SCIENCE AND SOCIETY

Epidemiology — identifying the causes and preventability of cancer?

Graham A. Colditz, Thomas A. Sellers and Edward Trapido

Abstract | It has been almost 25 years since Doll and Peto performed their landmark analysis of epidemiological data to identify the causes of cancers and possible modes of cancer prevention. Since then, there have been many additional studies of cancer incidence using various epidemiological techniques. These studies revealed expanded opportunities for cancer prevention through approaches that include vaccination, increased physical activity, weight control and avoidance of post-menopausal hormone therapy.

In 1981, Richard Doll and Richard Peto published a landmark study of the causes of cancer that was based, in part, on a review of cancer incidence across many countries. As early as 1964, an expert committee of the World Health Organization had concluded that common fatal cancers arise as a result of lifestyle and other environmental factors, including environmental carcinogens, hormonal factors and dietary deficiencies, and are therefore potentially preventable[1]. However, Doll had used international variation as a pointer to cancer preventability[2]. The comprehensive 1981 study by Doll and Peto was undertaken to refine and clarify the evidence for the existence of factors that, if avoided, could reduce cancer burden in the United States[3] or England[4]. By comparing the rates of cancer mortality in the United States with those of other countries, they estimated the degree to which cancer incidence and mortality could be reduced in the United States[5].

Based on the epidemiological observation that migrants tend to acquire the cancer rates of their new country, Doll and Peto concluded that differences in cancer rates can be attributed, in part, to environmental factors such as smoking, diet, reproductive behaviour, sexual behaviour, infection and occupational exposures. This was a bold conclusion because most cancer research at the time was focused on specific occupational exposures or genetic factors. This work, the most comprehensive analysis at the time of evidence on the risk (BOX 1) and preventability of cancer, improved understanding of the range of factors that can affect cancer incidence.

Drawing on epidemiological surveillance data, Doll and Peto compared the rates of different cancer types in high and low incidence populations, and estimated the proportion of cancers that could be attributed to non-genetic factors. As shown in TABLE 1, the ratio of highest rate to lowest rate of cancer was as high as 100-fold or more. Connecticut was used as the reference population for the United States because it has a longstanding population-based cancer registry that has recorded all cancer diagnoses in the state since the early 1940s. Based on comparisons of high and low incidence regions, Doll and Peto concluded that 75–80% of cancers diagnosed in the United States in 1970 theoretically could have been avoided.

What made the US population different from low-risk populations? The environmental (non-genetic) factors that differ between the United States and low-risk populations are many and diverse, and include factors such as birthweight, age at puberty, lifelong patterns of diet, weight gain, alcohol consumption, use of tobacco, use of pharmacological agents, and reproductive factors. This conclusion was provocative as, at the time, there was only a limited amount of data from rigorously performed epidemiological studies that related diet, obesity and alcohol intake to cancer risk. Other commonly studied environmental exposures, such as differences in air, water and food contamination between the United States and other populations, were also thought to be involved in determining cancer risk, but to a lesser extent than previously assumed (FIG. 1).

There were many criticisms of the Doll and Peto studies, such as that they placed too much emphasis on lifestyle factors — for example, smoking and diet — with too little emphasis on involuntary exposures such as occupational and environmental carcinogens[6]. In the United States, extensive epidemiological data had documented the carcinogenic hazards of workplace exposures, including asbestos, benzene, arsenic, nickel, polycyclic hydrocarbons and vinyl chloride. As a result, there was much public concern about occupational exposures and their health impact, so the Occupational Safety and Health Administration was established in 1970, to ensure safe and healthy working conditions. Over time, the agency has had a positive impact, decreasing cancer risk among industrial workers through reduced exposure to carcinogens[7]. However, as US workplaces have become less carcinogenic, the hazards of production for products such as steel — which is associated with exposure to crystalline silica, polycyclic aromatic hydrocarbons and various other carcinogenic chemicals — have been largely exported to countries that have cheaper labour and lower production costs[8]. Further studies are required to determine whether the global cancer risk associated with producing 1000 Kg of steel is the same today as it was in 1970, regardless of country. Although fewer US workers are exposed to occupational carcinogens, and industrial exposures have only a minor contribution to cancer in this country[9], it is still important to understand the cancer risk associated with occupational exposures. By studying occupational as well as lifestyle factors that contribute to cancer incidence, it becomes possible to find ways to prevent cancer on many fronts, worldwide. The relative importance of

Box 1 | **Epidemiology definitions**

Risk
The most basic type of risk is absolute risk, which is simply a person's chance of developing a specific disease over a certain time-period.

Structured-data summaries
Systematic reviews or structured-data summaries are conducted to address a clearly defined question, and involve the identification of all relevant primary research studies that address the question.

Meta-analysis
The statistical analysis of combined data as reported from a number of studies, identified through a systematic review of the literature[101].

Pooled analysis
After obtaining the original data from the component studies in a systematic review, the reviewer re-analyses the data using common analytical approaches.

Prospective studies
These studies follow people over time to relate exposure measures to subsequent risk of disease.

Retrospective studies
Individuals with disease are identified, and a comparison group without disease drawn from the same population is used to compare the odds of past exposure between the two groups.

Bias
Also known as systematic error, bias distorts an association that has been observed in an epidemiological study through a number of possible processes, such as differences between cases and controls in participant selection, information recall, or failure to continue the study until its conclusion.

Confounding
Refers to the mixing or muddling of effects that can occur when the relationship of interest is confused by the effect of something else[102].

Randomized controlled trials
Randomization is used to ensure that the groups under study are as similar as possible at the start of the study, with the exception of the treatment or intervention under study.

Statistical power
The statistical power of a study is the probability that the study will detect an association of a particular size if it truly exists in the general population[102].

occupational and environmental factors will vary from country to country, as will lifestyle factors.

Doll and Peto acknowledged that their estimate that cancer rates could be reduced by 75–80% was a theoretical maximum, and that it was unlikely that society could change enough — even over many years — to decrease cancer incidence by this amount. Their analysis, however, provided an important starting point for subsequent studies of cancer causes and strategies for cancer prevention[10,11], and has even led to some strictly defined timelines for reducing cancer incidence in the United States[12].

In the past 25 years, many subsequent epidemiological studies have confirmed the contribution of specific lifestyle factors to the aetiology of cancers[13] and have expanded the list of cancer risk factors to include obesity and lack of physical activity. Specifically, it is estimated that achievable changes in the preventable causes of cancer now account for more than 50% of all cancer cases in the United States[13]. Evidence indicates that in the United States, tobacco use accounts for some 30% of all cancer cases, alcohol consumption for 4% of cancer cases, obesity for 15% of cancer cases, physical inactivity for 5% of cases, viruses for 3% of cases and poor diet for 10–25% of cases. Many prescription drugs have been associated with cancer risk through epidemiological studies (BOX 2). By contrast, Doll and Peto proposed in their 1981 report that there could be a link between use of oestrogen and endometrial cancer, and they speculated that oestrogens might also increase breast cancer risk — this hypothesis has since been proven correct in recent studies[14,15].

Table 1 | **Examples of ratio of highest versus lowest rate of cancer in men and women**

Cancer type	High incidence area	Low incidence area	Ratio of highest rate to lowest rate
Oesophagus	Iran	Nigeria	300
Lung	England	Nigeria	35
Stomach	Japan	Uganda	25
Prostate	United States, African-Americans	Japan	40
Breast (female)	Canada	Israel, non-Jewish population	7
Corpus uteri	California, United States	Japan	30

Data from REF. 5.

