presentations at the annual meeting.

A secure portal, with access limited to consortium members, has been created to foster collaboration and information sharing. The portal serves as a repository and

archive for all consortium related activities including concept proposals, operational guidelines, historical documents and best practice documents, and is used by the scientific working groups for research activities. Individual working groups and projects are able to set up private work spaces within the portal with access limited to members of that working group. Several, but not all, working groups have used the portal in varying capacities for sharing updates, study policies and protocols, data, and manuscript versions, among their members.

Current Consortium projects and working groups

Overview

Most Consortium working groups have focussed on specific cancer outcomes (e.g. pancreatic cancer, or ovarian cancer), while others have focussed on particular exposures (e.g. diabetes, or alcohol), or a combination of the two (e.g. circulating carotenoids and breast cancer risk, or vitamin D and risk of rare cancers). Others have related to particular ethnic groups, notably the African American BMI and mortality pooling project, which includes over 200,000 African Americans from seven cohorts. Several working groups (including five at present) have conducted GWAS. A small number have addressed rare cancers, for instance male breast cancer and renal cell cancers, and a few have investigated causes of cancer mortality or general mortality.

Working groups are encouraged to remain open to new members joining provided that the new cohorts meet the minimum analysis-specific criteria: currently 29 of the working groups are open to new members. Reasons why working groups may at some point elect to close to new members are: that they were formed for a particular project and are now completing existing analyses before winding up; that their

funding (especially for laboratory assays) will not cover further cohorts joining; or that adding further cohorts, and the data transfer agreements and data harmonization entailed, would seriously delay analyses already well underway.

Operations and organization

The organization of individual projects and working groups varies. Most are led by the proposing investigator(s). Sometimes a steering committee is formed to manage activities and help with decision-making. The structure of working groups has varied widely. Some, for instance, have extensive written ground rules, publication guidelines, etc., while other are informal with no written rules. Some have continued for 10 years or more, accruing new analyses and purposes over time, while others have been formed for and conducted a specific investigation (e.g. a particular assay), published it, and disbanded. It is a strength of the Consortium that the collaborative arangements for working groups are flexible and in the hands of the members of each working group; additionally since most cohorts are members of many working groups, there is a great deal of expertise and experience from previous working groups to be drawn upon when creating new ones.

The Consortium supports the development of the next generation of cancer epidemiology researchers by encouraging junior investigators to assume leadership and other active roles in managing Consortium studies, for instance as leaders of spin-off projects. Funding for projects varies, as described below.

Details of the working groups and their accomplishments can be found on the Consortium's website (https://epi.grants.cancer.gov/Consortia/members/#members). The Consortium steering committee and NCI provide overall support to the working groups by fostering communication, providing networking opportunities, and

providing limited administrative resources. To facilitate data sharing and data harmonization the NCI has funded the Cancer Epidemiology Descriptive Cohort Database (CEDCD) (https://cedcd.nci.nih.gov/), and two data repositories, the Cohort Metadata Repository (CMR) (https://cmr.nci.nih.gov/) and the Cancer Epidemiology Data Repository (CEDR) (https://epi.grants.cancer.gov/CEDR/), described in Table 1. The CEDCD allows investigators to search for the types of data and biospecimen that were collected by each cohort study, numbers of cohort participants and numbers of incident cancers, in order that investigators planning a consortium project can determine which cohorts have data on the specific variables of interest and potential numbers of subjects and cancers that might be available.

The CMR is a tool that documents data harmonization processes, decisions and harmonized variables across cohorts that are participating in Consortium studies. It does not include individual-level data. Researchers interested in conducting pooled analyses in the consortium can view these metadata, including harmonized variables from specific projects and the specifications used to create them, to determine if they could use already-harmonized data sets for their analyses instead of undertaking a separate time-consuming harmonization effort.

The CEDR is a controlled acess database developed to enable sharing of actual research data, while protecting the privacy of research participants. Researchers can deposit, access, and analyze a variety of individual-level de-identified data, ensuring that use aligns with specific data use agreements and informed consent for each study.

Description of the participating cohorts

The cohorts taking part in the Consortium are shown in Table 2. They are contributing data from more than nine million study participants, and biospecimens, including germline DNA collected at baseline, from approximately 2 million participants.

