Dissecting the causal question underlying the association between cancer and ADRD

L. Paloma Rojas-Saunero

1. Introduction

There is significant interest in understanding the association between cancer and dementia. Many observational studies have consistently determined that individuals with a history or incident cancer have a lower risk of developing of Alzheimer´s disease or related dementias (ADRD) when compared to individuals with no history of cancer. These findings have motivated the identification of molecular and genetic mechanisms that explain this association. Specific molecular mechanisms, such as Pin1 enzyme over-expression promotes tumorigenesis, while its down-regulation is attributed to mechanisms that contribute to Alzheimer's Disease. These findings suggest new opportunities to repurpose cancer chemotherapeutics for Alzheimer´s Disease.

Nevertheless, observational studies are susceptible to several bias, and researchers have raised concerns related to the competing event of death as well as unmeasured confounding and ascertainment error that could drive these results. Though, to understand sources of bias the causal question should be explicitly specified. This particular question has at least two core challenges. First, defining the underlying biological mechanism we would ideally intervene upon and defining cancer diagnosis as the proxy for that mechanism. For example, researchers refer to biological mechanisms which onset may be prior to cancer diagnosis and to the effect of cancer treatment, both relate to different causal questions and would be translated into different study design and analytic choices. Second, defining the causal question (or estimand) of interest when the competing event of death is interest. Previous studies alluded to choosing a particular summary measure or modeling approach to account for death (incidence rate ratios (ref), hazard ratios (ref), multistate models (ref)). However, when competing events are present, there is more than one causal question that can be framed and each question has different interpretations and relies on different assumptions.

With these considerations in mind, we outline a case study focused on the question: *What is the direct effect of Pin1 over-expression at late-midlife in the risk of ADRD after 20 years of follow-up compared to having regulated expression of Pin1.* To identify this effect with real-world data, we illustrate the complexity of the research question by building progressively a directed acyclic graph (DAG) which help us connect the causal question to the observable data and the assumptions we rely on to identify the effect. We exemplify different scenarios with data collected from the Rotterdam Study, a population-based cohort study. We describe the challenges and how they translate into the analytic decisions. Last, we discuss how information on mortality and cause of death can provides insight about the direction of some sources of bias.

2. Overview of the causal structure and assumptions related to cancer diagnosis as a proxy for Pin1

We begin by illustrating a hypothetical scenario through the causal graph of Figure 1, that depicts a situation where Pin1, represented as P, was measured in all participants in late midlife (for example between age 50 and 60). Based on the measurement of Pin1, participant's exposure is classified as 1) Over-expressed Pin1 or 2) Regulated Pin1. Although Pin1 expression acts as a biomarker and could be considered an ill-defined intervention, we can allude that sometime in the future people with over-expressed Pin1 could follow a specific treatment that lowers Pin1 under a specified threshold. Furthermore, through out the follow-up

each participant followed a strict monitoring for ADRD onset, represented as Y_{20}, and no participant was lost to follow-up nor died during the study period (no competing events present).

Since Pin1 expression was not a randomized intervention, we consider a set of measured covariates L for which conditional exchangeability holds. Previous studies have described age, sex, educational level and race/ethnicity as the minimal adjusting set of covariates (Ospina). However, environmental and behavioral factors such as smoking, which are known to cause microenvironmental changes such as inflammation and changes in tissue remodelling, may translate into Pin1 over-expression and are also related to the development of ADRD. Therefore, had we measured all L we could identify the effect of P in Y_{20} by conditioning on L.

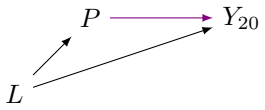


Figure 1: Effect of Pin1 in ADRD

Unfortunately, Pin1 expression measurement is not available at the moment in real-world settings, we can only rely on a proxy of this exposure. Since Pin1 over-expression is present in tumors and tumors are only measured through screening and diagnosis, we considered cancer diagnosis as the proxy for Pin1 over-expression, as most observational studies have defined (all refs). We depict this feature in Figure 2, where P^* represents *incident cancer diagnosis*, the measured proxy of P. In this DAG we colored the path $P^* \leftarrow P \rightarrow Y_{85}$ because although we are measuring the association between P^* and Y_{20} in the observed data, we are assuming that the captured effect is only through P.