Epidemiological analysis of cancer

The most widely used scheme to identify a causal relationship between a lifestyle factor, such as smoking, and cancer is through structured-data summaries, meta-analyses and pooled analyses (BOX 1), which synthesize epidemiological data across multiple studies to draw unified conclusions about the causes of cancer. This is often done with consideration for differences in the design of the study because prospective studies are less susceptible to certain types of bias than retrospective studies (BOX 1). For example, in a case–control retrospective study, subjects with the cancer of interest (cases) are identified along with subjects without cancer (controls), and exposure data are recalled for a relevant time period before diagnosis. One of the most significant potential sources of bias in retrospective studies is in 'differential recall' — when cases are better than controls at remembering past exposures, particularly if they suspect that a certain exposure might have caused their disease. This type of bias (BOX 1) leads to the false appearance of differences in exposure frequencies in retrospective case–control studies.

Alternatively, prospective cohort studies are less susceptible to recall bias because the exposure or lifestyle data are recorded uniformly for study participants, who are followed over time to record incidence of cancer. The structured-data summary was developed during the preparation of the first report of the US Surgeon General on Smoking and Health in 1964 (REF. 16). The International Agency for Research on Cancer (IARC) performed similar

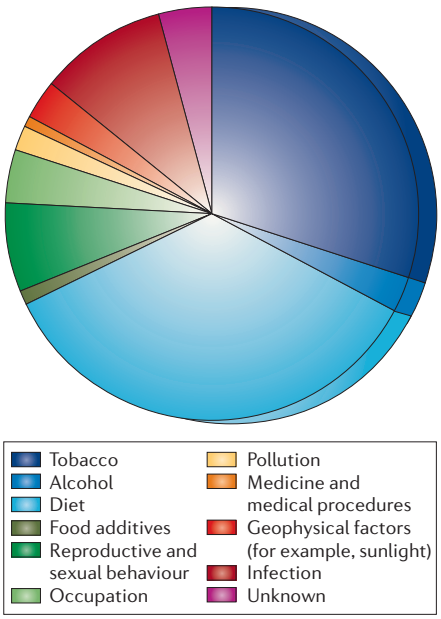

Figure 1 | **Proportion of cancer deaths attributed to non-genetic factors.** Estimated proportion of cancer in the United States that could have been avoided by changes in each category of non-genetic cancer causes, as estimated by Doll and Peto. Data from REF. 5.

structured-data summaries to associate other lifestyle factors with cancer, and has classified these in a monograph series (BOX 3). These reviews have provided a great deal of information on the environmental factors that increase cancer risk — since 1972, the IARC has published independent assessments of the carcinogenic risk posed to humans by more than 800 agents (see the IARC Press web site). Furthermore, since 1997, a new series of handbooks on cancer prevention have complemented the assessments of carcinogenic risk.

One challenge that epidemiologists face is that they cannot manipulate their systems — observational methods cannot certify results to the extent that is possible with laboratory experiments or randomized controlled clinical trials. Ethics and feasibility preclude the application of the experimental method when the exposure is potentially harmful[17], but epidemiologists do have a defined set of criteria that is considered sufficient to conclude that a cause-and-effect relationship exists[18]. These include biological plausibility (inhaling tobacco smoke into the lungs might cause cancer in those tissues), dose–response (heavier smokers have greater risks than lighter smokers), lack of temporal ambiguity (smoking precedes lung cancer) and coherence of all the evidence.

Consistency of findings across studies, and across study designs, also strengthens the case that exposure to a certain factor causes cancer. For example, recent data on breast and colon cancer incidence has been obtained from the Women's Health Initiative randomized controlled trial of use of oestrogen plus progestin among post-menopausal women. This data shows increases in the risk of breast cancer and reductions in the risk of colon cancer, and supports similar conclusions that have been obtained from earlier epidemiological investigations[19]. Furthermore, several recent studies have shown that some observational studies of drug efficacy have had comparable results to those of randomized clinical trials[20,21]. Epidemiological studies have also been essential in making the initial association between an environmental factor and cancer that could later be followed up and confirmed through laboratory analysis. For example, cancer rates differ dramatically between Western and Asian countries, as do differences in consumption of foodstuffs such as green tea. Based on this ecological correlation, laboratory studies have been done to evaluate whether green tea or some component of it can modify carcinogenesis[22] (BOX 4).

Identifying causes of cancer

Tobacco. Doll and Peto were not the first to perform a large epidemiological analysis of cancer causes. In 1964, the US Surgeon General's Report on Smoking and Health documented the causal relationship between smoking and lung cancer, and was the starting point for the national reversal on smoking within the United States and led to the ultimate reduction in the burden of cancer (FIG. 2). This report was based, in part, on the evidence from seven prospective studies that started as early as 1951. These studies documented 1,833 cases of lung cancer among smokers, a risk over 10 times higher than that of non-smokers. This strong association led the Surgeon General to conclude that smoking caused lung cancer[16]. These data were later supported by evidence from more than 20 retrospective studies.

From 1964 onwards, these epidemiological findings were translated to a broad range of cancer prevention strategies that aimed to reduce the use of tobacco. For many cancers, the reduction in risk after cessation of smoking takes years to be detected. For example, lung cancer risk in former smokers approaches that of 'never smokers' only 15–20 years after the individual stops smoking[23]. Not surprisingly, it was from the 1990s onwards that the incidence of smoking-related cancers, including lung cancer, dramatically fell, particularly among men.

Prevention efforts to reduce tobacco use in the US population have been in existence for a long time and continue to expand. In parallel with the increasing focus on prevention of tobacco use among youth[24] and increasing cessation among adults[25], researchers have identified the molecular mechanisms of tobacco-induced cancer[26,27]. Although such mechanisms could ultimately lead to new agents that increase or inhibit the carcinogenic mechanisms of tobacco, from a public health point of view, understanding the mechanism is less important than the knowledge that smoking causes lung cancer[5]. This is because it is always easier to prevent a disease through interrupting the root cause than to treat it through altering mechanistic processes. Although epidemiological studies have provided the information that is necessary to devise prevention efforts, data from studies such as the prospective American Cancer Society (ACS) Cancer Prevention Study II cohort study have been central to the estimation of the global burden of tobacco smoke[28]. Some 3 million deaths a year are estimated to be attributable to smoking, and this number has been estimated to increase to 10 million a year in 30–40 years time[29].

