## 1. Introduction

Many observational studies have consistently found that individuals with cancer have a lower risk of developing of Alzheimer´s disease or related dementias (ADRD) when compared to individuals with no history of cancer1–4. These findings have motivated substantial research toward mechanistic explanations, including searching for and hypothesizing that molecular and genetic mechanisms may explain this association5–12. These research inquiries inevitably lead to discussions of repurposing or augmenting current cancer chemotherapeutics for ADRD13.

Nevertheless, inferring any treatment or mechanistic effects from the observed cancer-ADRD inverse association is not straightforward, and researchers have raised concerns related to the competing event of death, unmeasured confounding, and ascertainment error that could explain these results9,14. However, understanding these or other sources of bias first requires that we make explicit the causal question. Moreover, making explicit the causal question is one step toward tying a research study to a question that is relevant to decision-making.

To illustrate the complexities of inferring hypothetical or available treatments’ effects on ADRD from the observed cancer-ADRD association, we focus on the? Pin1 enzyme. Previous animal? studies have shown that the? Pin1 enzyme over-expression promotes tumorigenesis, while its down-regulation is attributed to mechanisms that contribute to Alzheimer’s Disease11,12,15. If we one day could develop a drug that increases Pin1 expression specifically in brain tissue in hopes of preventing dementia, we could pose the question as: \_What is the direct effect of this Pin1-targeting drug on the risk of ADRD over time compared to standard treatments?\_

To explore how we might learn about this effect using real-world data on cancer and ADRD, we progressively build a causal directed acyclic graph to connect this particular 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

If this hypothetical Pin1-targeting drug was developed, the best way to understand its effect on dementia risk would be to do a well-conducted randomized trial in which we randomize eligible participants in late midlife (e.g., ages 50-60 years) to receive this drug or not, and closely monitor ADRD over a lengthy follow-up. Since there is no drug currently available that targets Pin1, at best we can use observational data on Pin 1 expression measurements.

For example, suppose that a biomarker test was available to measure Pin1 and we measured this biomarker from (stored) baseline blood samples in a population based-cohort that is recruited in all participants in late midlife. Since the biomarker Pin1 is measured within an observational study, confounding can explain an observed association between it and ADRD. In Figure 1, we show that, Pin1 expression ($P$) and ADRD at time $t$ ($Y\_t$) may share causes $L$, and that assessing the causal relationship requires adjusting for these confounders $L$. 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 remodeling, may translate into Pin1 over-expression and are also related to the development of ADRD.

For simplicity, we treat Pin1 expression as a point intervention and we fix the time-ordering of covariates (that is, we assume $L$ happens prior to $P$). In reality, it is possible that Pin1 expression changes over time and is affected by time-varying confounders (like smoking), which produces treatment-confounder feedback loops. Addressing such time-varying confounding would require repeated measurements of $L$ and $P$16.

```{tikz, fig.ext = 'png', echo = FALSE, fig.cap = "Effect of Pin1 in ADRD"}

\usetikzlibrary{arrows}

\begin{tikzpicture}[node distance=2cm, auto,>=latex, scale = 0.5]

\node (p) {$P$};

\node [below of = p, yshift = 1cm, xshift = -1cm](l) {$L$};

\node [right of = p] (y) {$Y\_{20}$};

\draw[->] (l) -- (p);

\draw[->, violet] (p) -- (y);

\draw[->] (l) -- (y);

\end{tikzpicture}

```

However, unfortunately, in current studies we do not even have a single measurement of Pin1 expression, let alone repeated measurements. Thus, 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-expression4,17–24. We depict this feature in Figure 2, where $P^\*$ represents \_incident cancer diagnosis\_, the measured proxy of $P$. In this causal graph we colored the path $P^\* \leftarrow P \rightarrow Y\_{t}$ because although we are measuring the association between $P^\*$ and $Y\_{t}$ in the observed data, we are assuming that the captured effect is only through $P$.

```{tikz, fig.ext = 'png', echo = FALSE, fig.cap = "Effect of Pin1 in ADRD, with cancer diagnosis as proxy of Pin1"}

\usetikzlibrary{arrows}

\begin{tikzpicture}[node distance=2cm, auto,>=latex, scale = 0.5]

\node (p) {$P$};

\node [below of = p, yshift = 1cm, xshift = -1cm](l) {$L$};

\node [above of = p, yshift = -1cm](pstar) {$P^\*$};

\node [right of = p] (y) {$Y\_{85}$};

\draw[->, violet] (p) -- (pstar);

\draw[->, violet] (p) -- (y);

\draw[->] (l) -- (p);

\draw[->] (l) -- (y);

\end{tikzpicture}

