9 Design of Retrospective and Case-Control Studies in Oncology

Katherine S. Panageas, Debra A. Goldman, and T. Peter Kingham

9.1 Introduction

The objective of research studies is to make inferences about hypothesized relationships within a population. These relationships include differences in survival among treatment groups, various risk factors for surgical outcomes, differences in quality of life, and genetic variations among cancer subtypes. The study design used to answer the research question is critical for the ability to draw conclusions and is directly related to the statistical analysis methods that can be applied. Properly designed and executed studies provide the strongest level of empirical evidence.

9.1.1 Randomized Controlled Trials

The gold standard study design for clinical research is the randomized controlled trial (RCT), which is the most likely to minimize inherent biases. In RCTs, using a large enough sample size, randomization ensures that each patient has an equal chance of receiving a given treatment and that treatment groups are comparable with respect to any known or unknown factors that may affect the outcomes. In addition to eliminating selection bias, randomization provides a simple foundation for straightforward statistical analyses compared with observational studies. Despite being considered the gold standard, RCTs have several drawbacks. First, they are expensive and time-consuming, and they require organizational infrastructure to

K. S. Panageas, Dr.P.H. • D. A. Goldman, M.S.
Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

T. P. Kingham, M.D. (✉)
Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: kinghamt@mskcc.org

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_9

develop and conduct. Second, RCTs for infrequent or rare outcomes require sizable sample sizes, so these outcomes may be more practical to examine using alternative study designs. Third, RCTs may pose ethical problems or may not be feasible owing to difficulties with recruitment and compliance. Fourth, results from RCTs may not be generalizable to real-world populations or circumstances, in which the environment cannot be strictly controlled [1, 2].

9.1.2 Observational Studies

Observational studies are alternatives to RCTs, and can be either prospective or retrospective. Observational studies may be used in settings when it is unethical to randomize patients to receive specific treatments, or to provide preliminary evidence for hypotheses of RCTs.

9.1.2.1 Prospective Observational Studies

In a prospective observational study, data collection and the events of interest occur in a group of individuals, some of whom have had, currently have, or will have the exposure of interest, such as a certain treatment, to determine the association between that exposure and the outcome. However, prospective observational studies are limited to conditions that occur relatively frequently and to studies with relatively short follow-up periods, so that sufficient numbers of eligible individuals can be enrolled and followed within a reasonable study period.

9.1.2.2 Retrospective Observational Studies

All retrospective research studies are classified as observational studies because the allocation to treatment or assignment of factors is not under control of the investigator. In retrospective studies, the study sample is generated from secondary or preexisting data. The disease experience of the group between a defined time in the past and the present is then reconstructed from medical records. Compared with prospective studies, retrospective studies are inexpensive, as they make use of available information. Further, retrospective studies of rare conditions are much more efficient because individuals experiencing these rare outcomes can be found among patient records rather than the investigators needing to prospectively follow a large number of individuals to identify a few cases. Studies have shown that the majority of publications in clinical subspecialty journals are based on retrospective observational studies [3–5]. Also, as most medical centers transition from paper to electronic medical records and as computing power advances to handle ever larger data sets, retrospective studies are becoming easier and more efficient to conduct.

Retrospective studies have long-established use in surgical oncology [6–8]. Single-institution data or large multicenter efforts examining past experiences can serve many beneficial purposes, including generating hypotheses to develop future prospective studies, to explore ideas in translational laboratory research projects, or to compare results with previous studies that enrolled a smaller or more heterogeneous patient population. Further, retrospective analyses can provide critically

relevant data for populations known to be poorly represented in clinical trials—especially of cancer—including older adults and individuals with eligibility-restricting comorbidities. These analyses also may identify adverse events that are potentially unrecognized in the often highly homogenous groups of study participants. Finally, both the safety and the efficacy of treatment afforded by longer observation periods and more prolonged therapy can be revealed by retrospectively examining previously treated patients [9]. However, because retrospective studies do not involve randomization, the potential for significant biases exists, such as sample selection and recall and referral biases, which can limit the applicability and generalizability of these studies.

9.2 Types of Retrospective Studies

The historical cohort study and the case-control study are two of the most common retrospective designs. A retrospective **cohort study** comprises a sample of individuals (e.g., surgically resected pancreatic cancer patients) in whom we assess the relationship between risk factors and outcomes, such as post-surgical complication rates, disease recurrence, and overall survival. Risk factors are considered the **exposure**, a broad term used to denote any factor that is potentially related to the outcome of interest [10]. In contrast, in a retrospective **case-control study**, the outcome (e.g., post-operative complications) is measured *before* the exposure. Controls are selected from a pool of patients who have not experienced the outcome (Fig. 9.1). It is critical that the control group be as similar to the cases as possible in terms of other factors, such as demographic and treatment details [11, 12]. The retrospective case-control study is an important research strategy encountered in the medical