Radiation. Resounding evidence from studies of occupational exposure and of atomic bomb survivors has clearly shown a dose–response relationship between exposure to radiation and the risk of cancer[30]. More recently, solar radiation has been identified as a significant risk factor

Box 2 | **Pharmaceuticals that modify cancer risk**

- Oral contraceptives reduce the risk of ovarian cancer[103] and endometrial cancer, and increase the risk of breast cancer in current users[51]
- Post-menopausal hormone therapy increases the risk of breast[14] and endometrial cancers, but reduces the risk of colon cancer[104]
- Diethylstilbestrol increases the risk of vaginal cancer[105]
- Non-steroidal anti-inflammatory drugs decrease the risk of colorectal cancer[106]
- Tamoxifen increases the risk of endometrial cancer[107]

Box 3 | **How are carcinogens classified, using the current IARC system?**

The International Agency for Research on Cancer (IARC) classification system states that a factor is considered to be a 'definite' carcinogen when an association has been established between exposure and outcome, and chance, bias and confounding can be ruled out with reasonable confidence. A factor is considered to be a 'probable' carcinogen when the association is established, but chance, bias and confounding cannot be ruled out with reasonable confidence. Finally, a factor is a 'possible' carcinogen when available studies are of insufficient quality, consistency, or statistical power to permit a conclusion of probable or definite association between the exposure and outcome[108].

for melanoma[31], with worldwide incidence of this malignancy increasing rapidly[32]. Epidemiological methods continue to be applied to examine the potential association of other sources of radiation with cancer risk, such as the association between exposure to radiation or to power lines and leukaemia, or a correlation between electromagnetic radiation from mobile phones and brain tumours[33]. Additional studies are underway to understand whether some of the familial components of risk for tumours such as breast cancer might be due to increased sensitivity to radiation in individuals who carry certain genetic defects, such as mutations in *ATM* (ataxia telangiectasia mutated), *BRCA1* (breast cancer 1) or *BRCA2* (REF. 34). Observational epidemiological studies have made significant contributions to our understanding of the influence that radiation has on cancer risk and have also provided clues into mechanisms of cancer pathogenesis.

Obesity and lack of physical activity. Trends in obesity over time are paralleled by increases in the rates of certain cancers[35]. Data from the National Health and Nutrition Examination Survey (NHANES), a nationally representative sample of US adults, shows that 65% of Americans are overweight or obese[36] — more than twice the percentage of the population that was considered to be obese in 1960. In 2002, the IARC Prevention Report on Weight Control and Physical Activity listed "obesity and lack of physical activity" as causes of cancer incidence and mortality[37]. Specifically, obesity was described as a cause of oesophageal, colon, uterine, kidney and post-menopausal breast cancer. Data from the ACS Cancer Prevention Study II, which followed more than 1 million men and women for an average of 16 years, showed an additional link to cancers of the prostate and pancreas, as well as to non-Hodgkin lymphoma and myeloma[38]. That study concluded that 16–20% of cancer deaths among women and 14% of cancer deaths among men were attributable to obesity[39]. Furthermore, the IARC monograph also reported that there was sufficient evidence to conclude that lack of physical activity increased the risk of breast and colon cancer — two of the most common cancers in the United States and Western Europe[37]. The ability of exercise to prevent these cancers is independent of the level of obesity. Given the epidemiology of obesity and growing understanding of the importance of obesity as a cause of cancer[38], in the coming years epidemiological methods for refining assessment of exposures (such as energy balance in light of energy expenditures versus dietary intake, or assessment of complex patterns of food consumption over a lifetime) will be developed.

Viruses and other infections. Viruses have long been known to cause certain types of cancer in animals, stimulating research into the role of infection in human cancers. Accordingly, there have been many studies of associations between infectious diseases and cancer, but these are often complicated because of confounding factors (BOX 1) and variations in laboratory assays. For example, although sexually transmitted viruses increase the risk for specific types of cancer, other diseases are also sexually transmitted. So, the effects of prevention strategies on cancer morbidity and mortality are not easily quantified.

Nonetheless, associations that were made through epidemiological studies led to the laboratory studies that eventually confirmed the causes of many cancer types. For example, epidemiological studies in the 1990s were the first to show that cervical cancer was caused by human papillomavirus (HPV) infection[40,41], and later studies associated HPV with vulvar, penile and anal cancers. Other epidemiological studies of viral infections followed suit — hepatitis B and C viruses were identified as a cause of hepatocellular cancer, and human lymphotropic virus type 1 was found to cause adult T-cell leukaemia. Epidemiological analyses helped establish the link between human immunodeficiency virus type 1 infection and Kaposi sarcoma or non-Hodgkin lymphoma, as well as between human herpes virus, Kaposi sarcoma and body-cavity lymphoma[42,43].

Epidemiologists were also the first to associate *Helicobacter pylori* infection with gastric cancer[44,45]. Further studies led to new treatment and preventative strategies for this disease, as subsequent randomized trials showed that eliminating the *H. pylori* infection resulted in a modest slowing of the pre-cancerous process but did not prevent cancer[46].

Efforts to contain the spread of viruses through sexual contact include behavioural and educational interventions (to modify sexual behaviour)[47], biomedical interventions (to develop and administer vaccines) and structural interventions through regulatory changes (to make contraceptives readily available). Epidemiology has been applied to evaluate prevention strategies after a cancer-causing agent has been identified. For example, through epidemiological analysis, it has been estimated that childhood administration of a vaccine against hepatitis B could reduce the global burden of liver cancer by 60% (REF. 48). Compelling evidence shows that vaccines against HPV[49] and hepatitis B[50] offer opportunities for preventing two cancers that have significant global burden.

Hormones. In 1981, the role of hormones and other drugs in cancer aetiology was still widely debated. Since that time, comprehensive studies and combined analyses have confirmed that current use of oral contraceptives increases the risk of breast cancer[51] and that use of post-menopausal hormones

Box 4 | **Observational studies compared with randomization?**

In studies of some drugs and behavioural interventions, investigators control the exposure and can randomly allocate participants to receive the therapy or a comparison (usually placebo) therapy. Through randomization, each group will be equally likely to develop cancer, so differences in risk can be attributed to the intervention. In observational studies, researchers record the exposures that are reported by participants and determine if they are associated with the development of cancer. For many exposures, such as oral contraceptive use, alcohol intake, occupational exposures, and so on, it is not possible to randomly allocate participants to receive the exposure of interest or a placebo. Through analysis, researchers must remove any differences in risk of cancer through stratification and multivariate adjustment.

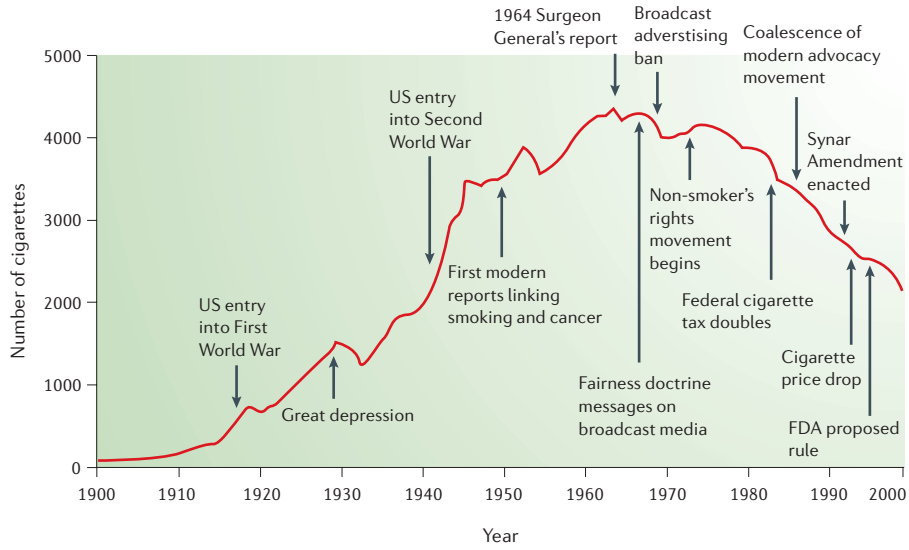

Figure 2 | **Cigarette consumption in the United States.** Per capita cigarette consumption among US adults from 1990–1999, and the major scientific, social and regulatory events that contributed to the decline in cigarette consumption. FDA, the United States Food and Drug Administration. Figure reproduced from REF. 109.

increases risk with increased duration of use[14]. Furthermore, hormone therapies that involve oestrogen plus progestin increase risk more than therapies that are based on oestrogen alone. Numerous other drugs have also been associated with increases and decreases in cancer risk (BOX 2). Continuing investigation is addressing formulations of post-menopausal hormones that might be less toxic to the breast, and the role of selective oestrogen receptor modulators to reduce the risk of breast and endometrial cancer. The impact of post-menopausal hormones on other cancers, such as ovarian, remains to be defined.