```

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 registries19,23,25,26) 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 causal graph 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 group27). 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 covariates28. 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.

Furthermore, we note that this is only one of issues with considering cancer diagnosis as the proxy for Pin1 expression in terms of information bias (ref Miguel’s paper that makes all the DAGs in Chapter 9 of his book? Or another earlier paper that talks about types of measurement error). We could add more complexity by considering further shared causes of cancer and ADRD diagnosis, or further mediators between Pin1 expression and cancer diagnosis, including but not limited to: screening guidelines, type of healthcare coverage, health-seeking behaviors, and healthcare availability.

```{tikz, fig.ext = 'png', echo = FALSE, fig.cap = "Zoom in the association between P and P star"}

\usetikzlibrary{arrows}

\begin{tikzpicture}[node distance=2cm, auto,>=latex, scale = 0.5]

\node (p) {$P$};

\node [draw, right of = p] (d) {$D = 0$};

\node [right of = d] (pstar) {$P^\*$};

\node [below of = p, yshift = 1cm, xshift = -1cm](l) {$C\_1$};

\node [below of = d, yshift = 1cm, xshift = 1cm](c) {$C\_2$};

\draw[->] (p) -- (d);

\draw[->, violet] (p) to [out=45,in=135, looseness=0.4] (pstar);

\draw[->] (d) -- (pstar);

\draw[->] (l) -- (p);

\draw[->] (l) -- (d);

\draw[->] (c) -- (pstar);

\draw[->] (c) -- (d);

\end{tikzpicture}

```

Up to this point we have outlined certain underlying assumptions of using incident cancer diagnosis as a proxy for Pin1 expression. However, as mortality increases steeply in late life, even in the setting of the ideal randomized trial of a hypothetical Pin1-targeting drug, we can only measure ADRD over follow-up in the individuals who survive long enough to have a diagnosis. For this reason, death is a competing event of ADRD because if a participant dies prior to ADRD diagnosis, death prevents observing ADRD at future time-points. In the trial, we would then need to first decide whether we are interested in a total effect (that is, the effect on ADRD even if partly mediated by the effect on death) or a direct effect (that is, the effect directly on ADRD not mediated by death) (ref). (In the discussion, we will return to the special case of if the drug was so highly targeted toward brain tissue that it was by design only having a direct effect.) Based on this decision, we would then adapt our analytic strategy (refs).

In our observational data setting, this becomes even more challenging, since we are measuring cancer diagnosis as the proxy for Pin1, and cancer is a leading cause of death29. 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. 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}$28. 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. Since the question of interest is focused on the direct effect of $P$ in $Y\_{20}$ as in Figure 4 (violet arrows) we need to conceptualize the different mechanisms through which $P$ could affect $D\_{15}$ and $Y\_{20}$. With this in mind we can conceive different causal questions (estimands) to represent this direct effect, such as the controlled direct effect (CDE) 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) and leave the separable direct effects question for discussion.

```{tikz, fig.ext = 'png', echo = FALSE, fig.cap = "Direct effect of Pin1 in the risk of ADRD at 20 years of follow-up, with cancer diagnosis as proxy of Pin1"}

\usetikzlibrary{arrows}

\begin{tikzpicture}[node distance=2cm, auto,>=latex, scale = 0.5]

\node (p) {$P$};

\node [below of = p, yshift = 1cm, xshift = -1cm](l) {$L$};

\node [above of = p, yshift = -1cm](pstar) {$P^\*$};

\node [draw, right of = p] (d) {$D\_{15}$};

\node [right of = d] (y) {$Y\_{20}$};

\node [below of = d, yshift = 1cm, xshift = +1cm](c) {$C$};

\draw[->, violet] (p) -- (pstar);

\draw[->] (pstar) -- (d);

\draw[->] (p) -- (d);

\draw[->] (d) -- (y);

\draw[->] (l) -- (p);

\draw[->] (l) -- (y);

\draw[->] (c) -- (d);

\draw[->] (c) -- (y);

\draw[->, violet] (p) to [out=45,in=135, looseness=0.4] (y);

\end{tikzpicture}