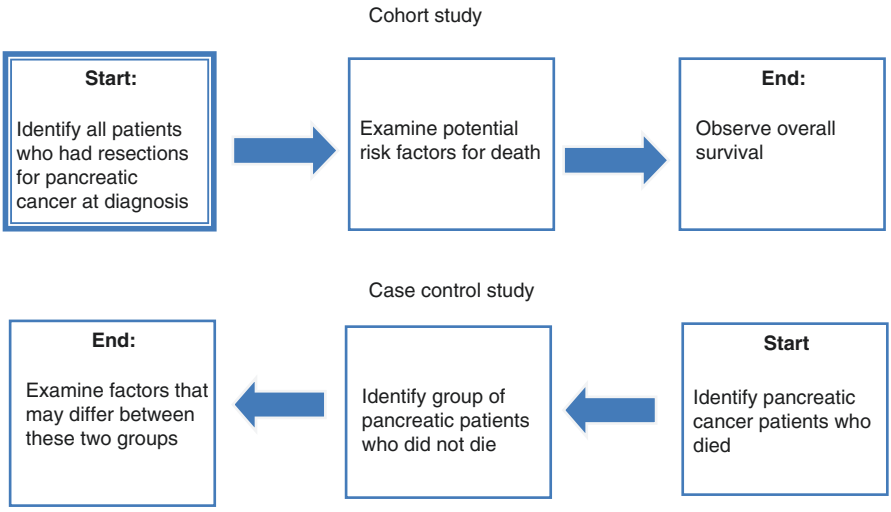

Fig. 9.1 Illustration of the differences between cohort and case-control studies

literature, and if carefully executed, can be an invaluable source of clinical information. Unfortunately, the retrospective viewpoint of case-control studies—looking "backwards" from an outcome event to an earlier exposure—is accompanied by numerous methodological hazards, including recall bias, which will be discussed later in this chapter.

Retrospective studies are often criticized on methodological grounds. Researchers must pay careful attention to selecting appropriate study groups, defining and detecting the outcome event, defining and ascertaining the exposure, assuring that the compared groups were equally susceptible to the outcome event at baseline, and performing careful statistical analysis. If systematic bias enters the research at any of these points, erroneous conclusions can result. In this chapter, we will cover design topics specific to retrospective studies, including validity, confounding, sample selection, sampling methods, missing data, and considerations for particular oncology outcomes.

9.3 Validity

The quality of a study depends on many factors, including internal and external validity. **Validity** is the degree to which a study result is likely to be true and free from bias [13]**.** As mentioned, **retrospective** study designs are inherently more susceptible to bias, given the lack of control over group assignment and the experiment environment. The study design and execution greatly determine the **internal validity. A study is internally valid if reported differences can be attributed to the exposure or intervention [14] and cannot be attributed to selection bias, information bias, or confounding. Confounding** is the distortion of the effect of one risk factor by the presence of another (Fig. 9.2). In randomized studies, confounding is typically accounted for in the randomization process. In retrospective studies, confounding can be controlled by restriction sampling, by matching on the confounding variable, or by accounting for it in the analysis using multivariable modeling.

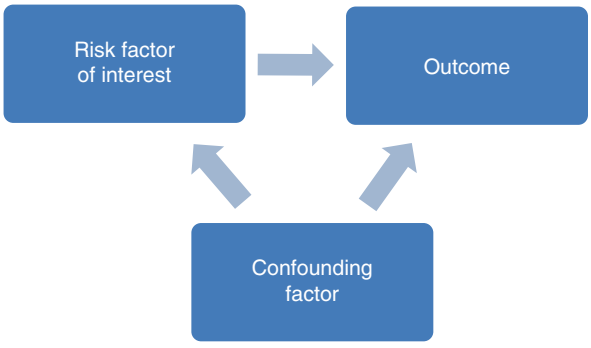

Fig. 9.2 Illustration of confounding

If a study is internally valid, it is important to assess whether it is has external validity as well. **External validity** refers to the extent to which the results can be generalized to other populations, other settings, and across time [15, 16]. For instance, if we build a model to predict survival in patients with incidental gallbladder cancer, we would want that model to be predictive for patients at other centers, in future years, and other circumstances. External validity is highly related to applying the appropriate sampling techniques, as we will now explore.

9.4 Sampling

Sampling refers to the process of selecting individuals to be included in a study. A **representative sample** is one in which the group sufficiently embodies the population that one is attempting to study, known as the **target population**. In retrospective samples, representativeness, generalizability, and sampling issues are important considerations. Unlike prospective research, in which one can control who is evaluated through enrollment and eligibility criteria, and one can control treatment environment and outcomes assessments, in retrospective research, one is limited by external factors that may have affected who is included in the study sample and who is not.

For example, in a study evaluating outcomes for gallbladder cancer over time, sampling may be limited to a single research institution. If the institution is a referral center for more complex cases or more advanced stage patients, the sample may not represent gallbladder cancer outcomes at other institutions or gallbladder cancer patients as a whole. Additionally, if data were retrieved from an institutional surgical database, sampling would be restricted to those patients who received consultation from a surgeon. Patients seen only by a medical oncologist would not be included, so findings could not be generalized to all patients with gallbladder cancer, but rather only to those who received surgery. Study site location is another important factor to consider. A study sample from a hospital in China is likely to contain mostly East Asian patients, whereas a study sample from the Netherlands is likely to contain mostly Western European patients. Differences in oncologic outcomes based on ethnic background are well documented [17–19] and present a challenge to generalizability.