Genetic susceptibility

Although the potential for cancer prevention by simply modifying lifestyle, based on epidemiological findings, was appealing, additional studies indicated that subgroups of the population might carry genetic susceptibility to certain cancers[52]. As early as 1922, experimental evidence indicated that certain strains of mice were more susceptible to certain types of cancer[53]. Consequently, most population-based studies in the 1930s and thereafter focused on the risk of cancer in relatives of patients with a specific type of cancer. For example, studies in the 1940s and 1950s indicated that breast cancer cases tended to cluster in families[54,55]. Compelling evidence for a genetic component to breast cancer came from the Cancer and Steroid Hormone Study. This multicentre, population-based case–control study, conducted by the Centers for Disease Control and Prevention, included 4,730 histologically confirmed cases of breast cancer in patients aged 20–54 years, and 4,688 controls. Initial analyses confirmed that cases were significantly more likely than controls to have a family history of the disease, especially cases who developed cancer at an early age[56]. A segregation analysis of the pattern of breast cancer in the case families provided evidence that the susceptibility was transmitted in a Mendelian manner[57]. Linkage analysis using DNA markers generated in the laboratory localized the first putative gene to a region of chromosome 17q21 (REF. 58), and *BRCA1* was subsequently identified through positional cloning[59]. Li and colleagues drew on the National Cancer Institute's (NCI's) cancer family registry to identify 24 kindreds with multiple cancers in young patients[60]. From this epidemiological investigation, the Li–Fraumeni syndrome was defined and subsequent genetic investigation identified germline *TP53* mutations. These few examples highlight the flow back and forth from epidemiology to laboratory investigation to refine understanding of aetiological pathways in cancer.

The successful mapping of the human genome has significantly improved our ability to map and identify additional cancer susceptibility genes. With the identification of additional cancer genes, we expect to increase the possibility of genetic testing and tailored interventions, such as screening or prophylaxis to match the genetic predisposition of individuals. Although the risks and potential benefits of such 'personalized medicine' remain to be documented[61], such a targeted approach could prove to be more cost-effective than a public health approach in which an intervention is applied to the whole population, if the exposures of interest are of modest prevalence. For example, might colon cancer screening recommendations be based on an individual's level of genetic predisposition to this cancer, rather than embarking on population-wide screening? Epidemiology studies can also be used to identify patients who are most likely to respond to certain therapeutic approaches.

So, observational epidemiological methods are particularly well-suited to examine the association between a particular genetic variation and cancer risk or response to therapy. Indeed, there are many large case–control and cohort studies underway to identify these associations. International consortia have been established to study the association between genetic factors and the risk of lymphoma, breast and prostate cancer, as well as many other cancer types. As humans have more than 30,000 genes, it is unrealistic to think that randomized controlled trials (BOX 1) can be conducted to correlate inherited genetic variation with outcomes, so we will increasingly need to rely on epidemiology to evaluate the promise of personalized medicine.

Prevention strategies

Knowledge of the causes of cancer does not guarantee that action will be taken to reduce exposure to carcinogenic factors or to increase preventive behaviours. Despite accumulating evidence that indicates that most cancer cases could be prevented, few national prevention efforts have been launched. To bring priority and a commitment of funding to cancer prevention, the scientific evidence must be properly summarized and presented to garner allocation of funding to build prevention efforts[62]. Among these activities are the estimation of population burden of cancer and the proportion that is attributable to preventable causes[28,63] (FIG. 1). Such estimates derive from analytical epidemiological data and surveillance data on risk factors. Using scientific evidence to identify the causes of cancer, population scientists build prevention strategies to translate these associations into further prevention strategies that require individual changes in behaviour, structural changes in society, and healthcare systems that sustain the adoption of lower risk behaviours across all members of society. Adoption of such strategies can reduce the risk of cancer[64].

Although the history of cancer control in the United States is long[65], tobacco studies have offered a unique window into the time-course of prevention efforts. Estimates of the benefits of smoking cessation are based on the declining prevalence of smoking following the 1964 Report of the Surgeon General. The prevalence of smoking among US adults was 42.4% in 1965, which decreased to 23.3% in 2000. At the same time, the prevalence of past smokers increased from 24.3% of the population in 1965 to 49.6% in 1993, but flattened after that to remain at 48.8% in 2000 (REF. 66). Converting these changing patterns of cigarette smoking to reduced mortality, Warner has estimated that the reduction in smoking between 1964 and 1985 resulted in the avoidance of 2.1 million smoking-related deaths between 1986 and 2000 (REF. 67). Despite these reductions in cigarette smoking, it is estimated that more than 750,000 avoidable smoking-related cancer deaths occurred between 1995 and 1999 (REF. 68). The decrease in risk after cessation from smoking spans decades, and substantial proportions of adults continue to smoke, which points to the need for continuing efforts to speed cessation from smoking at all ages.

So, for epidemiology and basic science to have an impact on the burden of cancer, research findings must be coupled with prevention strategies that bring the scientific advances to the public through broad-scale, population-wide efforts; through clinical interventions; or through a combination of both. Greater attention to the analysis and dissemination of the scientific evidence will speed the transfer of research findings to prevention efforts. More data on the cost-effectiveness of cancer prevention can further accelerate the translation. For example, smoking cessation and relapse prevention has been estimated to cost US$1,581 and $83 per year of life saved, respectively[69]. By comparison, in terms of cost per year of life saved, flashing lights at rail crossings cost $42,000, annual mammography costs $190,000, an ejection system for a B-58 bomber costs $1.2 million, and seat belts in school buses costs $2.8 million.

Epidemiology of early detection

A growing field for the application of epidemiological methods has been in screening and identifying biomarkers that can lead to earlier diagnosis. Although the only way to ensure that nobody dies from cancer is to prevent the disease from occurring in the first place, the magnitude of the burden of cancer in the United States makes prevention of all cancers unattainable in the next 10 years. Cancers that are detected at early stages have a significantly better prognosis than those that are detected at advanced stages. The 5-year survival rates for patients with most cancer types is, on average, near 90% for patients who are diagnosed with stage I tumours, and is 10% or less for patients who are diagnosed with stage IV tumours[70]. Screening strategies are routinely performed for cancers of the breast, prostate, colon and cervix, even though randomized clinical trials that demonstrate efficacy are lacking or controversial[71–73].

Unfortunately, a number of factors more broadly hinder the application of early detection. The sensitivity and specificity for approved tests are less than optimal. For example, even when mammograms are performed on state-of-the-art equipment and evaluated by expert radiologists, 5–10% of tumors are missed[74,75], and it has been estimated that 95% of women who have abnormalities reported in screening mammograms do not have breast cancer[76]. The Prostate Specific Antigen (PSA) test that is used to detect early prostate cancer is also less than ideal. One of the challenges is that it is difficult to identify a PSA concentration (cut-off point) that clearly separates men who probably have a prostate tumor from those who do not. For example, PSA concentrations vary by age and race of patients, and PSA concentrations are influenced by non-malignant processes such as benign prostatic hypertrophy[77]. Colonoscopy is considered the gold standard of an early detection test by gastroenterologists, but the procedure is not easy to apply, requires highly skilled personnel and is costly[78,79].