The majority (60%) of the cohorts are entirely or predominantly US-based, two are from Canada, 13 from Europe, six from south-east Asia, and one each from Mexico, Australia and Iran. Most (66%) studies have less than 100,000 participants (40% less than 50,000 and 14% less than 20,000), 16 have 100,000-300,000 and four have over 500,000 participants. The vast majority of cohorts restricted recruitment to adult ages, with 30 limited to people over age 35 years and 10 of those restricted to ages 50 and older. Three studies, all limited to women, restricted recruitment to younger participants (within the age range 25-55 years). Seventeen studies recruited only women and nine only men. Most studies were predominantly of whites.The vast majority of studies did not select on ethnic origin; exceptions were the Black Women's Health Study and Southern Community Cohort Study (exclusively and predominantly African Americans, respectively), the Mexican American (Mano a Mano) Cohort, the Multiethnic Cohort Study (oversampled on several minority groups), and the Singapore Chinese Health Study. One study, the Radiation Effects Research Foundation Life Span Study, began in the 1950s, three in the 1970s, 18 in the 1980s, 21 in the 1990s, and the remainder since 2000.

Primary objectives.

Twenty-eight cohorts were established to investigate multiple causes of cancer, five of these for specific cancers only; the remainder aimed also to investigate other diseases or causes of death. Eight cohorts had as their primary aim to assess the

influence of diet on cancer and/or other diseases, and ten others focused on vitamins, minerals or other medications as potential preventative agents for cancer. Other cohorts were established to investigate specific risk factors such as radiation, exogenous sex hormones and genetics.

Base populations.

The cohorts were recruited using a variety of sampling strategies and target populations (Table 2). The largest group of cohorts were sampled from geographic regions or countries, several were from occupational groups, and 10 were leveraged from established cancer or cardiovascular randomized clinical or screening trials, by extending follow-up and collecting new exposures and outcomes. To enrich for cancer risk, four cohorts sampled high-risk families or siblings of cancer cases, two enrolled people with precursor conditions, and four sampled from people with known risk factors for cancer. Four cohorts were established from breast screening services.

Outcome ascertainment.

All cohorts in the Consortium follow participants for cancer incidence, with the majority linking to cancer registries and/or using regular follow-up questionnaires or telephone calls. Self-reported cancers are generally validated through medical or pathology record review. To follow for overall mortality or cause of death, some cohorts link to existing data sources such as national, state or county death registries, or medical records; some also use active follow-up for these purposes. Several of the US cohorts including the VITamins and Lifestyle Study (Washington), the California Teachers Study, and the Multiethnic Cohort Study (Hawaii and California), purposely sampled regions covered by the NCI Surveillance,

Epidemiology, and End Results Program and other US population-based cancer registries, in order to obtain cancer incidence and survival data. Non-cancer and non-death outcomes are ascertained via linkages to administrative databases (e.g. for hospital discharges and outpatients, and military records), as well as through direct contact with study participants: for instance the three Shanghai cohorts schedule ongoing in-person visits with cohort participants to obtain repeat measurements and health status over time. Half of the cohorts have gained further data directly from subjects by repeat questionnaires or in person visits, although the frequency varies greatly. Most cohorts collected blood samples from a least a subset of participants either at recruitment, or less often later. Several cohorts have collected tumor samples, for one or more cancer sites, for cancers incident in their cohort.

Further details and contact information about the cohorts in the Consortium can be found on the website (https://epi.grants.cancer.gov/Consortia/members/#members) and in the Cancer Epidemiology Descriptive Cohort Database (https://cedcd.nci.nih.gov/).

Practical considerations and challenges in conducting consortium-based projects

Assembling a new pooling project is a complex and time-consuming endeavor. Consortium investigators have gained valuable experience in organizing such pooling successfully:-

Data acquisition and harmonization

For most analyses conducted within the Consortium, gathering the necessary data from the participating cohorts, harmonizing and analyzing them, has been done at

one or two centers. While the NCI has on occasion conducted harmonization of the database and/or the statistical analyses for specific projects, these tasks have mostly been carried out by members of the working groups themselves, with the procedures then re-used for any further pooled projects within the working group. The Consortium has not in the past had a central data repository; members have generally preferred to provide data sets from their cohorts to the team conducting the analysis, rather than to a central entity. However, NCI has recently instituted a controlled-access repository, the CEDR, described above, where investigators can deposit individual-level de-identified data, to make their data more readily accessible by others and avoid having to respond to repeated data requests.