As the above examples illustrate, some sampling issues are common or particular to oncology. For instance, treatment at a tertiary cancer center may be different from treatment at a community center, and studies of patients from a particular geographic region may not be generalizable to the disease as a whole. Ultimately, we may not be able to fully generalize our retrospective findings to the target cancer population; nevertheless, our findings make important contributions to the understanding of that disease. The sample that one can ultimately generalize to is considered the **accessible population**.

9.4.1 Selection Bias

If the study sample is not representative of the target population and the underlying exposure-outcome relationship, then the measures of association will be biased. Selection bias exists when a characteristic of the sample makes it different from the target population in a fundamental way that cannot be ignored [20]. The selection bias can affect both who is included in the study and the likelihood of people being retained or followed up within the study. Examples include differential patient referral or diagnosis; differential screening for disease or progression; selection of a comparison group that is not representative of the target population; or differential loss to follow-up in a cohort study, such that the likelihood of being lost to follow-up is related to one's outcome or one's exposure status [21, 22].

For example, in a retrospective study, we cannot control for variations in treatment, such as which patients received treatment, when patients received it, or which surgeon operated on which patients. In other scenarios, some patients may have received an additional diagnostic test whereas others did not, or some patients may have received genetic testing while others opted out or were not even offered the test. Taking the earlier example of gallbladder cancer outcomes, if only those patients who were seen by a surgeon were included, we would have introduced a selection bias into our study if we wanted to generalize to all patients with gallbladder cancer. Patients who were seen by a medical oncologist only and were not referred to a surgeon are part of the target population as defined. Thus, if the investigator's goal is to draw conclusions about all patients with gallbladder cancer, they should obtain data from other sources, such as medical oncology, so that all patients are represented. If the data are not available, then one may want to consider restricting the sample to only surgical gallbladder patients, recognizing that this limits the generalizability of the findings to gallbladder cancer patients who underwent surgical resection.

9.4.2 Information Bias

Information bias is a major limitation of retrospective studies, as the necessary data elements were not planned in advance. For example, reported post-operative complications depend on the complication being accurately documented in the medical record, and this information may not be available in the chart. In addition, the physician may have spoken with the patient and ordered a treatment from an outside pharmacy. Also, if one is trying to determine events from hospital billing records alone, not all medical events are documented in the International Classification of Diseases (ICD)-9/ICD-10 coding system, but only those that were related to medical billing charges. Therefore, some complications may be missing or incomplete. Though information bias itself may be unavoidable, using reproducible, systematic data collection methods will decrease the impact of errors arising from retrospective data capture.

9.4.3 Recall Bias

Recall bias is a specific type of information bias pertaining to the accuracy of data recalled from a time in the past. Recall bias occurs when patients are asked to recall symptom and/or treatment details that may have occurred months or years earlier. Examples of such bias include recalling age at menarche [23, 24] or the assessment of pain after a prior procedure [25].

9.4.4 The Denominator Problem

Being able to identify all patients eligible to be included in a retrospective study is a critical hurdle. Through a proxy, such as billing records, an institutional database may identify patients with a specific disease who had surgery. However, if patients were mistakenly billed for a different surgery (e.g., prostatectomy instead of prostate surgery), or the list of all possible billing codes is unknown, one could miss many patients. Further, it one's institution does not have electronic medical records or an institutional database, it may be extremely difficult or practically impossible to collect all possible patients. **Being unable to identify the number of potentially eligible patients is known as the denominator problem** [26]. This can be particularly troublesome for studies in which rates, such as post-operative complication rates or re-admission rates, need to be calculated. If not all patients were identified, these rates may be artificially higher than the true rate. The denominator problem is closely related to selection and information biases.

One common way to demonstrate how one's study sample reflects the total possible pool of patients is through flow charts. Flow charts are illustrations that demonstrate how one obtained the final sample from the initial group of patients. A flow chart enables others to get a sense of how common the inclusion criteria were and how exclusions shaped the final cohort. The following are examples of what information to include in a retrospective flow chart (Fig. 9.3):

9.4.5 Sample Selection Methods

Convenience sampling is a common selection method in retrospective research. In convenience sampling, one selects the cases that are easiest to obtain for the study. In retrospective research, this usually means that the sample is obtained from one's current institution, where one has access to the records, or is made up of patients that the researcher has treated. Because these patients are chosen for accessibility rather than representativeness, generalizability is a major problem in convenience sampling. It is important to note that, although one can employ probabilistic sampling techniques (e.g., systematic sampling) in a convenience sample to further refine it, if the larger cohort was not representative of the target population, the smaller study sample will not be generalizable either.