The level of acceptance of these screening modalities by the general population varies considerably, despite extensive research and outreach efforts to increase their use[80]. As a result of this, there remains a significant need to develop and evaluate screening tests for cancer that are cheap, easy to apply, do not require highly skilled physicians to conduct them, have high sensitivity and specificity, and are appropriate for all cancers, not just a select handful. For example, more people will die from lung cancer than from breast, colorectal and prostate cancer combined, yet there remains no validated screening strategy for it despite the evidence that, similar to other cancers, stage is associated with mortality.

Significant advances in early detection will require a paradigm shift from the reliance on imaging modalities that are capable of detecting tumors that have already reached a mass in the order of billions of transformed cells, to one that harnesses the fact that cancer is a process that reflects the accumulation of insults to cells, which are detectable years before the onset of clinical signs or symptoms — even as the tumour is too small to be imaged with existing technologies.

It is possible to exploit the discoveries into the molecular mechanisms of cancer that have been made over the past few decades to develop new early-detection assays[81–83]. Enabling technologies, such as gene-expression profiling[84,85], proteomics[86,87], methylation analysis[88] and high-throughput genotyping[89,90], are beginning to unravel the secrets of malignant transformation. Processes such as lymphangiogenesis and angiogenesis occur at early stages of the disease, and are important disease markers[91,92]. This knowledge, along with the growing catalogue of mutational events and altered protein concentrations or structures that are found in cancer cells, now makes it possible to pursue novel early-detection strategies, based on analysis of blood or urine[93–95]. It is unrealistic to think that a randomized trial will be conducted to evaluate every single potential early-detection biomarker. However, with carefully collected healthy populations, properly stored and annotated biospecimens, and adequate periods of follow-up to detect relevant outcomes (certainly cancer and mortality, but potentially also intermediate endpoints such as a colon polyps), epidemiological methods can contribute significantly to the translation of biomarkers that are discovered in the laboratory into determinants of cancer risk and diagnostics.

Within the NCI Early-Detection Research Network (EDRN), the phases of development and validation of biomarkers that are suitable for early detection have been defined. This definition uses a logical progression from discovery to retrospective longitudinal evaluation, and then to prospective evaluation using banked blood samples to document the extent and characteristics of disease that have been detected by the test, and the false-positive rates[96]. Studies that involve cohorts of stored samples, with subsequent follow-up and confirmation of incident cancers, will permit laboratory analysis of archived urine, blood or other tissues that were collected at a time before participants developed disease. These studies will permit discovery or validation of biomarkers that discriminate between subjects who remain cancer-free and those who develop cancer. The retrospective nature of such 'biorepositories' significantly reduces the

length of time to complete epidemiological studies, compared with a purely prospective trial, as the outcomes have already occurred and the specimens were collected even before that. For example, the observational arm of the Women's Health Initiative, which includes roughly 100,000 post-menopausal women and follow-up periods of 8–12 years, provides a powerful resource to examine the association of biomarkers of risk in combination with clinical and lifestyle data to test hypotheses, as does the biorepository of the Nurses' Health Study[97].

Since the first significant studies of cancer causes were performed by Doll and Peto, epidemiological methods have contributed greatly to our understanding of cancer pathogenesis and modes of prevention. Contemporary studies will probably contribute further insights about how the distribution of risk varies by genotype and how biomarkers of disease might be used in early detection of cancer, when the promise for cure is greatest.

Future directions

Method development and method incorporation. The epidemiology of obesity and growing understanding of the importance of obesity as a cause of cancer demand new epidemiological methods for refining assessment of exposures such as energy balance. Disease classification, based on genetic markers rather than traditional morphological features, and new statistical methods for assessing the vast amounts of data that are generated by genetic or molecular epidemiological studies (such as whole-genome scans)[98] will also change the scope of epidemiological research.

In many areas, however, advances in technology provide challenges for collaboration between the laboratory and epidemiological sciences. For example, through nanotechnology, pharmacogenomics, whole-genome scans, novel imaging techniques, and so on, epidemiological questions can be addressed that were technologically impossible, infeasible or too slow and costly to be undertaken previously. However, many current studies neglect the epidemiological properties of these technologies, such as sensitivity, specificity, predictive value, variability and reproducibility, and so on[99]. Therefore, another line of research will involve how to best incorporate novel technological advances into epidemiological research and how to bring the principles of epidemiology to the identification and refinement of new markers.

Aetiology and survivorship. There are multiple opportunities for epidemiology studies to extend from research into cancer pathogenesis**,** such as from studies to characterize the role of the microenvironment in tumour progression and recurrence. Understanding the biological mechanisms of tumour formation is an essential step in cancer detection, treatment and prevention. For example, in addition to the roles that epidemiology studies have in the validation of biomarkers for early detection and screening, biomarkers that are associated with tumour growth, progression, metastasis and response to therapy will also need to be validated in population studies.

The final important aspect of epidemiology research is the ability to form consortium studies — to combine data from different studies on different populations to form universal conclusions. Consortium studies are particularly important for rarer forms of cancer — most NCI-funded epidemiology investigations have been for research into breast, prostate, lung and colorectal cancer incidence, so studies into factors that cause less common cancers such as lymphoma and myeloma, brain cancers, and so on, could benefit from consortium approaches. These types of studies require new methods for ultra-rapid case ascertainment to obtain data while subjects are healthy enough to participate. Nevertheless, for patients with both common and uncommon cancers, consortium studies can provide sufficient statistical power (BOX 1) to evaluate cancer susceptibility, epigenetic factors, gene–environment interactions, gene–gene interactions, gene–gene–environment interactions, and so on[100]. Epidemiological studies of the interactions between behaviour, psychological and social mechanisms, and risk indicators could also fill large knowledge gaps in both aetiology and survivorship studies.

Graham A. Colditz is at the Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.

Thomas A. Sellers is at the H. Lee Moffitt Cancer Center & Research Institute, 12902 Magnolia Drive, Tampa, Florida 33612, USA.

Edward Trapido is at the Epidemiology and Genetics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, EPN 5112, 6130 Executive Boulevard, Bethesda, Maryland 20852, USA.

Correspondence to G.A.C.
e-mail: Graham.Colditz@channing.harvard.edu
doi:10.1038/nrc1784
Published online 22 December 2005

