Cancer epidemiology in practice: Working notes on cancer history-based selection and censoring

Leon A.M. Berge*[1,2], Tom K. Grimsrud[2], Ronnie Babigumira[1,2], Nathalie C. Støer[2], Nita K. Shala[1,2], Marit B. Veierød[1] and Jo S. Stenehjem[2]

1) Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway

2) Department of Research, Cancer Registry of Norway, Oslo, Norway

*Correspondence: Leon Alexander Mclaren Berge, University of Oslo, P.O. box 1122 Blindern, 0317 Oslo, Norway
E-mail: l.a.m.berge@medisin.uio.no Telephone: 47 99605034

SUMMARY

Researchers conducting observational studies of exposure-cancer associations must make decisions regarding the selection and censoring of study subjects with respect to their cancer history. Given available information, this may include cancers prior to start of follow-up (prevalent cases), and multiple primary cancers (MPCs) at the same or different time-points during follow-up. The choice of analytical strategy may have implications for statistical power and for potential biases. We discuss herein two approaches, the “all cancer approach” and the “first primary cancer approach”. Each contains a set of criteria, primarily for, but not limited to, cohort studies of exposure-cancer associations. Both approaches exclude subjects with prevalent cancers at start of follow-up to avoid information and selection bias but differ in terms of case definition and censoring criteria. In the “all cancer approach”, subjects are censored on non-eligible cancer subtypes of the cancer under study, though subjects with MPCs are not censored on other incident cancers preceding or occurring simultaneously (at time of diagnosis) with the cancer under study. To limit the influence of prior treatment regimens, the “first primary cancer approach” censors subjects upon any first primary incident cancer not under study, including any “simultaneous” MPCs, as well as non-eligible subtypes of the cancer under study. Depending on the cancer under study, these approaches may yield different estimates of the exposure-cancer association and may in this regard be utilized for sensitivity analyses. These alternative approaches are presented and discussed here as a potential resource and source of further discussion for peers in the field of cancer epidemiology.

CANCER HISTORY-BASED SELECTION AND CENSORING

As opposed to the experimental design of animal studies, in which virtually full control is exercised over exposures and potential confounding factors, human studies of disease aetiology are usually observational. In general, observational studies of exposure-cancer associations must contend with dynamic real-life events and may always suffer from some degree of systematic errors (biases), whether these spring from aspects of design, data collection or analyses. The interpretation of the effect estimates will partly depend on the criteria for how study subjects were selected, defined, and censored. Censoring means that follow-up of a subject stops at a certain timepoint beyond which an occurrence of the event of interest is unknown or disregarded (1).

Case definition in an observational cohort study of exposure-cancer associations does not rely only on diagnostic criteria. In general, the subjects are followed until the event of interest (i.e., the cancer, or cancer subtype, under study) occurs or until censoring. The cancer under study is usually a primary cancer (i.e., a cancer originating independently from an organ or specific systemic cell lines), as opposed to a secondary cancer (which usually denotes a metastasis from a cancer at another site) (Table 1). It may be the first primary cancer, or the second or third (etc.) primary cancer, as an individual may have two or more (i.e. multiple) cancers diagnosed during his or her lifetime (2). In subjects with multiple primary cancers (MPCs), two or more tumours arise in different or the same sites (3). Some subjects may even have what appear to be “simultaneous” MPCs (SMPCs), when two or more cancers are registered with the same date of diagnosis.

Criteria for how to select, define, and censor study subjects are an early and fundamental aspect of the analytical work process. We describe two approaches, the “all cancer approach” and the “first primary cancer approach”. Both approaches exclude subjects with a cancer history before start of follow-up and censor upon diagnoses of any other subtypes of the cancer under study (in terms of cancer subsite, histology/morphology, stage, or tumour differentiation) if the cancer under study is itself a cancer subtype. The “first primary cancer approach” censors on other cancer diagnoses than the one under study during follow-up, while the “all cancer approach” does not censor on other cancers during follow-up.

While the “all cancer approach” aims to maintain a representative study sample closer to real-life, there are several reasons why the “first primary cancer approach” should also be considered, which will be discussed later in this article.

Table 1. List of defined cancer terms.

