Hypothesis/Commentary

Cancer Epidemiology, Biomarkers & Prevention

Tumor-Based Case–Control Studies of Infection and Cancer: Muddling the When and Where of Molecular Epidemiology

Eric A. Engels, Sholom Wacholder, Hormuzd A. Katki, and Anil K. Chaturvedi

Abstract

We describe the "tumor-based case–control" study as a type of epidemiologic study used to evaluate associations between infectious agents and cancer. These studies assess exposure using diseased tissues from affected individuals (i.e., evaluating tumor tissue for cancer cases), but they must utilize nondiseased tissues to assess control subjects, who do not have the disease of interest. This approach can lead to exposure misclassification in two ways. First, concerning the "when" of exposure assessment, retrospective assessment of tissues may not accurately measure exposure at the key earlier time point (i.e., during the etiologic window). Second, concerning the "where" of exposure assessment, use of different tissues in cases and controls can have different accuracy for detecting the exposure (i.e., differential exposure misclassification). We present an example concerning the association of human papillomavirus with various cancers, where tumor-based case–control studies likely overestimate risk associated with infection. In another example, we illustrate how tumor-based case–control studies of *Helicobacter pylori* and gastric cancer underestimate risk. Tumor-based case–control studies can demonstrate infection within tumor cells, providing qualitative information about disease etiology. However, measures of association calculated in tumor-based case–control studies are prone to over- or underestimating the relationship between infections and subsequent cancer risk. *Cancer Epidemiol Biomarkers Prev; 23(10); 1959–64. ©2014 AACR.*

Introduction

Molecular epidemiology has made great strides in uncovering the causes of cancer. We wish to call attention to a type of retrospective case–control study that assesses infections as risk factors for cancer, and utilizes molecular tests on tumor tissue to assign exposure status for the cancer cases (1–6). The advantages and limitations of this "tumor-based case–control study" design and appropriate interpretation of such studies have not been described. In this Commentary, we review limitations in making inferences from tumor-based case–control studies, which arise due to (i) the retrospective timing of exposure assessment and (ii) differential exposure assessment for cases and controls. We illustrate how these limitations can lead to biased estimates of the strength of association between infections and cancer risk.

What Is a Tumor-Based Case–Control Study?

As an initial step in assessing whether a viral, bacterial, or parasitic infection causes cancer, tumor tissue from cases can be evaluated for markers of that infection that plausibly relate to exposure. For example, if the exposure is a viral infection, patient tumor specimens can be tested for the presence of the viral genome. Preliminary studies of this type, which evaluate solely human tumor tissues, are considered case series.

As a next step, investigators sometimes conduct a tumor-based case–control study. Importantly, the purpose of tumor-based case–control studies, as in other case–control studies, is framed as assessing the association between the infection and cancer. Although the tumor-based case–control study is similar to a case series, in that it uses the tumor tissues from cases to assess their exposure status, as a type of case–control study it always includes control subjects without the cancer of interest.

The use of tumor tissue to assess exposure in cancer cases may seem natural in tumor-based case–control studies. However, because control subjects do not have the cancer of interest, another tissue must be assessed for them. Therefore, tumor-based case–control studies differ from other retrospective case–control studies, which use the same tissues for assessing both cases and controls.

In a tumor-based case–control study, two approaches for selecting tissues for control subjects are readily available (Table 1). Among controls, the tissues evaluated for exposure can include normal tissue from the same body site that gives rise to the cancer (e.g., normal breast tissue for comparison with breast cancer tumors). Another common approach is to use "sentinel" specimens in contact with these body sites, for example, skin swabs to reflect

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland.