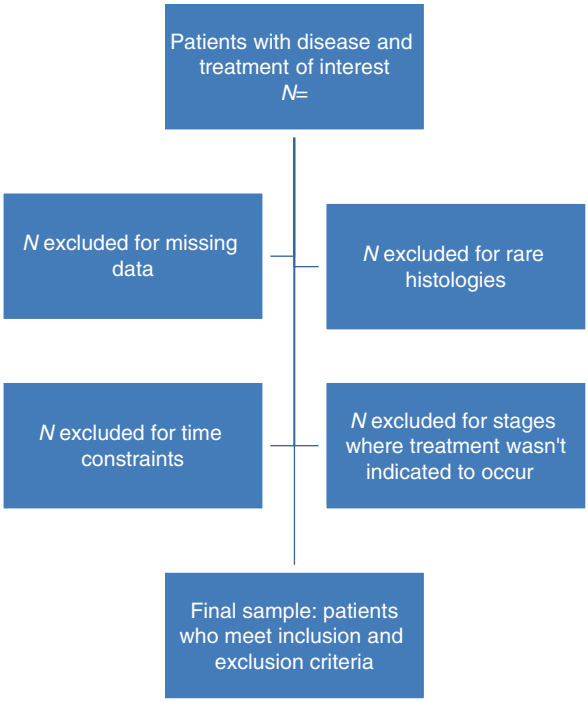

Fig. 9.3 Example flow chart

The first two techniques we will describe, simple random sampling and systematic sampling, are more commonly used in epidemiologic or population-based studies, in which one has a much larger cohort than is needed to answer the research question. However, these techniques can also be applied to retrospective clinical studies in which one does have enough resources or time to collect data on all patients.

In **simple random sampling,** patients are selected from a larger sample through random selection. The number of patients and range of values to be included in the study is decided a priori, a series of random numbers is generated, and each patient is assigned a random number. In this design, each patient has an equal chance of being selected. In **systematic sampling,** the full sample is taken from a defined time period and the patients are ordered chronologically. For instance, suppose we have all colon cancer cases diagnosed in the United States from 2004 to 2014. We order them from diagnosis date starting with January 1, 2004. The study sample is then selected using a systematic periodic rule, such as each 10th patient or every other patient in the list. One may use this technique when there is a large number of cases and it would be unfeasible to collect data on all patients.

Even with retrospective studies, it is important to balance the needs for resources and time with having a sufficient number of patients to enable one to confidently answer the research question. Although both simple random sampling and systematic sampling are valid for choosing a smaller sample of patients, if one does not have enough patients with the outcome of interest in the smaller sample, the overall validity of the study findings will be questionable.

Consecutive sampling refers to selecting all patients who meet the inclusion criteria within a specific time frame. Particularly for oncology, where many diseases and treatments are rare, consecutive sampling is an extremely popular technique. However, with consecutive sampling, heterogeneity, such as differences in treatment course or in patient characteristics, is introduced, and this must be balanced against the need to have a sufficient number of patients to study. In some instances, this heterogeneity (e.g., differences in neoadjuvant treatment before surgery) can be controlled for by adjusting for these factors in the model. Another strategy for handling heterogeneity is **restriction sampling.** Restriction sampling refers to limiting the sample to individuals within a certain range of values for a confounding factor, such as age, to reduce the effect of such a factor. For instance, suppose we wanted to study outcomes for gastrointestinal stromal tumors (GISTs). Since the approval of Gleevec® (imatinib mesylate; Novartis) in 2008, neoadjuvant treatment and subsequent outcomes have changed for GIST patients. Therefore, we may want to restrict our sample to patients treated after 2008 to avoid possible confounding due to known treatment outcome differences, or we may want to separately study patients from before and after 2008. Unfortunately, restriction sampling limits the generalizability of results to those within the same range of restricted values.

9.4.6 Matching

Matching is a technique used primarily in retrospective research projects to minimize differences between comparison groups. Although one can account for the differences by including these factors in a multivariable model, in some instances matching may be more efficient. For example, if the outcome of interest is relatively rare or the target sample is small, one may not be able to incorporate all the factors in the same multivariable model. Also, the control group one can pull from may be many times larger than the target group, and data collection may be unfeasible in such a large group. By matching, one can select a comparison group that is similar enough to the target group such that the relationship between the outcome and the exposure is not attributable to the confounding factors one bases the matches on.

In **frequency matching**, one matches based on the distribution of values. For instance, if 20% of the cases were stage 1 and 40% were stage 2, one would match the control group, so that approximately 20% of the controls were in stage 1 and 40% of the controls were in stage 2. In **individual matching**, one pairs each particular patient in the target sample with a patient in the control sample. For example, if a case was a 25-year-old female patient with adenocarcinoma, the control should also be a 25-year-old female patient with adenocarcinoma. Expanding on this strategy further, either the match can be exact, where the continuous variables are identical, or one can use **caliper matching**. In caliper matching, the values are allowed to differ within a specific range, called a caliper. It is common to set the caliper to 0.25 standard deviations [27], but other calipers have been used [28]. In individually matched samples, only those patients who matched would be included in the study; all others are dropped. As a result, exact matching may lead to excessive dropping. Therefore, caliper matching is a helpful strategy, particularly with regard to

covariates such as age, to prevent excessive sample loss and increase the likelihood of matching. Importantly, in individual matching, the comparison groups are no longer considered independent samples because the characteristics for the control group are dependent on what the characteristics were for the target group. Therefore, appropriate analytic methods to handle the dependency should be applied.

Propensity score matching is another technique that is used to reduce bias due to confounding variables. The propensity score is the probability of a patient receiving the treatment or experiencing the event conditional on specific factors or observed characteristics [5, 29–31]. In other words, if patients who are younger and have lower-stage cancer are more likely to receive treatment and are also less likely to die, these factors are confounding the relationship between the risk factor of interest and the outcome. Thus, these factors would be the ones to include in the propensity score. One can incorporate the propensity score into the study using several techniques: inverse probability weighting, stratification, covariate adjustment, and matching [32]. In propensity score matching, rather than matching being based on individual factors, one matches on the probability of being part of the target sample, which is determined before the matching process.