1. World Health Organization. *Prevention of Cancer* (WHO, Geneva, 1964).
2. Doll, R. *Prevention of Cancer — Pointers from Epidemiology* (Nuffield Hospital Trust, London, 1967).
3. Wynder, E. L. & Gori, G. B. Contribution of the environment to cancer incidence: an epidemiologic exercise. *J. Natl Cancer Inst.* **58**, 825–832 (1977).
4. Higginson, J. & Muir, C. S. Environmental carcinogenesis: misconceptions and limitations to cancer control. *J. Natl Cancer Inst.* **63**, 1291–1298 (1979).
5. Doll, R. & Peto, R. *The Causes of Cancer: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today* (Oxford University Press, New York, 1981).
6. Epstein, S. & Swartz, J. Fallacies of lifestyle cancer theories. *Nature* **289**, 127–130 (1981).
7. Weil, D. OSHA: Beyond the politics. *Frontline* (online), http://www.pbs.org/wgbh/pages/frontline/shows/workplace/osha/weil.html (2003).
8. Xu, Z. *et al.* Cancer risks among iron and steel workers in Anshan, China, part II: case–control studies of lung and stomach cancer. *Am. J. Ind. Med.* **30**, 7–15 (1996).
9. Monson, R. R. & Christiani, D. C. Summary of the evidence: occupation and environment and cancer. *Cancer Causes Control* **8**, 529–531 (1997).
10. Willett, W., Colditz, G. & Mueller, N. Strategies for minimizing cancer risk. *Sci. Amer.* **275**, 88–95 (1996).
11. Prevention Working Group. Cancer control objectives for the nation: 1985–2000. *NCI Monogr.* **2**, 3–11 (1986).
12. Byers, T. *et al.* The American Cancer Society challenge goals. How far can cancer rates decline in the U.S. by the year 2015? *Cancer* **86**, 715–727 (1999).
13. Colditz, G. A., DeJong, D., Hunter, D. J., Trichopoulos, D. & Willett, W. C. Harvard report on cancer prevention. Volume 1. Causes of human cancer. *Cancer Causes Control* **7**, 1–59 (1996).
14. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy. Combined reanalysis of data from 51 epidemiological studies involving 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **350**, 1047–1059 (1997).
15. Rossouw, J. E. *et al.* Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* **288**, 321–333 (2002).
16. US Public Health Service. *Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service* (US Department of Health, Education, and Welfare, Public Health Service, Centers for Disease Control, Washington DC, 1964).
17. Angell, M. The interpretation of epidemiologic studies. *N. Engl. J. Med.* **323**, 782–788 (1990).
18. Hill, A. B. The environment and disease: association or causation? *Proc. R. Soc. Med.* **58**, 295–300 (1965).
19. Fletcher, S. W. & Colditz, G. A. Failure of estrogen plus progestin therapy for prevention. *JAMA* **288**, 366–368 (2002).
20. Benson, K. & Hartz, A. J. A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med.* **342**, 1878–1886 (2000).
21. Concato, J., Shah, N. & Horwitz, R. I. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342**, 1887–1892 (2000).
22. Ahmad, N. & Mukhtar, H. Green tea polyphenols and cancer: biologic mechanisms and practical implications. *Nutr. Rev.* **57**, 78–83 (1999).
23. Peto, R. *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case–control studies. *Br. Med. J.* **321**, 323–329 (2000).
24. US Department of Health and Human Services. *Preventing Tobacco Use among Young People. A Report of the Surgeon General* (US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, Georgia, 1994).
25. Pierce, J. P., Fiore, M. C., Novotny, T. E., Hatziandreu, E. J. & Davis, R. M. Trends in cigarette smoking in the United States. Projections to the year 2000. *JAMA* **261**, 61–65 (1989).
26. Hecht, S. Tobacco smoke carcinogens and lung cancer. *J. Natl Cancer Inst.* **91**, 1194–1210 (1999).