Corresponding Author: Eric A. Engels, Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 9609 Medical Center Drive, Room 6E226 MSC 9767, Bethesda, MD 20892. Phone: 240-276-7186; Fax: 240-276-7806; E-mail: engelse@exchange.nih.gov

doi: 10.1158/1055-9965.EPI-14-0282

Table 1. Tumor-based case–control study design variants

Tumor-based case–control study description	Control subject tissues assessed for exposure	Comments
Case vs. control, normal tissue	Normal tissue from the same body site as affected in cases	Normal tissues can be difficult to obtain for some sites
Case vs. control, sentinel sample	Body fluid or other sample in contact with affected site in cases (e.g., cervical lavage for comparison with cervical cancer, skin swabs for comparison with skin cancer)	This approach requires that the exposure can be reliably assessed using the sentinel sample

the exposure of the skin, exfoliated cells or cervical lavage specimens to reflect the status of the cervix, or buccal rinse specimens to reflect the status of the oral cavity or oropharynx. Investigators can then utilize highly sensitive and specific techniques to detect infection in these tissues, as well as tumor tissues from cases, assaying for microbial DNA (e.g., using PCR), RNA (*in situ* hybridization), or proteins (IHC).

What Are the Problems with Tumor-Based Case–Control Studies?

The difficulties in tumor-based case–control studies arise from exposure misclassification. These can be understood in comparison with one standard epidemiologic design, the cohort study. In a cohort study, exposure is assessed at subjects' baseline visit (and perhaps additional time points), and subjects are followed subsequently for disease occurrence. Typically, an exposure exerts its causal effect during an "etiologic window" of time. Because cancers develop over a prolonged period, the etiologic window can be years (or even decades) before development of disease. Cohort studies attempt to assess exposure status in this window. Measuring exposure before development of disease in the cases establishes a temporal relationship between exposure and outcome (7). In addition, measuring exposure uniformly, that is, using the same tissue and assay method for future cases and noncases, reduces the likelihood that any exposure misclassification is differential by case–control status.

In a case–control study (the other major type of epidemiologic study), cases with disease are evaluated along with a sample of nondiseased control subjects. A nested case–control study can be viewed as an efficient way to sample subjects from a large cohort study (2), and in such a study, the investigator can assess the subjects' prior exposure status using previously collected biospecimens. In contrast, in a retrospective case–control study, exposures can only be evaluated using information or biospecimens collected at the time of selection. A key consideration is how well this retrospective assessment reflects subjects' earlier exposure status, especially during the etiologic window. In most retrospective case–control studies, the investigator uses similar biological specimens for cases and controls, and applies uniform methods to assess exposure. Thus, retrospective case–control studies suffer from a lack of information on the temporal relationship between exposure and outcome, but in most such studies, uniform exposure assessment helps ensure that exposure misclassification is nondifferential by case–control status.

A tumor-based case–control study, as we describe it, is similar to other retrospective case–control studies in starting with cases who already have the disease and a sample of people without the disease. Similar to other retrospective case–control studies, tumor-based case–control studies lack information on the temporal relationship between exposure and outcome, so they cannot provide an unambiguous measure of exposure during the etiologic window. In addition, the unique issue for tumor-based case–control studies is that the exposure status of case and control subjects is assessed using different tissues, which has the potential to yield a different assessment.

These two issues of tumor-based case–control studies both relate to exposure misclassification—we describe them as "when" and "where" issues:

1. The "when" of exposure assessment: retrospective assessment may not accurately reflect exposure during the etiologic window.
2. The "where" of exposure assessment: use of different tissues for cases and controls may not provide comparable exposure assessment.

Retrospective ascertainment of exposure ("when") is problematic for all retrospective case–control studies. For example, infection may have first occurred only after the etiologic window, leading to a false-positive assessment (i.e., low specificity) for exposure based on the retrospective assessment at the time of subject selection. Also, some infections may be cleared over time, leading to low sensitivity for retrospective assessment (8).

With respect to the "where" issue, use of various tissues in control subjects can have low sensitivity. For example, within an organ, infections may be focal and microscopic, and a random biopsy can miss the site of infection. Furthermore, even when the infection is present, the amount of microbial material in control tissues may be very small. In contrast, lack of sensitivity typically affects cases less. This situation is likely when viruses are the agents of interest, because if the virus caused the cancer, it is usually reasonable to assume that it will be present in at least one copy per tumor cell. Thus, clonal multiplication of infected tumor cells ensures a reasonable sensitivity for many molecular tests of tumor tissue in cases.