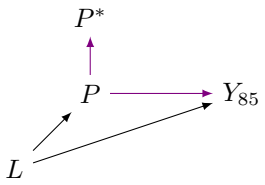


Figure 2: Effect of Pin1 in ADRD, with cancer diagnosis as proxy of Pin1

A major challenge related to assigning cancer diagnosis as the proxy of Pin1 is defining time zero. Not everyone who had Pin1 over-expression will be diagnosed with cancer by late-midlife, but in late-life. For this reason prospective cohort studies have considered cancer diagnosis as a time-varying or time-dependent exposure. This means that a participant within a prospective cohort study contributes to the "regulated Pin1" since study entry up to the time of cancer diagnosis and later on to the "Over-expressed Pin1" arm. Other studies have included participants with cancer diagnosis at the time of the diagnosis (for example from cancer registries such as SEER) and matched participants by age. In both cases we must remember that the the main interest is on the unmeasured P thus we should only adjust for covariates prior to P and be careful to adjust for post-baseline covariates of P or mediators between P and P^*.

One of the mediators between P and P^* is death prior to cancer diagnosis. Since individuals are at risk of dying from other causes (such as cardiovascular death), we can only measure P^* in the subset of individuals who have survived long enough to have a cancer diagnosis. We zoom in the relationship of $P \rightarrow P^*$ in Figure 3. In this DAG we include $D = 0$ between P and P^*, this illustrates that to observe P^*, we condition (box around $D = 0$) on surviving long enough to have a cancer diagnosis. Several risk factors that increase the risk of cancer might also cause death prior to cancer diagnosis, for example smoking may cause lung cancer and chronic obstructive pulmonary disease (a leading cause of death in this age group). Therefore, to isolate the effect of P in P^* (the violet arrow), we need to block all shared causes between P and D and between D and $P*$ both represented in Figure 3 as C_1 and C_2. In this way we assume a hypothetical scenario in which we could prevent death prior to cancer diagnosis, by conditioning on a rich set of covariates. This assumption must hold regardless of whether we use incident cancer as a time-varying exposure, or match cancer patients to participants free of cancer by age.

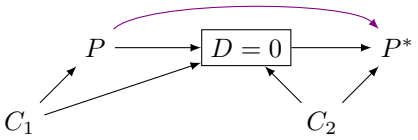


Figure 3: Zoom in the association between P and P star

Up to this point we have outlined the underlying assumptions of using incident cancer diagnosis as a proxy for Pin1 expression. We continue outlining the additional assumptions related to death as a competing event of ADRD.

3. Overview of the causal question and assumptions related to death as a competing event of ADRD

As illustrated in previous figures, we are interested in the effect of Pin1 in late midlife in ADRD risk after 20 years of follow-up. Nevertheless, mortality is substantial and increases steeply in late life, and a leading cause of death is cancer. This means that in order to measure ADRD over follow-up individuals have to had survive to cancer and to other leading causes of death. For this reason death is a competing event of ADRD because if a participant dies prior to ADRD diagnosis, death prevents from observing ADRD at future time-points. We can visualize this data feature in the causal diagram in Figure 4. In this DAG we include D_{15} as a representation of death at 15 years of follow-up, an arrow between P and D_{15} since P may increase the mortality risk irrespective of cancer diagnosis. We also include an arrow between P^* and D_{15} since cancer diagnosis, and subsequent treatment (or lack of treatment) may have an effect on death. Last, the arrow from D_{15} and Y_{20} represents the key feature of a competing events data structure: an individual who dies at 15 years of follow-up cannot subsequently develop ADRD, and since D_{15} and Y_{20} are events related to aging, C represent the shared causes of both events.