Cancer term	Definition
First primary cancer	The original or first cancer in a person (2).
Secondary cancer	A cancer that has spread (metastasized) from its site of origin to another site in the body (2).
Second, third etc. primary cancer	A new primary cancer in a person with a previous cancer. This occurs independently of other cancers months or years after the previous cancer was treated. May manifest as multiple primary cancers or "simultaneous" multiple primary cancers (2).
Multiple primary cancers (MPC)	Two or more cancers occur independently of time in the same person (3).
«Simultaneous» multiple primary cancers (SMPC)	Two or more cancers registered with the same date of diagnosis in the same person.

Table 2. Cancer history-based selection and censoring criteria of the "all cancer approach" and "first primary cancer approach".

Criteria	The "all cancer approach"	The "first primary cancer approach"
Cancer history before follow-up	Exclude all subjects with any prevalent cancer diagnoses	Exclude all subjects with any prevalent cancer diagnoses
Cancer history during follow-up	Do not censor on diagnoses of other cancers than the cancer under study	Censor on diagnosis of first primary incident cancer, if it is not the cancer under study
Cancer subtypes	Censor on related cancer subtypes	Censor on related cancer subtypes
SMPCs during follow-up	Do not censor subjects on any SMPCs	Censor subjects on any SMPCs

SMPCs = "Simultaneous" multiple primary cancers.

DEFINING THE STUDY SAMPLE

When studying the association between an exposure and cancer, the target (source) population is the group of subjects about which the inferences are desired (e.g., a certain age group of the population, an occupational group, a patient group, etc.) (4). From this, a set of inclusion and exclusion criteria are defined to establish a cohort and the study sample. Selection and censoring criteria may affect the external validity (i.e., the generalizability) of the study sample, which is the potential to apply its study results to the target population and beyond (5). In some cases, this depends on the representativeness of the study sample, the degree to which it reflects the composition of the target population or beyond, in terms of various biological, anthropometric or lifestyle parameters (6).

The "all cancer approach" aims to keep the study sample as representative of a real-life situation as possible during follow-up, by restricting the censoring to death, emigration, or end of follow-up. The "first primary cancer approach" aims to prevent potential biases by censoring on any first primary incident cancer not under study and any SMPCs, but at the expense of potentially reducing the representativeness of the study sample in terms of case definitions and time at risk. The choice of approach depends in part on the research question. For research questions focused on the association or potentially causal relationship between a certain exposure and cancer, the generalizability of potential findings relies more on an understanding of the conditions and mechansms specific to the cancer under study, as one could assume that the mechanisms of carcinogenesis are broadly applicable to other humans independent of selection of study subjects (7). This, rather than representativeness, can tell us whether the findings of a study can be applied beyond the study sample, and thus the "first primary cancer approach" may be more applicable. If, however, the objective is to describe a larger group and thus apply the results of the study sample to a specific target population (e.g., an occupational group), the "all cancer approach" may be preferable as representativeness becomes more important.

The procedures used to select subjects into the study or the analysis may also lead to selection bias when estimating the exposure-cancer association (8). Studies based on self-reported data will only include subjects alive and resident. Additionally, those who provide exposure information may be more motivated to provide answers, be generally healthier, or have higher social status than non-respondents (9). This, in turn, may bias the estimated association between the exposure and the cancer under study. This is less of a problem in register-based studies, in which all subjects are included from an independent source, though often with less targeted information. Methods to further reduce the effect of selection bias may include adjusting for selection covariates, inverse probability weighting, bias sensitivity analyses, or the use of frailty models (9,10).

SELECTION

In both the "all cancer approach" and "first primary cancer approach", all subjects who received a cancer diagnosis before start of follow-up (prevalent cancer) are excluded from the study (Table 2). Exclusion of subjects with prevalent cancer is a common criterion employed in many observational cohort studies of exposure-cancer associations (11,12). A previous cancer diagnosis is a good example of an ordeal which could lead to bias. It may be preferable to exclude subjects with prevalent cancer due to the potential of attribution or recall bias (i.e., exposure is misclassified differentially for those with and without disease) (2). In addition, increased medical surveillance, treatment interference from associated regimens of radiation therapy, chemotherapy or immunotherapy, as well as surgical resection, could, depending on the cancer form under study, affect the likelihood of a subsequent cancer diagnosis during follow-up (11). Such treatments can induce the occurrence of second primary cancers, like lung cancer induced by mediastinal radiation therapy in lymphoma patients (13,14). A previous cancer history (especially at a young age) may also be indicative of an increased genetic susceptibility to cancer, acquired or heritable.