Both the "when" and "where" issues can be present simultaneously in tumor-based case–control studies, because use of different tissues for cases and controls may combine to cause differential assessment of exposure for the earlier etiologic window. For example, among controls, infection may only be detected intermittently in sentinel samples, whereas infection may be persistently detectable in case tissues. This difference leads to lower sensitivity for detecting exposure during the etiologic window for controls than cases, leading to a spuriously positive association. Alternatively, changes in tumors and surrounding tissues as the cancer develops can reduce the ability of infections (especially bacteria) to persist. This "disease effect" decreases sensitivity selectively among cases, potentially leading to spuriously reduced associations relative to the truth.

Examples

Human papillomavirus and cancer

Substantial evidence links human papillomavirus (HPV) to cervical, anal, penile, vulvar, vaginal, and oropharyngeal cancers (1). The model whereby HPV is posited to cause cancer is through continued expression of viral oncoproteins (1). Some initial data supporting etiologic associations were derived from tumor-based case–control studies in which tumor tissues were assessed for HPV using molecular techniques (e.g., PCR for HPV DNA; refs. 1–4); HPV status of controls without cancer was evaluated using sentinel samples, for example, cells obtained through swabs, rinses, or lavages. Importantly, the etiologic window during which HPV infection first affects progression to cancer is many years before diagnosis (1, 8).

With respect to the "when" of exposure assessment, retrospective assessment of HPV infection is somewhat problematic. This issue may be especially difficult for controls, among whom HPV infection may have been transient. For example, among healthy young women, incident cervical HPV infections are frequently cleared. HPV clearance rates are not well characterized for other sites of infection. These considerations complicate interpretation of results from controls as reflecting their HPV status during the preceding etiologic window. For cases, in contrast, one may argue that evaluation of tumor tissues has high sensitivity for HPV infection during the etiologic window (i.e., if the virus caused the cancer, it must still be detectable in the tumor). However, detection of HPV in the tumor may have low specificity, because the virus may have been acquired later and be present in the tumor coincidentally. Evidence that HPV can be present coincidentally in tumors is provided by the observation that multiple HPV subtypes can sometimes be detected in a tumor, even though only one of the subtypes is causally related (9).

Alternative approaches may help increase the specificity of exposure assessment during the etiologic window in cases. For example, some viruses (including HPV) can integrate into host DNA (1). In such instances, a clonal pattern of viral integration within a tumor (demonstrated through tumor sequencing) supports that viral infection of the tumor cells occurred before the key neoplastic steps that led to clonal tumor growth, i.e., during the etiologic window. However, testing for clonality of viral integration cannot be used for normal tissues in controls, because the control tissues are not themselves clonal.

About the "where" of exposure assessment, it is important to consider that the researcher typically wishes to assess exposure for the entire person, the relevant organ, or the particular tissue at risk of cancer. One may posit that HPV detection is highly specific for concurrent infection for both cases and controls. However, for some sites of interest, sensitivity for detection of HPV infection likely varies between cases and controls based on differences in the evaluated biospecimens.

For oropharyngeal cancer, for example, the assessment of tumor tissue is highly sensitive for detecting HPV in case tumors, and it is probably very sensitive for detecting infection in the entire organ (i.e., the whole oropharynx) as well. However, it is difficult to determine the sensitivity of oral rinses used for assessing HPV exposure among controls. Oral rinses frequently miss HPV infection in oropharyngeal cancer cases (10, 11). If sensitivity of oral rinses is low in cases, it could be even lower in controls, because controls likely have only low-level or localized infection. These considerations point to differential sensitivity in assessment of oropharyngeal infection in cases and controls as a result of relying on different tissue types (e.g., tumor tissues vs. oral rinses). In contrast, the "where" issue would not be as important in assessing controls in a study of cervical cancer, because cytobrushes and cervical lavages have high sensitivity for detecting cervical HPV infection.