In this DAG we observe that, had we measured and adjusted for L, we could estimate the total effect of P in Y_{20} without further assumptions. However the total effect includes all pathways between P and Y_{20}. This means that if Pin1 has an effect on mortality through cancer or other mechanisms, we will observe a protective effect of Pin1 over-expression in ADRD, partially or fully mediated through death. For this reason methods that are used to approximate a total effect, such as the Aalen-Johansen estimator (underlying estimator for multistate models) or subdistribution hazard models, may not be useful in this setting. Since the question of interest is focused on the direct effect of P in Y_{20} as in Figure 4 (violet arrows), there are several causal questions (estimands) to represent this direct effect, such as: the controlled direct effect (CDE), the survivor average causal effect (SACE) and the natural separable direct effect. In this section we discuss the controlled direct effect as the causal question of interest since it translates to frequently used methods in this literature (such as Kaplan-Meier estimator and Cox-proportional hazard model).

THe CDE represents the effect of Pin1 in ADRD in a setting where we could have prevented death over the study period. It relies on the assumption that we have measured all C to block the pathway $Y20 \leftarrow C \rightarrow D_{15} \rightarrow P \rightarrow P*$. This assumption is defined as the independent censoring assumption conditional on covariates. In this setting death is treated as a censoring event and it can be interpreted as those who died would have the same risk of developing dementia if prevented from dying than those who remained alive and free of dementia at a given time point, conditional on shared causes of dementia and death.

4. Application to the Rotterdam Study

In this section we illustrate how previous DAGS may reflect on the analysis performed to unveil the effect of Pin1 in all cause-dementia. We use data collected in the Rotterdam Study, a population-based prospective cohort study among persons living in the Ommoord district in Rotterdam, the Netherlands. Recruitment and initial assessments were held between 1990 and 1993; it was later extended between 2000 and 2001 consisting

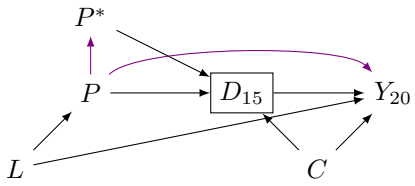


Figure 4: Direct effect of Pin1 in the risk of ADRD at 20 years of follow-up, with cancer diagnosis as proxy of Pin1

of individuals who had reached the age of 55 years or who had moved into the study area. Participants from first subcohort had follow-up visits between 1993-1995, 1997-1999, 2002-2005, and 2008-2010, second subcohort had follow-up visits between 2004 and 2005, and between 2011 and 2012. All participants had data on incident cancer diagnosis and incident dementia diagnosis through follow-up, collected from medical records of general practitioners (including hospital discharge letters) and through linkage with national registries. Date and cause of death was collected on a weekly basis via municipal population registries. Data on clinical outcomes was available until 2015. Further study details can be found in *Supplementary Material x*.

To match the analysis to our initial question, we considered as inclusion criteria being between 60 and 70 years old at study entry, without history of cancer diagnosis, free of cognitive decline or with previous history of dementia. Out of 10998 persons who participated at study entry, 3642 were considered eligible. Time to cancer diagnosis, time to dementia diagnosis and death status was measured for all participants. All participants were followed from study entry until dementia diagnosis, death or 20 years after their individual baseline date, whichever occurred first.