27. Hecht, S. S. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature Rev. Cancer* **3**, 733–744 (2003).
28. Ezzati, M. & Lopez, A. D. Estimates of global mortality attributable to smoking in 2000. *Lancet* **362**, 847–852 (2003).
29. Peto, R. *et al.* Mortality from smoking worldwide. *Br. Med. Bull.* **52**, 12–21 (1996).
30. Committee on the Biologic Effects of Ionizing Radiation. *Health Effects of Low Levels of Ionizing Radiation. BEIR V* (National Academy Press, Washington DC, 1990).
31. Armstrong, B. & Kricker, A. The epidemiology of UV induced skin cancer. *J. Photochem. Photobiol. B.* **63**, 8–18 (2001).
32. Parkin, D. M., Pisani, P. & Ferlay, J. Estimates of the worldwide incidence of 25 major cancers in 1990. *Int. J. Cancer* **80**, 827–841 (1999).
33. Ahlbom, A. *et al.* Epidemiology of health effects of radiofrequency exposure. *Environ. Health Perspect.* **112**, 1741–1754 (2004).
34. Bernstein, J. L. *et al.* Study design: evaluating gene–environment interactions in the etiology of breast cancer — the WECARE study. *Breast Cancer Res.* **6**, R199–R214 (2004).
35. Poledнak, A. Trends in incidence rates for obesity-related cancers in the US. *Cancer Detect. Prev.* **27**, 415–421 (2003).
36. Hedley, A. A. *et al.* Prevalence of overweight and obesity among US children, adolescents, and adults, 1999–2002. *JAMA* **291**, 2847–2850 (2004).
37. International Agency for Research on Cancer. in *Weight Control and Physical Activity* 315 (International Agency for Research on Cancer, Lyon, 2002).
38. Calle, E. E. & Kaaks, R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nature Rev. Cancer* **4**, 579–591 (2004).
39. Calle, E. E., Rodriguez, C., Walker-Thurmond, K. & Thun, M. J. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N. Engl. J. Med.* **348**, 1625–1638 (2003).
40. Schiffman, M. H. *et al.* Epidemiologic evidence showing that human papillomavirus infection causes most cervical intrepthelial neoplasia. *J. Natl Cancer Inst.* **85**, 958–964 (1993).
41. Franco, E., Rohan, T. & Villa, L. Epidemiologic evidence and human papillomavirus infection as a necessary cause of cervical cancer. *J. Natl Cancer Inst.* **91**, 506–511 (1999).
42. Buchschacher, G. & Wong-Staal, F. in *Cancer Principles and Practice of Oncology* (eds DeVita, V., Hellman, S. & Rosenberg, S.) 165–173 (Lippincott Williams & Wilkins, Philadelphia, 2005).
43. Howley, P., Ganem, D. & E, K. in *Cancer Principles and Practice of Oncology* (eds DeVita, V., Hellman, S. & Rosenberg, S.) 173–184 (Lippincott Williams & Wilkins, Philadelphia, 2005).
44. International Agency for Research on Cancer Working Group on the Evaluation of Carcinogenic Risks to Humans. in *Schistosomes, liver flukes and heicobacter pylori: views and expert opinions of an IARC working grooup on the evaluation of carcinogenic risks to humans* 177–240 (IARC Press, Lyon, France, 1994).
45. Webb PM & Forman D. *Helicobacter pylori* as a risk factor for cancer. *Baillieres Clin. Gastroenterol.* **9**, 563–582 (1995).
46. Correa, P. Is gastric cancer preventable? *Gut* **53**, 1217–1219 (2004).
47. US Institute of Medicine. *The Hidden Epidemic: Confronting Sexually Transmitted Diseases* (National Academy Press, Washington DC, 1997).
48. Stuver, S. Towards the global control of liver cancer. *Semin. Cancer Biol.* **8**, 299–306 (1998).
49. Koutsky, L. A. *et al.* A controlled trial of a human papillomavirus type 16 vaccine. *N. Engl. J. Med.* **347**, 1645–1651 (2002).
50. Chang, M. H. *et al.* Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. Taiwan Childhood Hepatoma Study Group. *N. Engl J. Med.* **336**, 1855–1859 (1997).
51. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53,297 women with breast cancer and 100,239 women without breast cancer from 54 epidemiological studies. Collaborative Group on Hormonal Factors in Breast Cancer. *Lancet* **347**, 1713–1727 (1996).
52. Rich, S. S. & Sellers, T. A. in *The Genetic Basis of Common Diseases.* (eds King, R. A., Rotter, J. & Motulsky, A. G.) 39–49 (Oxford University Press, Inc., New York, 2002).
53. Slye, M. Biological evidence for the inheritability of cancer in man: studies in the incidence and inheritability of spontaneous tumors in mice. Eighteenth report. *J. Cancer Res.* **7**, 107–147 (1922).
54. Jacobsen, O. *Heredity in Breast Cancer* (H. K. Lewis, London, 1946).
55. Anderson, V. E., Goodman, H. O. & Reed, S. C. *Variables Related to Human Breast Cancer* (University of Minnesota Press, Minneapolis, 1958).
56. Claus, E. B., Risch, N. J. & Thompson, W. D. Age at onset as an indicator of familial risk of breast cancer. *Am. J. Epidemiol.* **131**, 961–972 (1990).
57. Claus, E. B., Risch, N. & Thompson, W. D. Genetic analysis of breast cancer in the cancer and steroid hormone study. *Am. J. Hum. Genet.* **48**, 232–242 (1991).
58. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
59. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
60. Li, F. *et al.* A cancer family syndrome in twenty-four kindreds. *Cancer Res.* **48**, 5358–5362 (1988).
61. Peto, J. Cancer epidemiology in the last century and the next decade. *Nature* **411**, 390–395 (2001).
62. Atwood, K., Colditz, G. & Kawachi, I. Implementing prevention policies: relevance of the Richmond model to health policy judgments. *Am. J. Public Health* **87**, 1603–1606 (1997).
63. Peto, R., Lopez, A. D., Boreham, J., Thun, M. & Heath, C. J. Mortality from tobacco in development countries: indirect estimation from national vital statistics. *Lancet* **339**, 1268–1278 (1992).
64. Curry, S., Byers, T. & Hewitt, M. *Fulfilling the Potential of Cancer Prevention and Early Detection* (National Academy Press, Washington DC, 2003).
65. Hiatt, R. & Rimer, B. A new strategy for cancer control research. *Cancer Epidemiol. Biol. Prev.* **8**, 957–964 (1999).
66. Giovino, G. Epidemiology of tobacco use in the United States. *Oncogene* **21**, 7326–7340 (2002).
67. Warner, K. Effects of the antismoking campaign: an update. *Am. J. Public Health* **79**, 144–151 (1989).
68. Centers for Disease Control and Prevention. Annual smoking-attributable mortality, years of potential life lost, and economic costs — United States, 1995–1999. *MMWR Morb. Mortal. Wkly Rep.* **51**, 300–303 (2002).
69. Tengs, T. *et al.* Five hundred life saving interventions and thier cost-effectiveness. *Risk Analysis* **15**, 369–390 (1995).
70. Etzioni, R., Urban, N. & Ramsey, S. The case for early detection. *Nature Rev. Cancer* **3**, 243–252 (2003).
71. Smith, R. A. *et al.* American Cancer Society guidelines for breast cancer screening: update 2003. *CA Cancer J. Clin.* **53**, 141–169 (2003).
72. Smith, R. A. *et al.* American Cancer Society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers. Also: update 2001 — testing for early lung cancer detection. *CA Cancer J. Clin.* **51**, 38–75; quiz 77–80 (2001).
73. Saslow, D. *et al.* American Cancer Society guideline for the early detection of cervical neoplasia and cancer. *CA Cancer J. Clin.* **52**, 342–362 (2002).
74. Mushlin, A. I, Kouides, R. W. & Shapiro, D. E. Estimating the accuracy of screening mammography: a meta-analysis. *Am. J. Prev. Med.* **14**, 143–153 (1998).
75. Kouskos, E. *et al.* Missed cancers on mammograms: causes and measures of prevention. *Eur. J. Gynaecol. Oncol.* **25**, 230–232 (2004).
76. Elmore, J. G., Armstrong, K., Lehman, C. D. & Fletcher, S. W. Screening for breast cancer. *JAMA* **293**, 1245–1256 (2005).
77. Minardi, D. *et al.* Diagnostic accuracy of percent free prostate-specific antigen in prostatic pathology and its usefulness in monitoring prostatic cancer patients. *Urol. Int.* **67**, 272–282 (2001).
78. Seeff, L. C. *et al.* Patterns and predictors of colorectal cancer test use in the adult U.S. population. *Cancer* **100**, 2093–2103 (2004).
79. Frazier, A., Colditz, G., Fuchs, C. & Kuntz, K. Cost-effectiveness of screening for colorectal cancer in the general population. *JAMA* **284**, 1954–1961 (2000).
80. Schwartz, L., Woloshin, S., Fowler, F. Jr & Welch, H. Enthusiasm for cancer screening in the United States. *JAMA* **291**, 71–78 (2004).
81. Gorga, F. *The Molecular Basis of Cancer* (Bridgewater Review, 1998).
82. Spencer, S. L., Berryman, M. J., Garcia, J. A. & Abbot, D. An ordinary differential equation model for the multistep transformation to cancer. *J. Theor. Biol.* **231**, 515–524 (2004).
83. Abdel-Rahman, W. M. & Peltomaki, P. Molecular basis and diagnostics of hereditary colorectal cancers. *Ann. Med.* **36**, 379–388 (2004).
84. Garnis, C., Buys, T. P. & Lam, W. L. Genetic alteration and gene expression modulation during cancer progression. *Mol. Cancer* **3**, 9 (2004).
85. Sung, J. *et al.* Oncogene regulation of tumor suppressor genes in tumorigenesis. *Carcinogenesis* **26**, 487–494 (2005).
86. Misek, D. E., Imafuku, Y. & Hanash, S. M. Application of proteomic technologies to tumor analysis. *Pharmacogenomics* **5**, 1129–1137 (2004).
87. Cowherd, S. M., Espina, V. A., Petricoin, E. F. 3rd & Liotta, L. A. Proteomic analysis of human breast cancer tissue with laser-capture microdissection and reverse-phase protein microarrays. *Clin. Breast Cancer* **5**, 385–392 (2004).
88. Okuda T *et al.* The profile of Hmlh1 methylation and microsatellite instability in colorectal and non-small cell lung cancer. *Int. J. Mol. Med.* **15**, 85–90 (2005).
89. Rook, M. S., Delach, S. M., Deyneko, G., Worlock, A. & Wolfe, J. L. Whole genome amplification of DNA from laser capture-microdissected tissue for high-throughput single nucleotide polymorphism and short tandem repeat genotyping. *Am. J. Pathol.* **164**, 23–33 (2004).
90. Grieu, F., Joseph, D., Norman, P. & Iacopetta, B. Development of a rapid genotyping method for single nucleotide polymorphisms and its application in cancer studies. *Oncol. Rep.* **11**, 501–504 (2004).
91. Ohno, M. *et al.* Lymphagenesis correlates with expression of vascular endothelial growth factor-C in colorectal cancer. *Oncol. Rep.* **10**, 939–943 (2003).
92. Hicklin, D. J. & Ellis, L. M. Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *J. Clin. Oncol.* **23**, 1011–1027 (2005).
93. Demel, U. *et al.* Detection of tumor cells in the peripheral blood of patients with breast cancer, development of a new sensitive and specific immunomolecular assay. *J. Exp. Clin. Cancer Res.* **23**, 465–468 (2004).
94. Muller, V. & Pantel, K. Bone marrow micrometastases and circulating tumor cells: current aspects and future perspectives. *Breast Cancer Res.* **6**, 258–261 (2004).
95. Eissa, S. *et al.* Diagnostic value of urinary molecular markers in bladder cancer. *Anticancer Res.* **23**, 4347–4355 (2003).
96. Pepe, M. S. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93**, 1054–1061 (2001).
97. Colditz, G. A. & Hankinson, S. E. The Nurses' Health Study: lifestyle and health among women. *Nature Rev. Cancer* **5**, 388–396 (2005).
98. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
99. Ransohoff, D. F. Rules of evidence for cancer molecular-marker discovery and validation. *Nature Rev. Cancer* **4**, 309–314 (2004).
100. The National Cancer Institute Breast and Prostate Cancer Cohort Consortium. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nature Rev. Cancer* **5**, 977–985 (2005).
101. Mosteller, F. & Colditz, G. Understanding research synthesis (meta-analysis). *Ann. Rev. Public Health* **17**, 1–32 (1996).
102. Webb, P., Bain, C. & Pirozzo, S. *Esential Epidemiology* (Cambridge University Press, Cambridge, 2005).
103. Whittemore, A. S., Harris, R. & Itnyre, J. Characteristics relating to ovarian cancer risk: collaborative analysis of 12 US case–control studies. II. Invasive epithelial ovarian cancers in white women. *Am. J. Epidemiol.* **136**, 1184–1203 (1992).
104. Grodstein, F. *et al.* Postmenopausal hormone use and risk of colorectal cancer and adenoma. *Ann. Intern. Med.* **128**, 705–712 (1998).
105. Herbst, A., Kurman, R., Scully, R. & Poskanzer, D. Clear-cell adenocarcinoma of the genital tract in young females: registry report. *N. Engl. J. Med.* **287**, 1259–1264 (1972).