Data harmonization is frequently highlighted by Consortium investigators as a major bottleneck, and one for which the workload tends to be underestimated. Questions and response categories for a particular exposure often differ between cohorts, and when harmonizing the data the exposure categories may have to be limited to fewer categories in common.

The ease or difficulty of harmonization relates closely to the complexity, and degree of variety between studies, in the questions about, as well as the recording of, the risk factors. Relatively simple risk factors such as height and weight, tend to be relatively easy to harmonize, but even these can be problematic e.g. weights can be at different ages and because recording may have grouped weight into different, potentially incompatible, categories. More-complex variables such as diet and exercise have been more difficult, but nevertheless have been harmonized successfully – for instance, for exercise by converting the questionnaire responses in each study to Metabolic Equivalent of Tasks (METs). Socioeconomic variables can be difficult because they can be based on very different systems (e.g. salary,

education level, place of residence) in different countries. Serial exposures from baseline plus follow-up questionnaires can also be very difficult to harmonize because they reveal inconsistences between the serial responses. Age at menopause is exceptionally problematic to gain unambiguous data about, even within a single study and the more so for harmonization.

New working groups have often harmonized data ab initio, in part because a different set of cohorts may be included than in previous working groups, and new analyses may require new algorithms and variables. To try to avoid each new Consortium working group having to re-harmonize the same cohorts' data for each new project, the NCI developed the Cancer Metadata Repository described above. In 2013 the NCI supported a comprehensive harmonization of a large number of commonly used study variables for cohorts who chose to participate in the Diabetes and Cancer Initiative (n currently=28) and the code book from this harmonization is available in the CMR for other investigators. Greater use of such previously-developed study dictionaries and harmonization codes has the potential to speed up substantially the assembly of future pooled databases. The use of analytical platforms specifically adapted to harmonize and analyse epidemiological data from multiple sources, such as the Maelstrom research open-source software, (20), may further facilitate this process.

Legal issues

Over time, the legal constraints on data and material transfers have become more stringent. This has resulted in more-comprehensive and complex legal agreements (data and material transfer agreements) that stipulate the rules under which transfers of data or materials from individual cohorts are done. These agreements can be arduous and time-consuming to establish even for simple bilateral collaborations, but

they can lead to considerable delays for consortium projects involving dozens of institutions, each with different data sharing policies and often different national regulations, and with limitations imposed by different funders and employers. While the Consortium steering committee has established a pro forma template for data transfers, it remains to be determined if this will help facilitate future Consortium projects. Initiating the establishment of the necessary data and material sharing agreements as early as possible in a new consortium study is key to avoiding downstream delays.

Governance and coordination of individual consortium projects

There are no rules as to how individual pooling consortia assembled under the Consortium umbrella are governed. As noted above, some have been organized by an individual research group, others have a steering committee or multiple research groups leading different tasks or analyses. A common feature is a strong involvement of all investigators who wish to, in directing the research, such that in practice governance is by the participating cohorts, even though one or two groups may be central.

The collaborative nature of Consortium studies is also reflected in how scientific contributions are recognized in the authorship of resulting publications – typically with one to three co-authors from each participating cohort, and sometimes with a writing group of a few consortium members who have undertaken initial data interpretation and drafting of a manuscript.

Financial aspects

Obtaining sustained funding to maintain existing cohorts can itself be a major challenge, but without this, cohorts have limited capacity to contribute to a single, let alone multiple, consortial projects.

Setting up a new Consortium project also involves costs, both for the coordinating group and to a lesser extent for each participating cohort. While “data only” studies have sometimes been realized on a relatively slim or no budget by leveraging existing resources, studies involving assays of biological samples are inevitably costly. Retrieving biospecimens from cohort biobanks, preparation of aliquots, shipping, and performing the relevant assays is expensive, and biomarker-based consortium studies have typically required substantial grants.

When grants are obtained, these frequently include a small amount of support to each participating cohort, for instance, for preparation of the study database. NCI has developed and funded certain targetted initiatives within the Cohort Consortium to address NCI high programmatic and scientific research needs, but usually funding for projects has been obtained from investigator-initiated grants, from NCI or other sources. In the 171 published Consortium papers that stated funding sources, all but 2 cited the US NCI as a source (but this can include support of individual cohorts, not just the overall pooling) and 30 cited other US NIH Institutes, especially the National Institute on Aging [22] and the National Institute of Environmental Health Sciences [14], while the American Cancer Society contributed to funding of 43. Three quarters of the papers cited funding solely from the US, and a tenth by 3 or more countries.