4.1. Methods

We illustrate the association between cancer and dementia diagnosis under different scenarios that resemble the DAGs discussed above. First we considered the most simple scenario, were Pin1 over-expression is defined as *"cancer ever vs. never"* as if Pin1 measurement and cancer diagnosis happened at the same time if looked retrospectively. Second, we considered cancer as a *"time-varying"* exposure independent of death. This can be represented as follows, at five years from study entry the observed unadjusted risk of cancer was 6.5 (95%CI: 5.8, 7.4) and the risk of death was 2.6 (95%CI: 2.2, 3.2) for the entire cohort, this that those who died would have the same risk of cancer had they not died over follow-up. Third, we considered *"time to cancer"* diagnosis as the proxy for Pin1 over-expression. *Not sure how to express this*

To address confounding we fit inverse probability treatment weights, stabilized and truncated at 99th percentile. Weight fitting was different for each scenario. For the scenario "ever vs. never" weights for cancer diagnosis were defined as the inverse of the probability of cancer diagnosis conditional on confounders, and for individuals free of cancer as the inverse of not having cancer conditional on covariates. The scenario "time-varying cancer diagnosis" is similar to the previous scenario, weights are the product of the time-fixed IPT weights above and year-specific. For the scenario "time to cancer", weights represent the product of the time-fix IPT weights for time to cancer diagnosis, weights turn to one at the moment of cancer diagnosis. In all cases we estimated these probabilities assuming a logistic regression model for cancer diagnosis as a function of the following covariates: age at study entry, sex, educational attainment, cohort, smoking status. Further details on modeling specifications and weights assessment are presented as **Supplementary material x**.

To estimate the controlled direct effect in time-varying settings we compared the complement of a weighted Kaplan-Meier survival estimator in participants with incident cancer vs. no incident of cancer with time indexed in years. The weights in this case are time-varying by follow-up year, defined as a product of the time-fixed IPT weights above and a year-specific inverse probability of censoring (IPC) by death weights. For an individual still alive in year t, the time t IPC weight is the product of the inverse probability of

surviving in each year prior to t, conditional on measured common causes of death and dementia (that is, variables such as C in Figure 4). For an individual who has died by time t, the year t IPC weight is zero. We estimated survival probabilities using a logistic regression model for death as a function of baseline and time-varying covariates. Baseline covariates included age at study entry, sex, apoe4 status, educational attainment and the time-varying covariates smoking status, systolic blood pressure, BMI and prevalent and incident comorbidities such as: cancer, heart disease, stroke and diabetes. We additionally calculated the controlled direct effect considering death as an unconditional independent censoring event (as if there were no arrows from {C} to D_{15} and Y_{20}) for an illustrative purpose.

Estimates of the controlled direct effect at 20 years of follow-up are presented as risk differences (RD), risk ratios (RR) and hazard ratios (HR). We note that hazards, unlike risks, inherently condition on surviving both dementia and death, as such they will not have a causal interpretation in this case. We present them for comparison against risk ratios.

Since the conditional independent censoring assumption is untestable, we compute Peterson upper and lower bounds to represent: 1) the extreme scenario of independence, that refers to an scenario were those who died would never develop dementia (lower bound) and 2) complete dependency, that refers to an scenario where those who died would have a dementia prior to death (upper bound). The lower bound is calculated with the Aalen-Johansen estimator treating death as a competing event, and the upper bound is calculated with the Kaplan Meier estimator for the combined outcome of dementia or death.

We additionally present the causes of death for participants who developed cancer and for participants free of cancer diagnosis over follow-up. This information provides insights about the association between cancer diagnosis and death and about the potential misclassification of individuals with no cancer diagnosis (cancer as a cause of death).

All 95% confidence intervals were calculated using percentile-based bootstrapping based on 500 bootstrap samples. All analysis were performed using R, code is provided in supplementary material and available in https://github.com/palolili23/2021_cancer_dementia.

4.2. Results

Participants had a mean age of 64.46 (SD: 2.86), and 54% (n = 1981) were women. Further details on participants are presented in Table 1. Over follow-up, 24% (n = 878) developed cancer and 76% (n = 2764) remain free of cancer diagnosis, the median age of cancer diagnosis was 73 (IQR: 69-77). From the total sample, 12% (n = 431) had dementia over follow-up and median time to dementia was 79 (IQR: 75-83). Among participants with incident cancer, 6% (n = 50) had dementia diagnosis and 63% (n = 549) died over follow-up, 32% (n = 279) remain alive at 20 years since study entry. In contrast, among participants free of cancer diagnosis over follow-up, 14% (n = 385) were diagnosed with dementia and 23% (n = 624) died over follow-up, 63% (n = 1755) were alive at the end of follow-up.