One can choose between matching with or without replacement. In **matching without replacement**, a patient can be matched to another patient only once, whereas in **matching with replacement**, a patient may be included for multiple target patients. Just as with any repeated measure, the fact that the patient appears multiple times needs to be accounted for in the analysis. Also within propensity score matching, one can choose between so-called **greedy matching** and **optimal matching**. Optimal propensity score matching chooses the match that minimizes the within-pair difference of the propensity score. In contrast, in greedy matching, a patient is first selected at random. Next, the control patient with the closest propensity score to this random subject is selected for matching. The term 'greedy' is used because the matching is not redone if that control subject would serve as a better match for the next randomly selected patient. That is, the patient stays matched regardless of the optimal benefit to the sample as a whole [32].

Similar to individual matching, with propensity score matching, one can set a threshold, or caliper, to decide how close the match should be. In **nearest-neighbor matching**, no restriction is made on the distance between the propensity score of the target and the control. In nearest-neighbor matching within a specified caliper distance, the propensity score is restricted by the caliper, or the maximum acceptable distance. This is similar to how calipers are used in traditional individual matching. Also, like individual matching, propensity score matching creates dependence between the cases and controls, so alternative analysis methods that account for the conditional nature of these samples should be employed [28, 29].

Propensity score matching was employed in a study on the relationship between protective lung ventilation during pulmonary resection and post-operative complications [33]. In this study, multiple factors were thought to be associated with the likelihood of receipt of protective lung ventilation and the occurrence of post-operative complications. Therefore, these factors could confound the

relationship between ventilation use and complications. Matched cohorts were created from clinically relevant factors including, but not limited to, the factors that differed between patients on ventilation versus those not on ventilation. After propensity score matching, the authors assessed whether the cohorts were well balanced. The authors then performed their primary analyses using these balanced cohorts.

Control patients can be matched to target patients at a rate of one control per case, or there can be multiple controls for one case. The former is referred to as **1:1 matching**, and the latter as **1:n matching**. The overall sample size increases with additional controls, which can increase the strength or power of the findings. However, the benefit of using additional matches depends on the distribution and size of the pool of possible controls, and little added benefit may exist beyond 1:1 or, at most, 1:3 matching [24, 27, 28]. Further, increasing the number of required matches per patient increases the chance that the case may not be matched given a fixed pool of controls. It should be noted that, in propensity score matching, it is possible to have a variable number of matches for each control, which has been shown to reduce bias [34].

Once patients are matched, it is critical to check that the characteristics one matched on are balanced between the two groups to ensure the matching was done correctly. Although matching can reduce confounding between groups, it introduces an additional layer of complexity into the analysis methods. Further, matching can also account only for known, measurable confounding factors. If the groups differ in fundamental ways that cannot be controlled for, a selection bias may be present that limits the validity of one's study.

9.5 Missing Data

Available relevant data may be limited in retrospective studies as the data were recorded or collected for clinical or other purposes outside the scope of the current study. Take, for example, a study investigating patients undergoing re-resection for incidental gallbladder cancer, which occurs when the cancer is diagnosed on pathology after a routine laparoscopic cholecystectomy. The goal of the project is to predict residual disease on re-resection using variables discovered at the earlier surgery. However, the pathology reporting and tissue collection are different for what is thought to be a standard cholecystectomy than for a known gallbladder cancer resection. For instance, the surgeons will perform a portal lymphadenectomy if gallbladder cancer is a known diagnosis. Thus, lymph node status is one factor that may be known only for those patients with cancer that is diagnosed prior to surgery. For patients in whom the lymph nodes were not removed at the incidental procedure, one cannot assume that they were negative for cancer. Additionally, patients referred for a gallbladder cancer surgery may be coming from multiple outside institutions to a tertiary cancer center or a specialist in a different hospital. Because pathology reports are not standardized across institutions or even within institutions, specific information regarding lymphovascular invasion (LVI) or perineural invasion may be missing as well.

Analyzing only those patients who have all their information is known as **complete case analysis**. Complete case analysis is a common strategy for handling missing data [35], but should only be used when the data are **missing at random (MAR)**. The term MAR refers to the situation where missing data are unrelated to the outcome. When summarizing variables, researchers should check for the proportion of cases with missing values. Unfortunately, there is no standard cut-off for the number or proportion of patients for which one should formally check how missing data affects the relationship between risk factors and outcomes. Regardless of amount, efforts should always be made to capture all missing data, which may require re-reviews of the medical records by an additional independent researcher.

Using the above example, if a small number of patients, such as one or two patients, are missing tumor stage or grade, one cannot logistically perform any formal checks, as this is too small a sample from which to make statistical inferences. Therefore, in this example, researchers should assess whether there were particular reasons for the lack of reporting. If one can reasonably assume that the missingness is a function of the retrospective nature of the study and not the result of any factors related to the study itself, then investigators can exclude these patients.