However, the selection of subjects without prevalent cancer should be a less important criterion in cohorts young at start of follow-up and with lower cancer incidence. This could include a situation where exposure information is based on complete and lifelong employment records, where subjects are followed up from the first day of any employment. Under such circumstances, later surveys recording more detailed work history data, might be less prone to recall bias and could justify the inclusion of subjects with prevalent cancer diagnoses other than the cancer under study. The exclusion of subjects with prevalent cancer could also mean that the study does not include certain subjects with relevant exposure. If a hypothetical study of the association between smoking and a certain cancer excluded subjects with prevalent cancer, including lung cancer, then several subjects with likely exposure to smoking may be excluded. It is also worth noting that many other serious health conditions prior to start of follow-up may introduce similar types of bias in studies of exposure-cancer associations. However, such studies rely on information from cancer registries, which typically do not include information on other diseases. On the other hand, when the representativeness of the study sample is important, it may also be more prudent for the "all cancer approach" to include all subjects with prevalent cancer, other than the cancer under study.

CENSORING

As mentioned briefly above, the subjects are followed until the event of interest (i.e., the cancer under study) occurs or until censoring. By principle, censoring should be non-informative, which means the cause behind censoring should be unrelated to any aspect of the study, in other words, censored subjects should have the same chance of survival (i.e., not experiencing the event of interest) as those who continue to be followed. For example, in a study of the effect of hormone replacement therapy on conception among women, the effect of women who drop out (and are censored) due to a failure to conceive would bias the results (15). Without a diagnosis of the cancer under study, subjects are censored at the end of the study period or upon leaving the study (e.g., emigration). However, other events during follow-up may affect the chance of subsequently receiving the relevant cancer diagnosis, and while death is a definitive example of this, subjects may also be censored at other time-points. Hence, censoring is practiced in time-to-event analyses to avoid overestimating time-at-risk (1). It is important to handle censoring carefully as it can potentially bias the results and reduce statistical power.

In a study of exposure-cancer associations, the "all cancer approach" would censor a subject on the date of death, emigration, or end of the follow-up period, whichever occurred first. However, if the cancer under study is a subtype of a cancer (e.g., in terms of cancer subsite, histology/morphology, stage or tumour differentiation), then subjects are censored upon diagnoses of the subtypes not under study (Table 2). A diagnosis of another subtype of the cancer under study may alter a subject's subsequent risk of the cancer under study due to circumstances irrelevant for evaluation of exposure-related risks, including increased surveillance and more advanced examinations, and associated therapy regimens. Hence subjects are censored to prevent increasing follow-up time beyond this point. An example is non-small cell lung cancer as a subtype of total ("any") lung cancer. If the cancer under study is non-small cell lung cancer, subjects are censored upon being diagnosed with any other lung cancer subtype, because it may affect the likelihood of having a subsequent non-small cell lung cancer diagnosis.

In the "first primary cancer approach", we additionally censor on the diagnosis of the first primary incident cancer unless it is the cancer under study (Table 2). This approach is in line with the criterion for excluding subjects with cancer history prior to follow-up and is meant to avoid undue influence from past tumours and from increased medical surveillance and associated therapy regimens during follow-up. The exposure under study may also be a weak carcinogen, which means the "first primary cancer approach" may be more likely to identify a potential exposure-cancer association by limiting the effect of prior cancer treatments. With the "first primary cancer approach", follow-up times may be shorter, and the age at end of follow-up lower for both cases and non-cases, compared to the "all cancer approach". It may also change the number of cases and non-cases, as a subject diagnosed with another cancer prior to the cancer under study would be ineligible as a case. In the "first primary cancer approach" we thus accept a potentially reduced statistical power and a reduction in representativeness in terms of both external and internal comparisons (between study participants) as we discriminate between subjects based on cancer history and censor individuals who technically could be followed longer for a potential diagnosis of the cancer under study.