Results for all scenarios are present in Table 2. Had we defined cancer diagnosis as *ever vs. never*, and relying on death as independent censoring event (unconditional), we observe a significant protective effect of ever having cancer in the risk of dementia [RR: 0.7 (95%CI), HR: 0.54 (0.4, 0.74)]. However this effect is diminished [RR: 1.03 (95%CI), HR: 0.92 (0.67, 1.25))] if we relax the assumption of independent censoring conditional on baseline covariates related to death.

Had we defined cancer diagnosis as a *time-varying* exposure we observe higher risk of dementia had participants had a cancer diagnosis over time, had we prevented death conditional on covariates [RR: 1.53 (95%CI), HR: 1.6 (1.05, 2.43))].

5. Discussion (brainstorming points)

- We observed how definition of cancer diagnosis as the proxy of Pin1 over-expression changes results. Most papers focus on a methodology but not on the question behind. Only with this in mind we can consider the confounders of the association. If the question was instead related to different cancer

treatments it would require a different design and definition of confounders. As opposed to Ospina's paper that says this:

"Confounders that would explain the observed inverse cancer-AD association would be those that raise risk of cancer but reduce risk of AD, ruling out many common lifestyle and social factors associated with increased risk of both conditions, such as smoking or alcohol consumption. We considered age, sex, and educational level as sociodemographic factors that should be included in a minimal adjustment set in all studies on this association."

- Previous studies classified "competing risk bias" vs. "survival bias" which is unclear. We first need to pick an estimand. If we are interested in the CDE we rely a strong assumptions. But to consider that unconditional independecy makes no sense. Results change substantially if we relax this assumption with covariates. Also bounds show us the extent of extreme scenarios. Also a large proportion of cancer patients died prior to dementia diagnosis (63% (n = 549)), the leading cause of death was cancer in this group.
- Besides, all estimands can be presented as risks, but depending on the estimand it treats death differently, and under different assumptions, and time-varying hazards (period specific hazards) are not useful.

Efforts to prevent and treat cancer should converge with similar efforts to prevent other aging- associated diseases. We need to figure out what the key aging-dependent changes are and how to modulate these factors safely.

- Knowing the cause of death provides information about the direction of missclassification. Among individuals free of cancer, we observed % of individuals who died with cancer as a primary cause.
- Explicitly outlining the estimands and the assumptions that connect the causal question to the observed data provide an opportunity to improve the design of observational studies and the interpretation of their findings, plus better insight of potential sources of bias.
- This is a crucial since these studies are providing insights that are guiding other fields of research in the area, from bench science to biostatistics and epidemiological methods.
- The CDE has an interpretation that relates to an scenario where death was eliminated. Future work on separable effects may help disentangle the different mechanisms that affect dementia and death.
- In the future, we may be able to measure this biomarker and collect data retrospectively from stored blood samples, but we need to design the study very carefuly.
- We could have change Pin1 to other molecular mechanism. This also extends to other questions that study the effect of one disease in the risk of another disease to understand the biological mechanisms behind the.

Tables

Table 1: Descriptive characteristics of individuals who had a cancer diagnosis and of those free of cancer diagnosis over follow-up.