Cohort consortium publications

The Consortium working groups have published 188 articles since the Consortium began in 2003 (https://epi.grants.cancer.gov/Consortia/publications.html). These articles are well cited, with an average of 9 citations each per year. Consortium papers have, on average, been cited over three times as often per year as the average NIH-funded paper in their field. They have particularly contributed on the role of common genetic variants in risk of various cancers(14-19, 21), a greater understanding of multifactorial contributors to common cancers (22, 23), understanding of the etiology of several less-common cancers or cancer subtypes (24-27), analyses with sample sizes sufficient to investigate risk factors for cancer in African Americans (28), and improved understanding of the shape of dose-response for risk factors for all-cause and cancer-specific mortality (23, 27-29).

The vast majority of the papers [159 [85%]] were on cancer epidemiology (Table 3), mainly focussing on associations between genetic (51%) or lifestyle/anthropometric (27%) risk factors and cancer incidence [155] or less often for cancer survival [4] or cancer mortality [2]. These were mainly of nested case-control or cohort design, but in some instances also included non-cohort-based case-control studies to maximize numbers, for instance for genetic analyses (Table 4). The remainder of the papers covered epidemiology of other outcomes including mortality (most commonly all-cause mortality) (4%), biomarkers (2%), mechanisms including DNA methylation and mosaicism (3%), and methodology for pooling projects (7%).

The 94 papers on cancer genetics have included 29 on GWAS, nine of which were meta-analyses of GWAS. Twenty of the GWAS found novel risk loci and eleven confirmed previously reported risk loci. Forty-eight publications investigated candidate genes, individually or in pathways: 25 found significant associations with risk, while 23 had null findings. The remaining 17 studied pleiotropy [5], gene-

environment interactions [8], or addition of genetics in risk models [4]. Studies of anthropometry, smoking and alcohol often found moderate associations with cancer risk, while studies of diet often gave less marked or null findings. A large number of studies examined the associations between pre-diagnostic biomarkers and cancer, most of which found no association with risk.

Although the majority of cohort participants overall are white, there are sufficient non-whites that 6 papers were published solely on Chinese and 2 solely on African Americans, as well as 12 with analyses stratified by ethnic group.

Among the 159 cancer-related articles, the sites most frequently studied were breast (26%), prostate (22%) and pancreas (18%), but many papers also included less common sites (Table 3). In total, 67 of the papers were on cancers defined by the US National Cancer Institute as 'rare' (<15 cases per 100,000 per annum) and many of common cancers included subdivisions that require large numbers e.g. 17 papers on breast cancer subdivided by hormone receptor status, and papers on in-situ breast cancer and Type II endometrial cancer. The exposures analysed were in general relatively common (Table 3), but again large numbers enabled investigation of uncommon subsets within these (e.g. there were 2 papers on risks in metabolically healthy obese subject), and of subdivisions by more than one variable e.g. lung cancer risks in relation to vegetable intake and smoking, and liver cancer risks in relation to OC use and oophorectomy. For genetics, the need for large numbers was to enable the search for the relatively small elevations in relative risk that are typically present for individual SNPs , thus necessitating pooling even for analyses of common cancers.

Pooling successfully enabled analyses of large numbers of events, allowing greater power even for rare events than would have been possible in a single study. The average number of studies included in a manuscript was 10 and the average number of events was 7,026. There was a tendency for the publications on rarer cancers to include a larger number of studies and a smaller average sample size.

The average number of authors listed on the papers was 44, reflecting the team effort required for projects that combine studies; less than 15% of the papers had 15 or fewer authors. The Cohort Consortium offered opportunities to many researchers, including many junior investigators. One hundred and eighteen unique investigators were first authors on the 188 papers.

Conclusions

The Cohort Consortium has come a long way from a group of investigators from nine like-minded cohort studies, all but two from the US, in 2001, to a collaboration now of 58 cohorts from 20 countries. We have learned that cohort investigators are willing and eager to engage in large scale pooling projects, easily self-organize, and work productively with each other on studies that have yielded important findings with implications for understanding of etiology, clinical guidelines and public health measures. Over 180 publications have resulted, with productivity across a very wide range of aspects of cancer epidemiology research, as well as providing a forum for discussing and improving methods and best practice for cohort studies and pooling analyses.

