Tumor-based case-control studies of infection and cancer:
muddling the when and where of molecular epidemiology

Eric A. Engels¹, Sholom Wacholder¹, Hormuzd A. Katki¹, Anil K. Chaturvedi¹

¹ Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

Address correspondence to:
Eric A. Engels
Infections and Immunoepidemiology Branch
Division of Cancer Epidemiology and Genetics
National Cancer Institute
9609 Medical Center Dr, Room 6E226 MSC 9767
Bethesda, MD 20892
Telephone 240-276-7186, Fax 240-276-7806

The authors have no conflicts of interest to disclose. This research was supported by the Intramural Research Program of the National Cancer Institute.

Running title: Tumor-based case-control studies of infection and cancer

Key words: infections, cancer etiology, study design, methodology
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 in studying cancer cases), but they must utilize non-diseased 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 timepoint (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 regarding disease etiology. However, measures of association calculated in tumor-based case-control studies are prone to over- or under-estimating the relationship between infections and subsequent cancer risk.
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 1) the retrospective timing of exposure assessment, and 2) 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. While 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, since 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, e.g., skin swabs to reflect 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 polymerase chain reaction [PCR]), RNA (in situ hybridization), or proteins (immunohistochemistry).

**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 to one standard epidemiologic design, the cohort study. In a cohort study, exposure is assessed at subjects’ baseline visit (and perhaps additional timepoints), and subjects are followed subsequently for disease occurrence. Typically, an exposure exerts its causal effect during an “etiologic window” of time. Since 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). Additionally, measuring exposure uniformly—i.e., using the same tissue and assay method for future cases and non-cases—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 non-diseased 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. By 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 retrospective case-control studies, the investigator’s goal is to use similar biological specimens for cases and controls and apply 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 non-differential 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. Additionally, 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 derived from tumor-based case-control studies, in which tumor tissues were assessed for HPV using molecular
techniques (e.g., PCR for HPV DNA) (1-4); HPV status of controls without cancer was evaluated using sentinel samples, e.g., 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 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 (e.g., 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.

Regarding 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 cancers 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, 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 (e.g., 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, since 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) (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 non-cancer 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, since *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 biological sample, they may still produce a biased measure of association. If the misclassification does not differ between cases and non-cases, then the bias is typically toward the null. A cohort study can assess exposure at a defined timepoint prior to 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.

Due to this vexing “where” issue, tumor-based case-control studies may be expected to yield biased measures of association (e.g., odds ratios) 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 odds ratios 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).

Unfortunately, a measure of association, such as an odds ratio, 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 odds ratios 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 odds ratio. However, meta-analyses that summarize the evidence linking an infection to cancer include these studies and derive odds ratios, thereby converting them into *de facto* tumor-based case-control studies. The meta-analyses then frequently present summary odds ratio 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). While 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. Due to 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.
Reference List


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

<table>
<thead>
<tr>
<th>Tumor-based case-control study description</th>
<th>Control subject tissues assessed for exposure</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case vs. control, normal tissue</td>
<td>Normal tissue from the same body site as affected in cases</td>
<td>Normal tissues can be difficult to obtain for some sites.</td>
</tr>
<tr>
<td>Case vs. control, sentinel sample</td>
<td>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)</td>
<td>This approach requires that the exposure can be reliably detected in the sentinel sample.</td>
</tr>
</tbody>
</table>
Table 2. Issues in tumor-based case-control studies of infection and cancer that lead to biased measures of association

<table>
<thead>
<tr>
<th>Issue</th>
</tr>
</thead>
</table>
| **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).*                                                                                                                                                                                                                                     |
| **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).*  
Case and control tissues may not reflect overall infection status of person or tissue of interest.                                                                                                                                                                                                 |

Table 2 note  
* A disease effect causes both “when” and “where” issues, because it causes differential assessment with regards to exposure status during the etiologic window (see text for further details).
Tumor-based case-control studies of infection and cancer: muddling the when and where of molecular epidemiology


Cancer Epidemiol Biomarkers Prev  Published OnlineFirst July 25, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-14-0282

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

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

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