Helicobacter pylori and gastric cancer

Helicobacter pylori, a bacterium, is an established cause of gastric cancer (1). *H. pylori* infects the stomach lumen adjacent to the epithelial lining, and induces chronic inflammation and gastric atrophy. These changes set the stage for development of gastric cancer. However, as cancer develops, the stomach becomes less supportive of *H. pylori* infection, leading to loss of the bacterium (i.e., a disease effect; refs. 1, 12–14). Some tumor-based case–

control studies of *H. pylori* and gastric cancer have evaluated tumors from cases and random gastric biopsies from noncancer control subjects (e.g., patients with gastritis), using various staining procedures to identify the bacterium (5, 6).

The challenge for any retrospective case–control study of gastric cancer is the accuracy of its assessment of *H. pylori* infection status during the etiologic window ("when" issue). For a tumor-based case–control study, use of gastric biopsies for controls plausibly has reasonable sensitivity both for current infection and infection during the earlier etiologic window, because *H. pylori* often causes diffuse and longstanding gastritis. However, use of tumor tissue for exposure assessment in cases likely would have lower sensitivity because of the disease effect (both a "when" and "where" issue).

Discussion

Tumor-based case–control studies, which utilize diseased tissue to assess exposure status in cases, have been used to study infections as cancer risk factors. One reason why investigators may consider tumor-based case–control studies attractive is that it is natural to interpret molecular evidence for a microbe in tumor tissue as reflecting infection that predates development of the cancer. It therefore seems straightforward to define exposure for cases in that way, and it then only appears necessary to obtain some comparison tissues from controls. However, we argue that this approach suffers from difficulties in assessing exposure during the etiologic window (the "when" issue) and a lack of comparable tissues for control subjects from which to assess exposure (the "where" issue). We summarize limitations of tumor-based case–control studies in Table 2 in the context of evaluating associations between infections and cancer.

Exposure assessment is a challenge for any epidemiologic study. Some assays for infection (e.g., serum antibody assays for HPV or *H. pylori* infection) may lack sensitivity or specificity. Therefore, even if the assays are applied uniformly for all subjects, using the same types of biologic sample, they may still produce a biased measure of association. If the misclassification does not differ between cases and noncases, then the bias is typically toward the null. A cohort study can assess exposure at a defined time point before development of cancer, and assessment is uniform for cases and controls. All retrospective case–control studies face challenges from the "when" issues that we describe, even though most typically utilize the same method of exposure assessment in cases and controls. The subset of tumor-based retrospective case–control studies uniquely suffer from the "where" issue, which results in differential exposure misclassification.

Because of this vexing "where" issue, tumor-based case–control studies may be expected to yield measures of association (e.g., ORs) between infection and cancer that are upwardly or downwardly biased away from the true association. For example, for many viral infections, differential exposure assessment yields good sensitivity in case tumors but lower sensitivity in control tissues. The difference in the tissue types will lead to inflated ORs even for high-quality assays. A bias in the opposite direction can occur in tumor-based case–control studies of other infections (e.g., *H. pylori*), if the infection is lost over time as the tumor develops (12, 15).

Table 2. Issues in tumor-based case–control studies of infection and cancer that lead to biased measures of association

Issue
When issues: exposure assessment does not reflect status during the etiologic window
Infection may be transient
Infection may occur after etiologic window
Case tissues may lose infection over time (disease effect)[a]
Where issues: exposure assessment differs for cases and controls, leading to differential exposure misclassification
Infection may be focal or low-level in control tissues, making it difficult to detect
Case tissues may lose infection over time (disease effect)[a]
Case and control tissues may not reflect overall infection status of person or tissue of interest

[a]A disease effect causes both "when" and "where" issues, because it causes differential assessment with regard to exposure status during the etiologic window (see the text for further details).

Unfortunately, a measure of association, such as an OR, is usually interpreted as providing information on the relative risk for developing cancer in exposed and unexposed people. Given the biases that we describe, we believe it is inappropriate to make quantitative interpretations of ORs from tumor-based case–control studies, such as, "exposure is X-fold more common in cancer cases than controls," or even more problematically, "exposure is associated with an X-fold increased risk of developing cancer." Unfortunately, we have seen both interpretations in the published literature. Even when the true association is quite strong, the bias in the estimates derived from tumor-based case–control studies can have important downstream ramifications. For example, clinicians may attempt to use such results for risk stratification, or public health researchers may use them to predict disease burden. Those models will produce unreliable results if the strength of the association is incorrectly estimated.