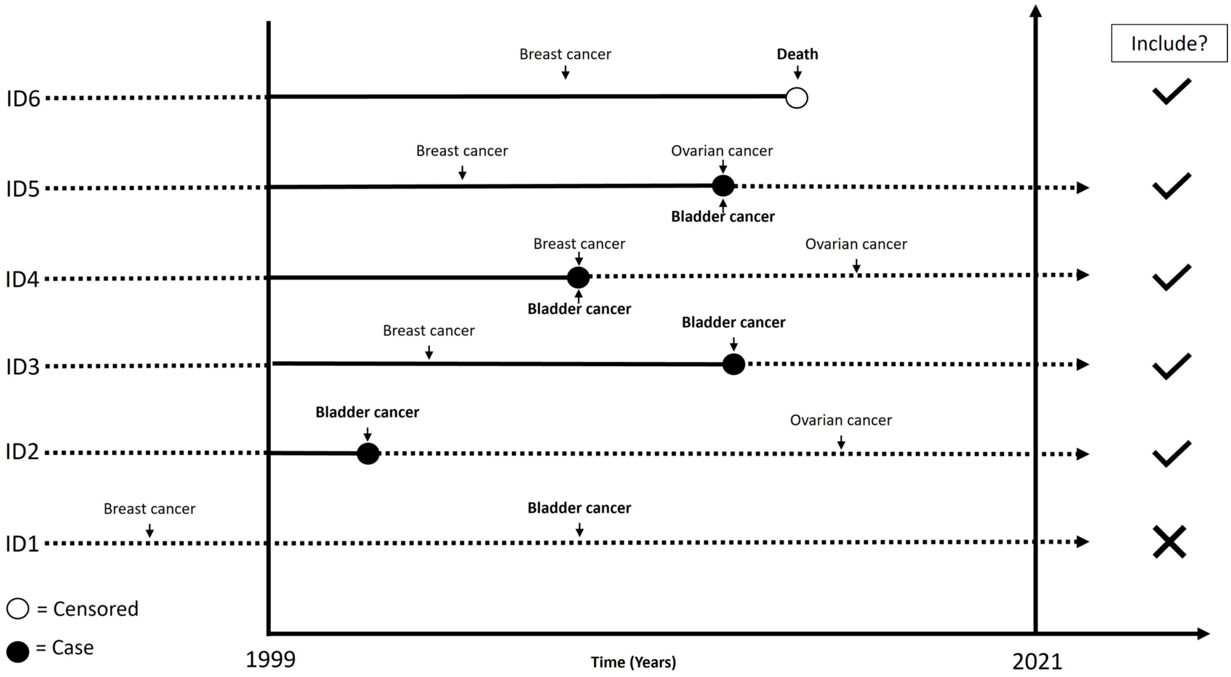

Figure 1. Temporal sequence from start (1999) to end of follow-up (2021) in a study of bladder cancer incidence among women, illustrating the cancer history of six bladder cancer cases that could potentially be included in the study and the application of the selection and censoring criteria of the "all cancer approach". The left vertical line is the start of follow up. The solid horizontal lines are the follow-up times for each subject. A filled circle indicates a case, and an open circle indicates censoring. The right vertical line marks the end of the study.

"SIMULTANEOUS" MULTIPLE PRIMARY CANCERS (SMPCS)

The definitions of MPCs change over time and may vary between studies, though two rulesets are commonly used; those of the Surveillance Epidemiology and End Results (SEER) program (16), primarily used by North American cancer registries, and those of the International Association of Cancer Registries and International Agency for Research on Cancer (IACR/IARC) (17), used more internationally, including by European cancer registries (3,18). Some subjects could, according to the data provided, be diagnosed with two or more SMPCs (i.e., primary cancers registered on the same date). This may be a result of data registration artifacts, or a thorough examination (workup) in cases with suspected malignant disease to clarify the origin site of the cancer, as well as any metastases and comorbidity.

In such cases, the "true" first primary cancer is often not known due to the other primary or "tied" cancer diagnosis. In the "all cancer approach", follow-up of subjects is stopped at occurrence of SMPCs which include the cancer under study, in the same manner as with a first primary or second primary incident diagnosis of the cancer under study and are hence defined as cases (Table 2). Maximizing the number of cancer cases in analyses, including those with MPCs, is seen as conducive to producing estimates with potentially less selection bias, at least in studies of cancer survival (19). Longer cancer survival times, coupled with increasingly aging populations also means that the number of subjects with MPCs (and SMPCs) is expected to increase, also making their inclusion conducive to the external validity of an analysis, in terms of both risk and survival (20).

When applying the "first primary cancer approach", any subjects with SMPCs, including those with the cancer under study, are censored upon the SMPCs and are not defined as cases (Table 2). This is due to a lack of knowledge of what tumour constitutes the primary diagnosis. We may also be unable to decide whether genetic susceptibility for the cancer under study is more pronounced in subjects with SMPCs than those with cancers occurring further apart in time (3). It has been found that subjects with MPCs are more likely to be Caucasian, have less aggressive tumours that present at earlier stages, have longer survival times, and have a strong family history of cancer (21). Depending on the cancer type under study, however, this may potentially affect only a small number of subjects, and could thus be a minor issue.