	Incident cancer	No incident cancer
n	878	2764
sex = Male (%)	520 (59.2)	1141 (41.3)
age_0 (mean (SD))	64.61 (2.87)	64.42 (2.86)
education (%)		
Higher	114 (13.0)	269 (9.7)
Intermediate	412 (46.9)	1225 (44.3)
Lower	347 (39.5)	1251 (45.3)
Unknown	5 (0.6)	19 (0.7)
apoe4 (%)		
Not carrier	622 (73.8)	1874 (71.1)
One allele carrier	203 (24.1)	685 (26.0)
Two allele carrier	18 (2.1)	78 (3.0)
smoke1 (%)		
Current	260 (29.6)	664 (24.0)
Former	425 (48.4)	1301 (47.1)
Never	193 (22.0)	799 (28.9)
bmi1 (mean (SD))	26.39 (3.50)	26.65 (3.80)
oh1 (mean (SD))	13.25 (17.97)	10.13 (15.50)
sbp1 (mean (SD))	138.74 (20.98)	138.67 (20.80)
ht1 = No history of hypertension (%)	369 (42.0)	1141 (41.3)
hd_prev = No history of heart disease (%)	800 (92.6)	2503 (92.5)
hd_v = No incident heart disease (%)	665 (75.7)	1967 (71.2)
diabetes_prev (%)		
History of diabetes	91 (10.4)	284 (10.3)
No history of diabetes	553 (63.0)	1981 (71.7)
Unknown	234 (26.7)	499 (18.1)
diab_v = No incident diabetes (%)	713 (81.2)	2215 (80.1)
stroke_prev = No history of stroke (%)	863 (98.3)	2717 (98.3)
stroke_v = No incident stroke (%)	777 (88.5)	2403 (86.9)
cancer_v = No incident cancer (%)	0 (0.0)	2764 (100.0)

Table 2. Risk difference and risk ratio for the risk of dementia

Proxy	model	cancer_v=0	cancer_v=1	rd	rr	hr
Ever vs. Never	Unadjusted	20.3	14.4	-5.9	0.71	0.55 (0.41, 0.73)
Ever vs. Never	IPTW	20.3	14.3	-6.0	0.70	0.54 (0.4, 0.74)
Ever vs. Never	IPTW + IPCW	21.1	21.7	0.7	1.03	0.92 (0.67, 1.25)
Time-varying cancer	Unadjusted	19.4	18.8	-0.6	0.97	0.98 (0.73, 1.32)
Time-varying cancer	IPTW	18.3	23.1	4.9	1.27	1.37 (0.91, 2.05)
Time-varying cancer	IPTW + IPCW	19.0	29.1	10.1	1.53	1.6 (1.05, 2.43)
Time to cancer	IPTW	19.5	18.5	-1.0	0.95	0.96 (0.7, 1.31)
Time to cancer	IPTW + IPCW	20.3	21.9	1.7	1.08	1.07 (0.77, 1.49)

Table 3. Peterson bounds on the risk of dementia

Proxy	model	cancer_v=0	cancer_v=1	rd	rr
Time-varying cancer	Lower bound	15.0	9.4	-5.6	0.63
Time-varying cancer	IPTW	18.3	23.1	4.9	1.27
Time-varying cancer	Upper bound	42.2	90.6	48.4	2.15
Time to cancer	Lower bound	16.9	5.7	-11.2	0.34
Time to cancer	IPTW	19.5	18.5	-1.0	0.95
Time to cancer	Upper bound	41.1	93.8	52.7	2.28

Figures

Figure 1: Distribution of participants under each health status, by age over follow-up