Critical to the success of the Consortium have been the provision of a central administrative and coordinating infrastructure by NCI, and the trust between cohort investigators, expectation of reciprocity, and willingness to pool, that have built up

over the years. This is particularly critical because cohort studies are extraordinarily long-term, and individual investigators have often invested decades of work into them, so that the decision whether to join pooling efforts, which may invalidate future solo publications (e.g., on a rare tumour), is not a simple one. Trust has been greatly increased by the facilitative, non-prescriptive, bottom-up ethos of the Consortium, that it is driven by the research ideas of its members, with leadership by different research groups on different projects.

The Consortium has provided an efficient means to bring together multiple diverse cohorts to address scientific questions the cohort investigators could not address on their own. The underlying cohorts in aggregate represent a huge investment, and have taken decades to collect follow-up data. The Consortium has, over a relatively short period produced important added value based on very large numbers.

As a voluntary, investigator-led, collaborative framework, the future directions of the Consortium will depend on the wishes of its members. However, some general comments can be made. The opportunities that the Consortium has exploited, and will exploit in future, depend on funding constraints and incentives, the intrinsic strengths and limitations of the component cohorts, scientific opportunities and priorities, and the research interests of the members. Thus, for instance, there were many more gene-environment interactions, uncommon exposures and uncommon cancers that the Consortium could address in future, although the latter has the difficulty of a lack of funding opportunities, (e.g. site-specific charities for common cancers such as breast tend to be much better resourced than those for rarer tumors). Furthermore, for very rare cancers or subtypes of cancers, even a consortium of this size may contain too few cases to generate precise risk estimates. The exposures that can be investigated with large numbers are limited to those for

which there are sufficient cohorts (and especially large ones) that collected data on these variables and to the level of detail that can be harmonised. Another limitation of the Consortial approach is the workload to involve large numbers of groups and investigators (although this is also a strength, because of the large number, and variety, of experienced investigators contributing to the quality of each publication).

The strengths of consortial research, in particular compared with starting new cohorts, also include the long collective length of follow-up, and therefore the large numbers of incident cancers after recruitment already accrued, for immediate analysis, and the very low cost of combined analysis compared with the cost of initiating new cohorts. It would help to enable this if funding agencies made available funding initiatives aimed specifically at (a) maintenance and continuing follow-up of existing high quality cohorts and (b) large-scale pooled cohort analyses, especially for uncommon cancers, which collectively account for a large burden of mortality and morbidity but are individually difficult to fund.

The Consortium needs to continue to expand, both to increase numbers for analyses of interactions, subgroup analyses and rare tumors and exposures, and to include more non-white populations, to enable larger analyses for these groups for which numbers within the Consortium are currently much fewer than for whites. The Consortium also needs to consider how to encourage and enable research on questions that currently are not addressed because no investigator has initiated investigation of the topic. Much of the 2017 annual meeting was spent discussing which research directions (e.g. rare tumors, rare exposures, gene-environment interactions) should be given priority in the next 5 years, and the steering committee is now formulating plans to address the priority areas.

There is great potential for the Consortium to contribute to future research, with opportunities provided by the wealth and growing volume of data and biospecimens available from the underlying cohorts, the changing landscape of cancer risk factors within and across populations (e.g., the growing worldwide obesity epidemic), and the extraordinary technological advances occuring in the assessment of genetic, metabolomic, and other molecular characteristics. The large sample sizes of the Consortium are advantageous both for discovering novel associations and for validating findings from individual studies. The move towards precision medicine and prevention (30, 31) will need reliable, stable risk estimates from cohort studies in ever-finer subdivisions of the population. As noted above, large cohort-based databases will also be needed to investigate rare cancers, rare exposures, and gene-environment interactions, to stratify risks by tumor subtypes, and to refine population-level and individual-level risk prediction. There are also great opportunities in survivorship research, and in biomarker research. The Consortium welcomes new members with cancer-oriented cohorts of 10,000 or more participants and an interest in collaborative research.

Acknowledgments

The authors acknowledge the great contribution of those who founded the Consortium, and of all the cohorts who have joined it since. They thank the members of the Steering Committee and the Principal Investigators of the cohorts who made comments on the manuscript and gave information about their cohorts.