However, if patients with complete information differ from those with missing information with respect to the outcome, we cannot simply perform a complete case analysis. In the above example, suppose LVI data are missing in 25 cases, or 10% of the total study sample. In this situation, one should check whether patients with complete LVI information differ from those with incomplete LVI information with respect to the outcome, residual disease. Next, one should check whether the patients with unknown LVI status differ from those who are positive for LVI or those who are negative for LVI with respect to residual disease. If patients with incomplete LVI data differ from those with complete data, or if patients with incomplete LVI data differ from those with positive or negative LVI, then the data are not missing at random, as an underlying difference exists in those unknown cases [35, 36]. If the data are not missing at random for a particular factor, we cannot include that factor in the analysis, as our sample is not representative.

Alternatively, if the data were to be missing at random with no discernible clinical reason or observed differences with respect to the outcome, single and multiple imputation are two strategies for probabilistically assigning values to patients with missing data. Single-value imputation provides a single value, such as the mean estimate in patients with complete data, for all patients with missing data. In multiple imputations, missing values are determined based on the distribution of other known values in the data set or known values for that patient. Both of these strategies require assumptions and complex probabilistic methods, so researchers should proceed with caution when employing them [36].

Ultimately, when missing data is related to the outcome in a retrospective study, the safest strategy is to not include the factor with missing data in the model or assessment of outcome, and only include those factors where complete data is

available. Although this limits the applicability of one's study to specific factors, it prevents biased estimates or erroneous conclusions. This will strengthen the generalizability to other samples and the overall validity of the study.

9.6 Considerations for Particular Oncology Outcomes

9.6.1 Peri-operative and Post-operative Outcomes

Reporting peri-operative or post-operative outcomes can be a retrospective study in itself, or it can be part of a larger study on outcomes. Peri-operative outcomes may be used in a variety of ways: for learning-curve studies to assess improvements in a new surgical technique, such as laparoscopic cholecystectomy; to assess how one surgical technique compares with another; or to see how peri-operative and post-operative diagnoses later influence survival. Peri-operative outcomes should be clearly defined prior to data collection. For instance, if an operation contains multiple procedures, the researcher needs to decide whether to consider the full operation time or only the time spent on the particular procedure. Analyzing complications has also become important to enable the generation of quality improvement programs and because, in many diseases, complications are associated with oncologic outcomes. A reasonable time period should be defined for which post-operative complications can be attributed to the surgery under study. Overall, when examining these short-term outcomes, clear definitions and methodology are essential for data accuracy and reproducibility.

9.6.2 Survival Outcomes

Survival endpoints are a critical component of many retrospective research studies. Simply estimating overall survival and other survival endpoints for specific cancers is fundamental for understanding their disease course. From these endpoints, we can establish a baseline from which to compare treatment outcomes or identify prognostic biomarkers. When the study objectives are to compare survival between two groups, it is important to report the survival data of the full cohort, as this is one way to check for sampling bias. That is, if the survival estimate of the cohort differs from previously published or clinically understood estimates, the sample may not be representative of the target population. Alternatively, there may be a problem with the way data were collected or the way time was measured.

Essential to correctly estimating survival is knowing when to start counting towards survival. This time point will depend on the patient groups one is comparing and what one is trying to estimate. Suppose we are investigating survival in patients who had laparoscopic liver resection compared with survival in those who underwent open liver resection. At first, it may seem acceptable to measure from time of diagnosis. However, not all patients underwent resection

directly after their diagnosis. In fact, some patients received neoadjuvant chemotherapy, so months may have passed before these patients received surgical treatment. If one were to count the months between diagnosis and surgery as attributable to the effects of surgery, this would bias the findings in favor of patients who waited longer between diagnosis and surgery [37]. Ultimately, one should start the survival clock when the comparison of interest occurred. This allows one to attribute the time between the comparison and event outcome to the comparison of interest.

What counts as an event in a survival study is another factor to consider. As mentioned, retrospective studies suffer from information bias, so the cause of death is not always known. Although in some cancers one may be able to find the cause of death by the course of disease, this is not always the case. Also, patients may receive their primary treatment, such as surgery, at a tertiary cancer center, but then receive adjuvant chemotherapy or further treatment at a local institution, or vice versa. The investigator's current institution may possess only the death certificate or notification of death, but no notes of treatment after the initial diagnosis. This omission makes attributing survival to the cancer of interest difficult. Therefore, unless cause of death can be determined for the majority of patients who died, disease-specific survival as an endpoint should be used with caution.

When investigating disease progression or recurrence outcomes, it is important to consider how to regard death. In many studies, death will be regarded as an event. However, death, particularly in less functional or highly comorbid populations, may be due to causes other than progression of the cancer. Thus, one may want to regard death as a competing event and perform a competing risks analysis. In the first case, one assumes that a death is equivalent to a progression, or that progression had occurred at the time of death. In the latter case, one assumes that the patient's disease had not progressed and that the death prevented the progression from occurring. Assumptions are made in both cases, and which option to use depends on the disease and the study goals.

Lastly, in all studies of survival outcomes, one must consider how to count the patients lost to follow-up. In survival analyses, patients are counted in the survival models up until the point they are censored. In prospective studies, this is usually at the study close or on the off-study date. However, in retrospective studies, cutoff dating may not be so straightforward. In the United States there is no way to freely check death records for individuals, and the families of patients are not legally required to tell treating hospitals of a patient's death. Therefore, one cannot assume that all patients were alive on the last day that survival data were collected. Making this assumption would artificially prolong survival estimates. Alternatively, just because a patient was treated at outside institutions after the initial treatment does not mean that he or she was lost to follow-up on the date of the initial treatment. Assuming so would artificially truncate survival. Instead, one should use the last date a patient was known to be alive, using either clinic visit records, outside reports sent in, or phone conversations recorded with the hospital staff.