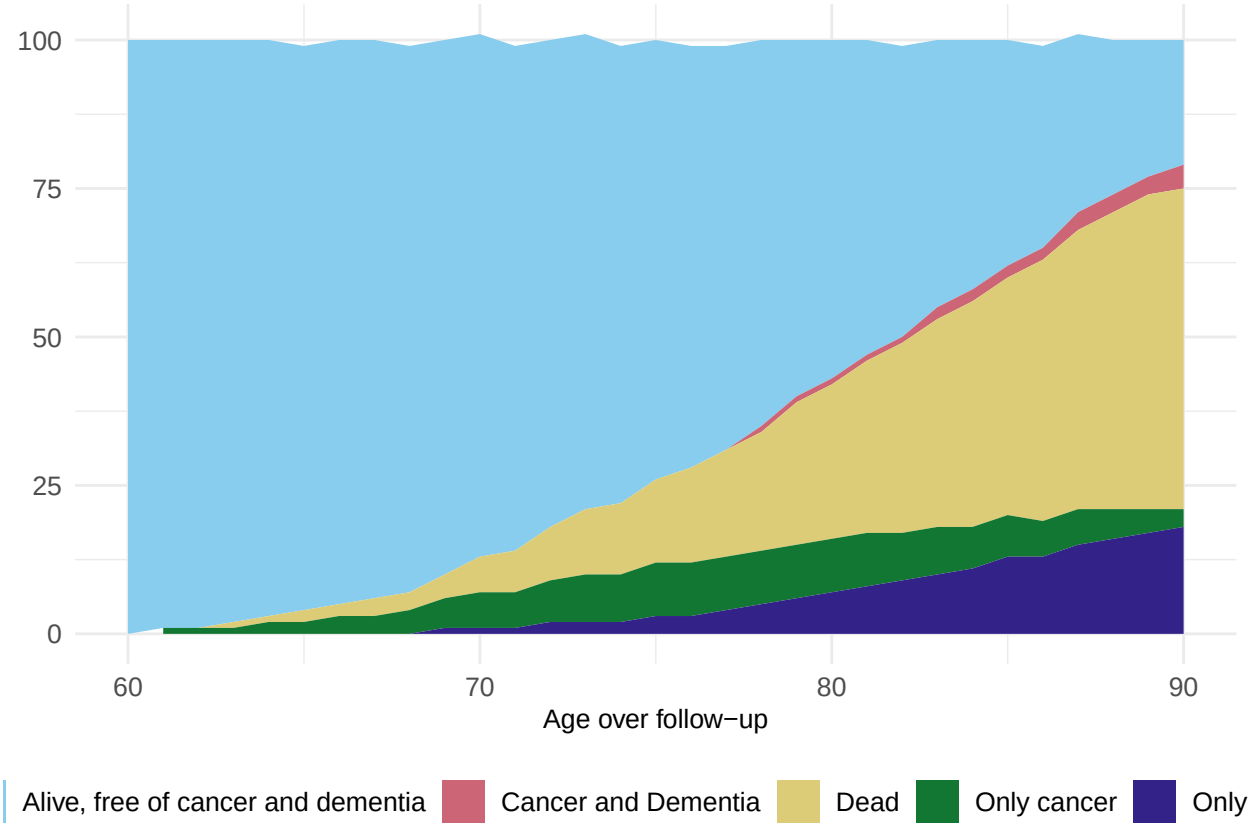
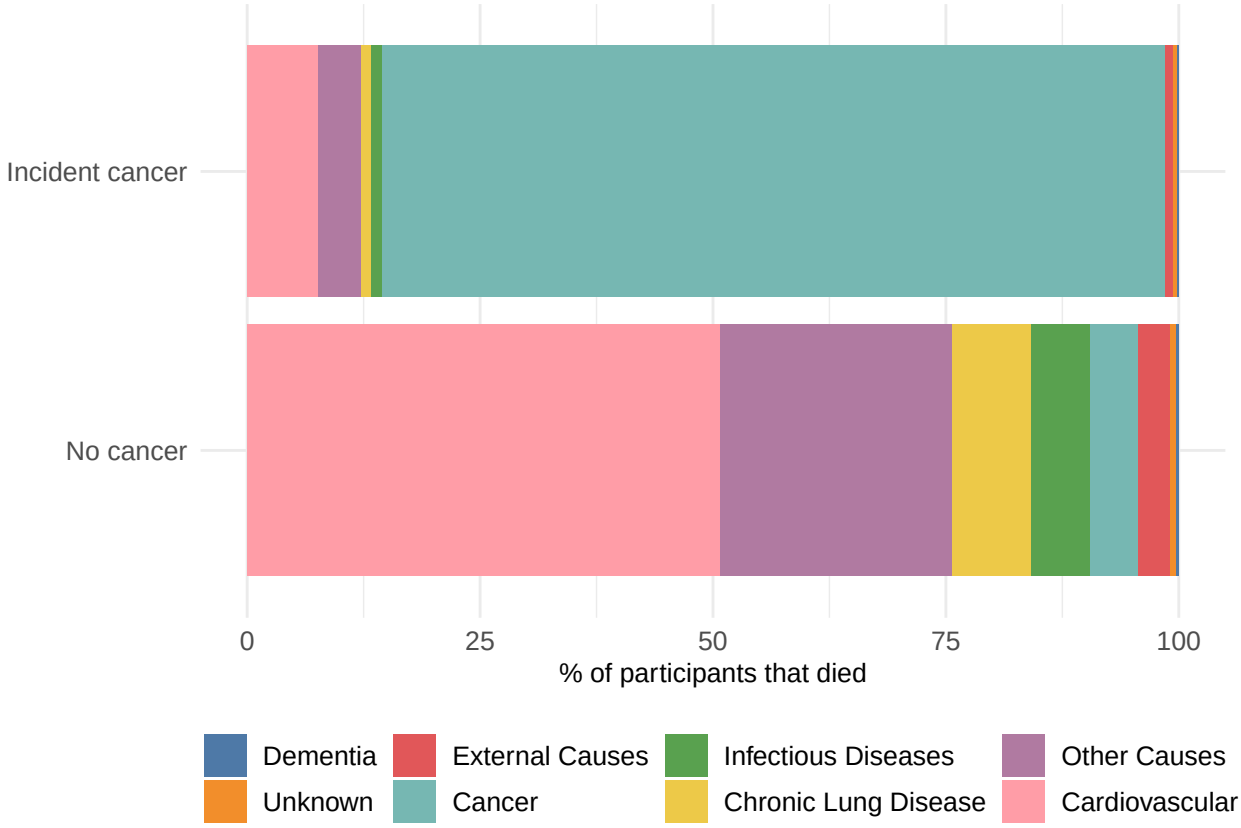