Reference List

1. Doll R, Hill AB, The mortality of doctors in relation to their smoking habits; a preliminary report. *Br Med J* 1954; 1:1451-1455.
2. Hammond EC, Horn D, The relationship between human smoking habits and death rates: a follow-up study of 187,766 men. *J Am Med Assoc* 1954; 155:1316-1328.
3. Case RAM, Hosker ME, McDonald DB, Pearson JT, Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry *Br J Ind Med* 1954; 11:75-104.
4. Breslow NE, Day NE, Statistical methods in cancer research. Volume II-The design and analysis of cohort studies. *IARC Sci Publ* 1987:1-406.
5. Bernstein L et al., High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). *Cancer Causes Control* 2002; 13:625-635.
6. Hennekens CH et al., Use of permanent hair dyes and cancer among registered nurses. *Lancet* 1979; 1:1390-1393.
7. Colditz GA, Hankinson SE, The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer* 2005; 5:388-396.
8. Kolonel LN et al., A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* 2000; 151:346-357.
9. Rosenberg L, Adams-Campbell L, Palmer JR, The Black Women's Health Study: a follow-up study for causes and preventions of illness. *J Am Med Womens Assoc (1972)* 1995; 50:56-58.
10. Swerdlow AJ et al., The Breakthrough Generations Study: design of a long-term UK cohort study to investigate breast cancer aetiology. *Br J Cancer* 2011; 105:911-917.
11. The Million Women Study Collaborative, Group. The Million Women Study: design and characteristics of the study population. *Breast Cancer Res* 1999; 1:73-80.
12. Collins R, What makes UK Biobank special? *The Lancet* 2012; 379:1173-1174.
13. Smith-Warner SA et al., Methods for pooling results of epidemiologic studies: the Pooling Project of Prospective Studies of Diet and Cancer. *Am J Epidemiol* 2006; 163:1053-1064.
14. Hunter DJ et al., A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat Rev Cancer* 2005; 5:977-985.
15. Beckmann L et al., Comprehensive analysis of hormone and genetic variation in 36 genes related to steroid hormone metabolism in pre- and postmenopausal women from the breast and prostate cancer cohort consortium (BPC3). *J Clin Endocrinol Metab* 2011; 96:E360-367.
16. Gu F et al., Eighteen insulin-like growth factor pathway genes, circulating levels of IGF-I and its binding protein, and risk of prostate and breast cancer. *Cancer Epidemiol Biomarkers Prev* 2010; 19:2877-2887.
17. Canzian F et al., Comprehensive analysis of common genetic variation in 61 genes related to steroid hormone and insulin-like growth factor-I metabolism and breast cancer risk in the NCI breast and prostate cancer cohort consortium. *Hum Mol Genet* 2010; 19:3873-3884.
18. Schumacher FR et al., Genome-wide association study identifies new prostate cancer susceptibility loci. *Hum Mol Genet* 2011; 20:3867-3875.

19. Husing A et al., Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J Med Genet* 2012; 49:601-608.
20. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V, Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. *Int J Epidemiol* 2017; 46:1372-1378.
21. Yeager M et al., Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007; 39:645-649.
22. Barrdahl M et al., Post-GWAS gene-environment interplay in breast cancer: results from the Breast and Prostate Cancer Cohort Consortium and a meta-analysis on 79,000 women. *Hum Mol Genet* 2014; 23:5260-5270.
23. Maas P et al., Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol* 2016; 2:1295-1302.
24. Helzlsouer KJ, Committee VS, Overview of the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers. *Am J Epidemiol* 2010; 172:4-9.
25. Arem H et al., Physical Activity and Risk of Male Breast Cancer. *Cancer Epidemiol Biomarkers Prev* 2015; 24:1898-1901.
26. Wentzensen N et al., Ovarian Cancer Risk Factors by Histologic Subtype: An Analysis From the Ovarian Cancer Cohort Consortium. *J Clin Oncol* 2016; 34:2888-2898.
27. Campbell PT et al., Body Mass Index, Waist Circumference, Diabetes, and Risk of Liver Cancer for U.S. Adults. *Cancer Res* 2016; 76:6076-6083.
28. Sonderman JS et al., Multiple Myeloma Mortality in Relation to Obesity Among African Americans. *J Natl Cancer Inst* 2016; 108.
29. Berrington de Gonzalez A et al., Body-mass index and mortality among 1.46 million white adults. *N Engl J Med* 2010; 363:2211-2219.
30. Collins FS, Varmus H, A new initiative on precision medicine. *N Engl J Med* 2015; 372:793-795.
31. Rebbeck TR, Precision prevention of cancer. *Cancer Epidemiol Biomarkers Prev* 2014; 23:2713-2715.