106. Thun, M. J., Henley, S. & Patrono, C. Nonsteroidal anti-inflammatory drugs as anticancer agents: mechanistic, pharmacologic, and clincal issues. *J. Natl Cancer Inst.* **94**, 252–266 (2002).
107. Bernstein, L. *et al.* Tamoxifen therapy for breast cancer and endometrial cancer risk. *J. Natl Cancer Inst.* **91**, 1654–1662 (1999).
108. Bergman-Jungestrom, M., Gentile, M., Lundin, A. C. & Wingren, S. Association between *CYP17* gene polymorphism and risk of breast cancer in young women. *Int. J. Cancer* **84**, 350–353 (1999).
109. United States Department of Health and Human Services. in *Reducing Tobacco Use: A Report of the Surgeon General — Executive Summary*. 7 (United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, Georgia, 2000).

Acknowledgements
We are most thankful for the critical contribution of peer reviewers.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ATM | *BRCA1* | *BRCA2* | PSA | *TP53*
National Cancer Institute: http://www.cancer.gov
breast cancer | colon cancer | endometrial cancer | lung cancer | non-Hodgkin lymphoma | prostate cancer

FURTHER INFORMATION
American Cancer Society Cancer Prevention Study II: http://www.cancer.org/docroot/RES/content/RES_6_1_History_and_Accomplishments.asp
IARC Handbooks on Cancer Prevention: http://www.iarc.fr/IARCPress/general/prev.pdf
IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: http://monographs.iarc.fr/
IARC Press: http://www.iarc.fr/IARCPress/index.php
National Center for Health Statistics, National Health and Nutrition Survey: http://www.cdc.gov/nchs/nhanes.htm
NCI Early-Detection Research Network: http://www3.cancer.gov/prevention/cbrg/edrn/
US Surgeon General Reports on Smoking and Health: http://profiles.nlm.nih.gov/NN/ListByDate.html
Access to this interactive links box is free online.

SCIENCE AND SOCIETY

Locus-specific mutation databases: pitfalls and good practice based on the p53 experience

Thierry Soussi, Chikashi Ishioka, Mireille Claustres and Christophe Béroud

Abstract | Between 50,000 and 60,000 mutations have been described in various genes that are associated with a wide variety of diseases. Reporting, storing and analysing these data is an important challenge as such data provide invaluable information for both clinical medicine and basic science. Locus-specific databases have been developed to exploit this huge volume of data. The p53 mutation database is a paradigm, as it constitutes the largest collection of somatic mutations (22,000). However, there are several biases in this database that can lead to serious erroneous interpretations. We describe several rules for mutation database management that could benefit the entire scientific community.

Progress has been made over recent years in the cloning of the genes involved in both monogenic and polygenic disorders, including complex diseases such as cancer[1]. For each of these genes, numerous alterations of various types have also been described, ranging from point mutations to large deletions. The future development of new high throughput methods for the detection of mutations will lead to an enormous increase in the detection of new mutations[2]. It is difficult to evaluate the number of mutations reported in the literature to date (more than 50,000 have been collected in various databases), but a similar number could remain unpublished. It is also impossible to predict how many new mutations will be detected over the next 10 years, and the reporting and analysis of these mutations will therefore constitute a major challenge for the future[3,4]. Nevertheless, a number of points can be predicted. First, knowledge of these mutations will be important for treatment decisions as well as for basic science. And second, changes in our environment will lead to variations in the mutational events that modify our genome. Such changes will alter the distribution and/or pattern of mutations leading to the discovery of new and specific hot-spot mutations, so databases will need to be constantly updated. A good example of this is the specific mutation of *TP53* at codon 249 that is only found in hepatocellular carcinoma (HCC) that occurs in countries with a high exposure to the food contaminant aflatoxin B1. In these regions, the high specificity of the 249 mutation has enabled the development of a very sensitive diagnostic procedure that should not be applied when exposure to aflatoxin B1 is not (or no longer) evident[5].

The practical value of mutation analysis

All studies performed to date show that mutations are, in general, not randomly distributed. Hot-spot regions have been demonstrated, corresponding to a region of DNA that is susceptible to mutations (such as CpG dinucleotides), a codon encoding a key residue in the biological function of the protein, or both (BOX 1). Identification of these hot-spot regions and natural mutants is essential to define crucial regions in an unknown protein. In large genes such as neurofibromin 1 (*NF1*; 59 exons, 2,818 amino acids), retinoblastoma 1 (*RB1*; 27 exons, 928 amino acids), adenomatosis polyposis coli (*APC*; 15 exons, 2,843 amino acids), breast cancer 1 (*BRCA1*; 24 exons, 1,863 amino acids) and the titin gene (*TTN*; 363 exons, approximately 25,000 amino acids), detection of point mutations by direct sequencing analysis is difficult because of the size of the target gene. Identification of a hot-spot region allows analysis to be focused on this region, keeping in mind that a negative result should be viewed with caution.

It has also been clearly demonstrated that alterations in a single gene can cause various types of disorders. For example, mutations in *RET* are associated with multiple endocrine neoplasia types IIA[6] and IIB[7], familial medullary thyroid carcinoma[8] and a non-cancerous disorder known as Hirschsprung disease[9,10]. Mutations seem to be localized in specific domains of the protein for each of these disorders. The site of specific alterations at various positions in a given gene is also associated with different clinical features, as in the case of colon cancer and mutations in *APC*. A mutation in the C-terminus of the protein is specifically associated with a secondary abnormality, congenital hypertrophy of the retinal pigment epithelium[11], whereas mutations in the N-terminus are associated with an attenuated phenotype[12]. Analysis of mutations can also lead to the definition of risk factors. For instance, von Hippel–Lindau (VHL) families with mutations in *VHL* that result in truncated proteins have an increased frequency of renal-cell carcinoma (83%) compared with families with *VHL* missense mutations (54%)[13]. In diseases that are characterized