```

The CDE represents the effect of Pin1 over-expression 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.

Therefore, if we combine the challenges in section two related to cancer diagnosis as a proxy for Pin1 over-expression, and having death as a competing event of ADRD we observe the complexity of the DAG in Figure 5. Though this is yet a simplified version since we omitted time varying $P\*$, $L$ and $C's$. We further summarize the three challenges raised, and the ways to mitigate or better understand them given current available data.

THIS SECTION IS STILL MISSING A CLEARER EXPLANATION OF WHEN WE ARE ABLE TO “TEST” VS ESTIMATE. THE MEASUREMENT ERROR BY DEFINITION MEANS WE ARE ONLY AT BEST TESTING THE ARROW WE CARE ABOUT. PLEASE GO BACK AND MAKE THIS CLEARER! IT COULD BE SOMETHING THAT BECOMES MORE PROMINENT IN THE TABLE?

```{tikz, fig.ext = 'png', echo = FALSE, fig.cap = " Direct effect of Pin1 in the risk of ADRD at 20 years of follow-up, with cancer diagnosis as proxy of Pin1"}

\usetikzlibrary{arrows}

\begin{tikzpicture}[node distance=2cm, auto,>=latex, scale = 0.5]

\node (p) {$P$};

\node [below of = p, yshift = 1cm, xshift = -1cm](l) {$L$};

\node [right of = p] (dp) {$D\_{<P\*}$};

\node [right of = dp] (pstar) {$P\*$};

\node [right of = pstar] (d) {$D\_{15}$};

\node [right of = d] (y) {$Y\_{20}$};

\node [above of = dp, yshift = -1cm, xshift = +1cm](c) {$C\_1$};

\node [above of = d, yshift = -1cm, xshift = +1cm](cd) {$C\_2$};

\draw[->] (p) -- (dp);

\draw[->] (dp) -- (pstar);

\draw[->, violet] (p) to [out=-45,in=-135, looseness=0.3] (pstar);

\draw[->] (pstar) -- (d);

\draw[->, violet] (p) to [out=-45,in=-135, looseness=0.3] (y);

\draw[->] (d) -- (y);

\draw[->] (l) -- (p);

\draw[->] (l) to [out=0,in=270, looseness=0.4] (y);

\draw[->] (c) -- (dp);

\draw[->] (c) -- (pstar);

\draw[->] (cd) -- (d);

\draw[->] (cd) -- (y);

\end{tikzpicture}

```

## 3?. Application to the Rotterdam Study

In this section we conduct an analysis of the cancer-ADRD association that is structured to the best of the available data’s abilities to inform the possible effect of Pin1 on 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 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, `r total\_n` 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 `r cancer\_5y` and the risk of death was `r death\_5y` 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 `r mean\_age`, and `r women\_prop` were women. Further details on participants are presented in Table 1. Over follow-up, `r yes\_cancer` developed cancer and `r no\_cancer` remain free of cancer diagnosis, the median age of cancer diagnosis was `r cancer\_time`. From the total sample, `r dem\_total` had dementia over follow-up and median time to dementia was `r dem\_time`. Among participants with incident cancer, `r cancer\_dem` had dementia diagnosis and `r cancer\_death` died over follow-up, `r cancer\_alive` remain alive at 20 years since study entry. In contrast, among participants free of cancer diagnosis over follow-up, `r no\_cancer\_dem` were diagnosed with dementia and `r no\_cancer\_death` died over follow-up, `r no\_cancer\_alive` 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: `r table\_results[2,6]` (95%CI), HR: `r table\_results[2,7]`]. However this effect is diminished [RR: `r table\_results[3,6]` (95%CI),

HR: `r table\_results[3,7]`)] 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: `r table\_results[6,6]` (95%CI), HR: `r table\_results[6,7]`)].