9.6.3 Treatment Response

In retrospective studies, the schedule of treatment administration and subsequent follow-up is not standardized. As a result, some patients may have received additional cycles of treatment, fewer cycles of treatment, or missed treatments in a heterogeneous fashion. Similarly, some patients may have had scans done every 6 weeks, some at 8 weeks, and some at 12 weeks. If one is looking at the time-to-treatment response, if patients' responses were not measured at the same time, then the time to response will be artificially altered due to the underlying differences in when measurement occurred. Further, treatment scheduling or drug dosing may have changed over time. To counteract this effect, one can use restrictive sampling to include only those patients with relatively homogeneous treatment schedules and response measurement samples. However, as discussed earlier, restrictive sampling limits the study's generalizability to all patients and to the real-world setting. Thus, time-to-treatment response is a difficult endpoint for a retrospective analysis. Alternatively, one could use response rate by a specific cutoff point, such as 12 weeks, and include all 6-, 8-, and 12-week assessments. Ultimately, treatment response studies must strike a delicate balance between real-world treatment experience and validity.

In the majority of prospective studies, the RECIST 1.1 (Response Evaluation Criteria in Solid Tumors) system is used to measure tumor response, which makes the findings reproducible and internally valid. In contrast, in most retrospective studies, tumor response is determined by the actual radiology report. Reporting may not be standardized and may differ across time or among different radiologists. Therefore, it may be challenging to determine what constitutes a response in a particular patient. One option for correcting this inconsistency is to have a radiologist perform a research re-read using standardized methodology. However, this option may be costly or not feasible in some institutions. When no re-read is conducted, one should record the language on the reports that constitutes a response, stable disease, and progression. These language categories should be reported in the methodology of the manuscript so that data collection is reproducible. In either scenario, deciding on a definition of treatment response before analysis begins is critical.

9.6.4 Residual Disease

In retrospective studies, residual disease status is typically obtained from the surgeon's operative report. Therefore, accuracy of this outcome is largely dependent on the consistency of definitions between surgeons. As an example, take primary debulking surgery for ovarian cancer. One surgeon may say "no residual disease present," another may say "no residual disease present greater than 5 mm," and another may say "no residual disease present greater than 1 cm." Fortunately, in primary

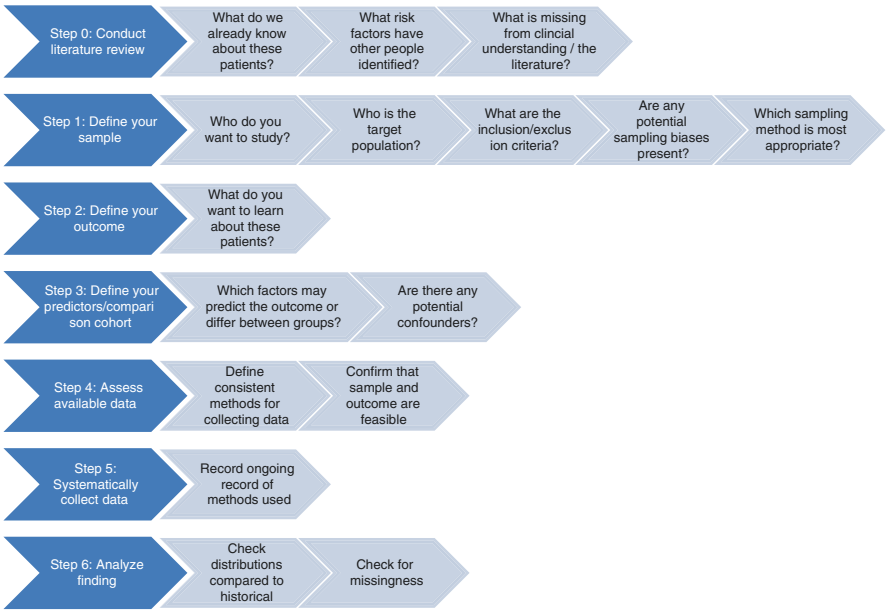

Fig. 9.4 Study design guide

debulking surgery, 1 cm is a generally agreed-upon cutoff, so one may assume that all these patients are free of residual disease. However, not all diseases have an agreed-upon cutoff. In other studies, one may define residual disease by site, such as not present, loco-regional, or distant residual. Therefore, to avoid biasing the findings, the best strategy for residual disease projects is to define residual disease and residual disease location sites before collecting and analyzing data. Additionally, one should consider collecting three elements for residual disease: presence/absence, size, and location. From this, one can use the data gathered to quantify the breadth of responses, while also allowing for appropriate categorization should there be disagreement in the literature. As in the above study outcomes, clear definitions and systematic data collection are the key tools for making a retrospective study internally valid and reproducible.