EXAMPLE

The following is an example of how selection, censoring, and SMPCs would be handled in a hypothetical study investigating the risk of bladder cancer incidence among women, applying both the "all cancer approach" and "first primary cancer approach".

The "all cancer approach"

We exclude ID1 as this subject has a prevalent cancer diagnosis prior to start of follow-up (bladder or any other cancer type). ID2 is included in the study as a case

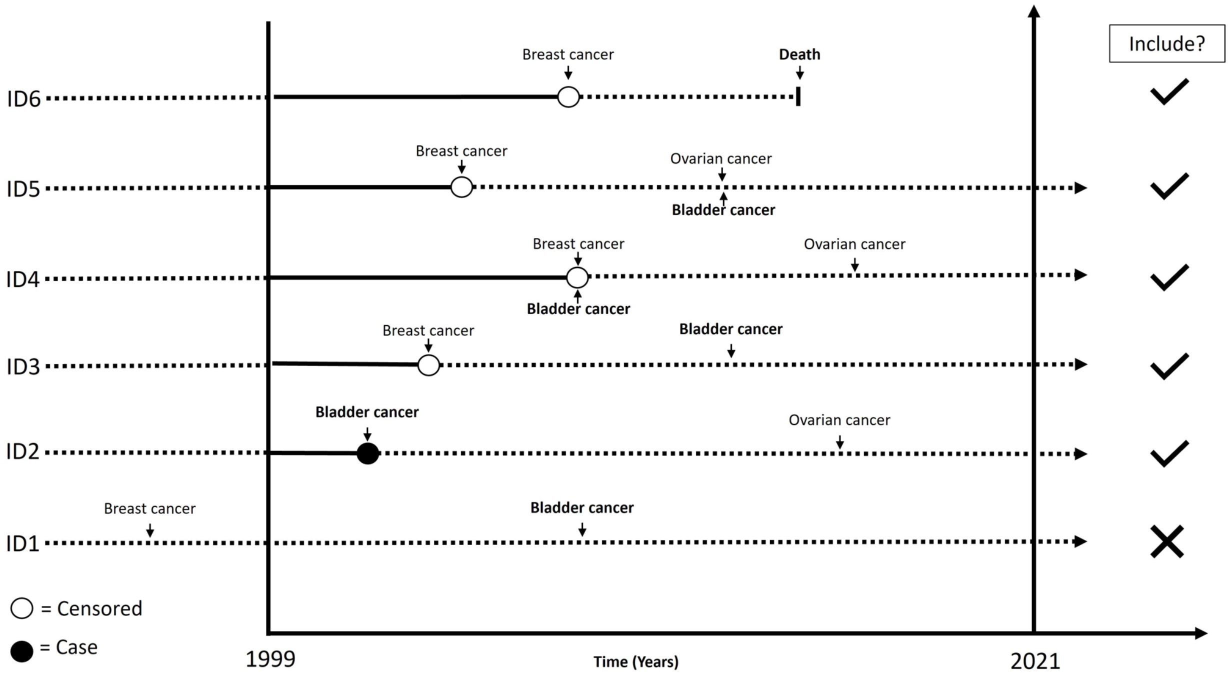

Figure 2. Temporal sequence from start (1999) to end of follow-up (2021) in a study of bladder cancer incidence among women, illustrating the cancer history of six bladder cancer cases that could potentially be included in the study and the application of the selection and censoring criteria of the "first primary cancer approach". The left vertical line is the start of follow up. The solid horizontal lines are the follow-up times for each subject. A filled circle indicates a case, and an open circle indicates censoring. The right vertical line marks the end of the study.

and follow-up time stops at the bladder cancer diagnosis. ID3 is also included in the study, and follow-up time stops at the bladder cancer diagnosis, despite the first primary incident breast cancer diagnosis. ID4 is included in the study as a case and follow-up time stops at the bladder and breast cancer SMPC. ID5 is included in the study as a case and follow-up time stops at the bladder and ovarian cancer SMPC, despite the first primary incident breast cancer diagnosis. Finally, ID6 is included in the study, and is censored upon death, despite the first primary incident breast cancer diagnosis (Figure 1).