Figure 2. Causes of death for participants with and without incident cancer



cancer_definition	Models	Covariates
Cancer ever vs. never	IPTW denominator	cancer_v ~ bs(age_0, 3) + sex + education + as.factor(smoke1) + cohort,
Cancer ever vs. never	IPTW numerator	1
Cancer ever vs. never	IPCW Denominator	competing_plr ~ bs(age_0, 3) + sex + education + apoe4 + as.factor(smoke1) + ht1 + bs(sbp1, 3) + bs(bmi1,3) + as.factor(diabetes_prev) + cancer_v + cohort,
Cancer ever vs. never	IPCW Numerator	1
Time-varying cancer	IPTW denominator	cancer_v ~ bs(age_0, 3) + sex + education + as.factor(smoke1) + cohort,
Time-varying cancer	IPTW numerator	1
Time-varying cancer	IPCW Denominator	cancer_v + bs(time, 3) + bs(age_0, 3) + sex + education + apoe4 + as.factor(smoke) + bs(sbp, 3) + bs(bmi, 3) + ht + ht_drug + hd_v + stroke_v + diab_v + cohort
Time-varying cancer	IPCW Numerator	competing_plr ~ cancer_v + bs(time, 3) + cohort
Time to cancer	IPTW denominator	cancer_v ~ bs(age_0, 3) + sex + education + as.factor(smoke1) + cohort,
Time to cancer	IPTW numerator	cancer_v ~ bs(time, 3),
Time to cancer	IPCW Denominator	same as previous
Time to cancer	IPCW Numerator	same as previous