Conclusions

Retrospective studies allow researchers to study outcomes in a real-world setting at reduced costs compared with those for prospective trials. However, retrospective studies suffer from unique biases that researchers must pay careful attention to. It is critical that patients be selected and data captured methodically in order to make the findings internally valid, generalizable, and reproducible. We leave readers with a baseline checklist of questions to consider when designing a retrospective study, to enable them, as researchers, to better design and more easily execute these types of studies (Fig. 9.4).

References

1. Gay J. Clinical study design and methods terminology. 2010. http://people.vetmed.wsu.edu/jmgay/courses/glossclinstudy.htm. Accessed 1 May 2017.
2. Porter GA, Skibber JM. Outcomes research in surgical oncology. Ann Surg Oncol. 2000;7(5):367–75.
3. Funai EF, Rosenbush EJ, Lee MJ, Del Priore G. Distribution of study designs in four major US journals of obstetrics and gynecology. Gynecol Obstet Investig. 2001;51(1):8–11.
4. Scales CD Jr, Norris RD, Peterson BL, Preminger GM, Dahm P. Clinical research and statistical methods in the urology literature. J Urol. 2005;174(4, Part 1):1374–9.
5. Solomon MJ, McLeod RS. Surgery and the randomised controlled trial: past, present and future. Med J Aust. 1998;169(7):380–3.
6. Kærn J, Tropé CG, Abeler VM. A retrospective study of 370 borderline tumors of the ovary treated at the Norwegian Radium Hospital from 1970 to 1982. A review of clinicopathologic features and treatment modalities. Cancer. 1993;71(5):1810–20.
7. Sartwell PE. Retrospective studies: a review for the clinician. Ann Intern Med. 1974;81(3):381–6.
8. Van den Beuken-van Everdingen M, De Rijke J, Kessels A, Schouten H, Van Kleef M, Patijn J. Prevalence of pain in patients with cancer: a systematic review of the past 40 years. Ann Oncol. 2007;18(9):1437–49.
9. Markman M. A unique role for retrospective studies in clinical oncology. Oncology. 2014;86(5-6):350.
10. Hayden GF, Kramer MS, Horwitz RI. The case-control study. A practical review for the clinician. JAMA. 1982;247(3):326–31.
11. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57–61.
12. Sauerland S, Lefering R, Neugebauer E. Retrospective clinical studies in surgery: potentials and pitfalls. J Hand Surg. 2002;27(2):117–21.
13. The Cochrane Collaboration. Glossary. 2017. http://community-archive.cochrane.org/glossary/5#letterv. Accessed 1 May 2017.
14. Rochon PA, Gurwitz JH, Sykora K, Mamdani M, Streiner DL, Garfinkel S, Normand S-LT, Geoffrey M. Reader's guide to critical appraisal of cohort studies: 1. Role and design. BMJ. 2005;330(7496):895.
15. Cook TD, Campbell DT. The design and conduct of quasi-experiments and true experiments in field settings. In: Dunnette MD, editor. Handbook of industrial and organizational psychology, vol. 223. Amsterdam: Elsevier; 1976. p. 336.
16. Steckler A, McLeroy KR. The importance of external validity. Am J Public Health. 2008;98(1):9–10.
17. Ademuyiwa FO, Edge SB, Erwin DO, Orom H, Ambrosone CB, Underwood W. Breast cancer racial disparities: unanswered questions. Cancer Res. 2011;71(3):640–4.
18. Albain KS, Unger JM, Crowley JJ, Coltman CA, Hershman DL. Racial disparities in cancer survival among randomized clinical trials patients of the Southwest Oncology Group. J Natl Cancer Inst. 2009;101(14):984–92.
19. Du XL, Fang S, Vernon SW, El-Serag H, Shih YT, Davila J, Rasmus ML. Racial disparities and socioeconomic status in association with survival in a large population-based cohort of elderly patients with colon cancer. Cancer. 2007;110(3):660–9.
20. York RO. Conducting social work research: an experiential approach. London: Pearson College Division; 1998.
21. Aschengrau A, Seage GR. Essentials of epidemiology in public health. Burlington, MA: Jones & Bartlett Learning, LLC; 2013.
22. Weiss NS. Clinical epidemiology: the study of the outcome of illness. Oxford: Oxford University Press; 1996.
23. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol. 1990;43(1):87–91.

24. Damon A, Bajema CJ. Age at menarche: accuracy of recall after thirty-nine years. Hum Biol. 1974;46:381–4.
25. Lowe JT, Li X, Fasulo SM, Testa EJ, Jawa A. Patients recall worse preoperative pain after shoulder arthroplasty than originally reported: a study of recall accuracy using the American Shoulder and Elbow Surgeons score. J Shoulder Elb Surg. 2017;26(3):506–11.
26. Schatman ME, Campbell A, Loeser JD. Chronic pain management: guidelines for multidisciplinary program development. Boca Raton, FL: CRC Press; 2007.
27. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhyā. 1973;35:417–46.
28. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150–61.
29. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037–49.
30. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
31. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79(387):516–24.
32. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399–424.
33. Amar D, Zhang H, Pedoto A, Desiderio DP, Shi W, Tan KS. Protective lung ventilation and morbidity after pulmonary resection: a propensity score-matched analysis. Anesth Analg. 2017;125(1):190–9.
34. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat. 1993;2(4):405–20.
35. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91(1):4–8.
36. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66.
37. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1(11):710–9.