Meta-analyses are an important source of confusion in equating tumor-based case–control studies with standard case–control studies. Many primary studies comparing case tumors with normal tissues present their findings in terms of the proportion of specimens of each tissue type that exhibits evidence for the infection of interest, and do not present measures of association, such as an OR.

However, meta-analyses that summarize the evidence linking an infection to cancer include these studies and derive ORs, thereby converting them into *de facto* tumor-based case–control studies. The meta-analyses then frequently present summary OR estimates that are interpreted to reflect associations between the exposure and risk of subsequently developing the cancer. Examples of such meta-analyses include evaluations of HPV and bladder, breast, and laryngeal cancers (16–18); Epstein–Barr virus and breast cancer (19); simian virus 40 and mesothelioma, brain tumors, sarcomas, non-Hodgkin lymphoma, and colon cancer (20); and *H. pylori* and liver cancer (21, 22). Although systematic reviews of published studies are always informative, we suggest that investigators refrain from quantitative summaries, or they should clearly state that differences between groups should not be interpreted as measures of risk.

If tumor-based case–control studies do not reliably provide a measure of association between risk factor and outcome, they can still serve a useful role, especially early in a line of research when little is known. They can provide qualitative laboratory evidence for a possible relationship between the exposure and cancer. For instance, because persistent expression of viral oncogenes is often thought to be required in viral carcinogenesis, frequent detection of infection in case tumors helps support an etiologic role; such evidence is captured under Bradford Hill's criteria for causality as "coherence" (7). The key distinction is that tumor-based studies do not yield reliable quantitative measures of association, which are required under Bradford Hill's criterion of "strength of association" (7). Tumor-based case–control studies can also serve as helpful benchmarks in assessing the reliability of laboratory testing, by demonstrating consistent differences in detection of infection between different tissues (e.g., by ruling out PCR contamination).

As a final point, we note that the tumor-based case–control study design differs from other case–control studies that use tumor tissue only to separate cases into subgroups, not for exposure assessment. A classic example is the division of breast cancer cases according to tumor expression of hormone receptors (23), and the approach has also been used in case–control studies to divide cancer cases into infection-positive and negative subgroups (24). Investigators have then investigated whether the case subgroups exhibit different risk factors, by uniformly assessing all of the cases and controls for another exposure of interest. Because of its uniform exposure assessment, this approach does not suffer from the "where" issue we describe for tumor-based case–control studies.

In summary, standard cohort and case–control study designs provide the strongest framework for measuring associations between infections and cancer. Given the "when" and "where" issues faced by tumor-based case–control studies, their results, particularly measures of association, should be interpreted with caution.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: E.A. Engels, S. Wacholder, H.A. Katki, A.K. Chaturvedi
Development of methodology: S. Wacholder
Writing, review, and/or revision of the manuscript: E.A. Engels, S. Wacholder, H.A. Katki, A.K. Chaturvedi

Grant Support

This work was supported by the Intramural Research Program of the National Cancer Institute.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 13, 2014; revised July 8, 2014; accepted July 21, 2014; published OnlineFirst July 25, 2014.