The "first primary cancer approach"

We again exclude ID1 due to a prevalent cancer diagnosis prior to start of follow-up. ID2 is included in the study as a case and follow-up time stops at the diagnosis of bladder cancer. ID3 is also included in the study but is censored at the diagnosis of the first primary incident breast cancer diagnosis (stopping follow-up time), which means it is not considered a case for the purposes of the analyses. ID4 is included in the study and is censored upon occurrence of the bladder and breast cancer SMPC and is not considered a case. ID5 is also included in the study but censored at the first primary incident breast cancer diagnosis, and thus not considered a case. As before, ID6 is included in the study, but is now censored on the first primary incident breast cancer diagnosis (Figure 2).

CONCLUSIONS

We have presented herein two approaches for the selection, definition, and censoring of study participants in observational cohort studies of exposure-cancer associations. While efforts have been made to discuss and justify the different considerations of each approach, they are not meant to be interpreted as a strict set of methodological instructions and may benefit from additional sensitivity analyses. Rather, they are presented here for discussion and consideration by peers. A synthesis of such methodological work processes rarely makes its way into mainstream publications, but when shared, might serve as an additional and useful resource within the field of cancer epidemiology.

REFERENCES

1. Porta M. A Dictionary of Epidemiology (6th edn.) – Censoring. Oxford University Press, 2016.
2. NCI Dictionary of Cancer Terms [Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms.]
3. Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, et al. Multiple primary tumours: challenges and approaches, a review. *ESMO Open* 2017; **2** (2): e000172-e.
4. Porta M. A Dictionary of Epidemiology (6th edn.) – Target Population. Oxford University Press, 2016.

5. Westreich D. Causal Impact: From Exposures to Interventions. Epidemiology by Design. New York: Oxford University Press, 2019.
6. Westreich D. Observational Cohort Studies. Epidemiology by Design. New York: Oxford University Press, 2019.
7. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013; **42** (4): 1012-4.
8. Porta M. A Dictionary of Epidemiology (6th edn.) – Selection Bias. Oxford University Press, 2016.
9. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta Obstet Gynecol Scand* 2018; **97** (4): 407-16.
10. Stensrud MJ, Valberg M, Røysland K, Aalen OO. Exploring selection bias by causal frailty models: the magnitude matters. *Epidemiology* 2017; **28** (3): 379-86.
11. Cronin-Fenton DP, Antonsen S, Cetin K, Acquavella J, Daniels A, Lash TL. Methods and rationale used in a matched cohort study of the incidence of new primary cancers following prostate cancer. *Clin Epidemiol* 2013; **5**: 429-37.
12. Wang YH, Li JQ, Shi JF, Que JY, Liu JJ, Lappin JM, et al. Depression and anxiety in relation to cancer incidence and mortality: a systematic review and meta-analysis of cohort studies. *Mol Psychiatry* 2020; **25** (7): 1487-99.
13. Travis LB, Gospodarowicz M, Curtis RE, Clarke EA, Andersson M, Glimelius B, et al. Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease. *J Natl Cancer Inst* 2002; **94** (3): 182-92.
14. Dores GM, Metayer C, Curtis RE, Lynch CF, Clarke EA, Glimelius B, et al. Second malignant neoplasms among long-term survivors of Hodgkin's disease: a population-based evaluation over 25 years. *J Clin Oncol* 2002; **20** (16): 3484-94.
15. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ* 1998; **317** (7172): 1572.
16. Adamo M, Johnson C, Ruhl J, Dickie L. 2014 SEER program coding and staging manual. National Cancer Institute, Bethesda, MD 20850-9765.
17. International rules for multiple primary cancers (ICD-0 third edition). *Eur J Cancer Prev* 2005; **14** (4): 307-8.
18. Coyte A, Morrison DS, McLoone P. Second primary cancer risk – the impact of applying different definitions of multiple primaries: results from a retrospective population-based cancer registry study. *BMC Cancer* 2014; **14**: 272.
19. Weir HK, Johnson CJ, Thompson TD. The effect of multiple primary rules on population-based cancer survival. *Cancer Causes Control* 2013; **24** (6): 1231-42.
20. Mariotto AB, Rowland JH, Ries LAG, Scoppa S, Feuer EJ. Multiple cancer prevalence: a growing challenge in long-term survivorship. *Cancer Epidemiol Biomarkers Prev* 2007; **16** (3): 566-71.
21. Amer MH. Multiple neoplasms, single primaries, and patient survival. *Cancer Manag Res* 2014; **6**: 119-34.