References

1. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. A review of human carcinogens. Part B: Biological agents. Lyon: IARC; 2009.
2. Strome SE, Savva A, Brissett AE, Gostout BS, Lewis J, Clayton AC, et al. Squamous cell carcinoma of the tonsils: a molecular analysis of HPV associations. Clin Cancer Res 2002;8:1093–100.
3. Smith EM, Ritchie JM, Summersgill KF, Hoffman HT, Wang DH, Haugen TH, et al. Human papillomavirus in oral exfoliated cells and risk of head and neck cancer. J Natl Cancer Inst 2004;96:449–55.
4. Castellsague X, Diaz M, de SS, Munoz N, Herrero R, Franceschi S, et al. Worldwide human papillomavirus etiology of cervical adenocarcinoma and its cofactors: implications for screening and prevention. J Natl Cancer Inst 2006;98:303–15.
5. Kokkola A, Valle J, Haapiainen R, Sipponen P, Kivilaakso E, Puolakkainen P. Helicobacter pylori infection in young patients with gastric carcinoma. Scand J Gastroenterol 1996;31:643–7.
6. Kim HY, Cho BD, Chang WK, Kim DJ, Kim YB, Park CK, et al. Helicobacter pylori infection and the risk of gastric cancer among the Korean population. J Gastroenterol Hepatol 1997;12:100–3.
7. Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295–300.
8. Wacholder S. Chapter 18: statistical issues in the design and analysis of studies of human papillomavirus and cervical neoplasia. J Natl Cancer Inst Monogr 2003;31:125–30.
9. Li N, Franceschi S, Howell-Jones R, Snijders PJ, Clifford GM. Human papillomavirus type distribution in 30,848 invasive cervical cancers worldwide: variation by geographical region, histological type and year of publication. Int J Cancer 2011;128:927–35.
10. D'Souza G, Kreimer AR, Viscidi R, Pawlita M, Fakhry C, Koch WM, et al. Case-control study of human papillomavirus and oropharyngeal cancer. N Engl J Med 2007;356:1944–56.
11. Herrero R, Castellsague X, Pawlita M, Lissowska J, Kee F, Balaram P, et al. Human papillomavirus and oral cancer: the International Agency for Research on Cancer multicenter study. J Natl Cancer Inst 2003;95:1772–83.
12. Helicobacter and Cancer Collaborative Group. Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. Gut 2001;49:347–53.

13. Hu PJ, Mitchell HM, Li YY, Zhou MH, Hazell SL. Association of Helicobacter pylori with gastric cancer and observations on the detection of this bacterium in gastric cancer cases. Am J Gastroenterol 1994;89:1806–10.
14. Karnes WE Jr, Samloff IM, Siurala M, Kekki M, Sipponen P, Kim SW, et al. Positive serum antibody and negative tissue staining for Helicobacter pylori in subjects with atrophic body gastritis. Gastroenterology 1991;101:167–74.
15. Danesh J. Helicobacter pylori infection and gastric cancer: systematic review of the epidemiological studies. Aliment Pharmacol Ther 1999;13:851–6.
16. Gutierrez J, Jimenez A, de Dios LJ, Soto MJ, Sorlozano A. Meta-analysis of studies analyzing the relationship between bladder cancer and infection by human papillomavirus. J Urol 2006;176:2474–81.
17. Li X, Gao L, Li H, Gao J, Yang Y, Zhou F, et al. Human papillomavirus infection and laryngeal cancer risk: a systematic review and meta-analysis. J Infect Dis 2013;207:479–88.
18. Li N, Bi X, Zhang Y, Zhao P, Zheng T, Dai M. Human papillomavirus infection and sporadic breast carcinoma risk: a meta-analysis. Breast Cancer Res Treat 2011;126:515–20.
19. Huo Q, Zhang N, Yang Q. Epstein-Barr virus infection and sporadic breast cancer risk: a meta-analysis. PLoS ONE 2012;7: e31656.
20. Vilchez RA, Kozinetz CA, Arrington AS, Madden CR, Butel JS. Simian virus 40 in human cancers. Am J Med 2003;114:675–84.
21. Xuan SY, Xin YN, Chen AJ, Dong QJ, Qiang X, Li N, et al. Association between the presence of H pylori in the liver and hepatocellular carcinoma: a meta-analysis. World J Gastroenterol 2008;14: 307–12.
22. Lorenzon L, Ferri M, Pilozzi E, Torrisi MR, Ziparo V, French D. Human papillomavirus and colorectal cancer: evidences and pitfalls of published literature. Int J Colorectal Dis 2011;26:135–42.
23. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 2011;103:250–63.
24. Gillison ML, D'Souza G, Westra W, Sugar E, Xiao W, Begum S, et al. Distinct risk factor profiles for human papillomavirus type 16-positive and human papillomavirus type 16-negative head and neck cancers. J Natl Cancer Inst 2008;100:407–20.