[should include some final comments on this being interpreted as back to the original question about Pin1, rather than being results about the particular estimates from the models too, but I think we need to discuss the analyses first.]

## 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. And if we consider treatments given at a different time in life too. And how we want to think about which effect is of interest given competing events. 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 or any closely-related direct effect kind of question? 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 (`r cancer\_death`), 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.

<!-- Causes such as aging, smoking, irradiation etc are causing micro environmental changes, like inflamation and changes in tissue remodelling. This promotes selection for adaptive mutations.Youthful tissues -->

- 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.

- We could have change Pin1 to other molecular mechanism (MAYBE). 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.

* I don’t know if it is here or elsewhere, but somewhere we need to make super clear that up to this point you also are still not ESTIMATING the effect from the target trial at all. You are helping rule out or think through a bunch of backdoor paths, but the measurement error layers can be reconceptualized as saying that we are either bounding (?) the effect size MAYBE (have to think about that), or more humbly we are just testing the sharp null hypothesis. (Again being humble about what we learn is okay! This isn’t your doing. This is you trying to make sense of what others are doing. Let’s be super clear what they are able to learn with what assumptions.)

My comments on the graphs:

* I’m going to try to explain a slightly different vision for how to build a DAG in figures, but it doesn’t affect the logical flow yet so will not try to explain that til the next time we meet.
* Figure 2: should we also start adding a bit on the measurement error structures here too (see Chapter 9 of the CI book)? This is just the simplest best-case scenario of it being a proxy, in terms of measurement. Or maybe you can also just say that along the way of talking about Figure 3, that Figure 2 is the easiest version of proxy status to work with but there is a lot to unpack and Figure 3 is one kind of example. Added this to description of figure 3
* If you agree with me that we start with the target trial, then Figure 1 is the challenge of using observational data, Figure 2-3 are the challenges of using proxy data, and Figure 4 is the challenge of competing risk. Kind of a nice flow? Maybe make Figure 4 JUST the part relevant to the competing risk to begin. Great, do you thing now is okay?
* Again I have a slightly different vision for how to begin it all, but I think the current Figure 4 could become a Figure 5 that better says we are putting all the pieces together in one place. But then I think you might want to also make clear that there is a D between P and P\*. And then this might also mean getting time-varying Ds and P\*s in the mix. I don’t think you want the time-varying issue to appear in the figures that introduce each bias, but I think it’d be helpful in the final stage to be like “whoa whoa whoa don’t forget that our simplified DAGs above are trying to succinctly summarize some really complex things.” Your doubts about covariate selection and even what estimator we use come down to this kind of complexity. So let’s add that complexity and then show under what asusmptions you would choose to adjust for smoking or not. In this case I think we will end up showing that in both cases you are biased (because you want to adjust for smoking pre-time zero of the target trial of when you’d give the drug, but not thereafter) but then it becomes very clear to your readers why this is a tradeoff with the type of data we have. Right? Yes, added Figure 5.
* I have some minor ideas on using colored arrows but I think it is better to wait until the document is further along.

\newpage

1. Ma, L. L. *et al.* Association between cancer and Alzheimer’s disease: Systematic review and meta-analysis. *J. Alzheimer’s Dis.* **42**, 565–573 (2014).

2. Hanson, H. A., Horn, K. P., Rasmussen, K. M., Hoffman, J. M. & Smith, K. R. Is Cancer Protective for Subsequent Alzheimer’s Disease Risk? Evidence from the Utah Population Database. *Journals Gerontol. - Ser. B Psychol. Sci. Soc. Sci.* **72**, 1032–1043 (2017).

3. van der Willik, K. D., Schagen, S. B. & Ikram, M. A. Cancer and dementia: Two sides of the same coin? *Eur. J. Clin. Invest.* **48**, 1–12 (2018).

4. Ospina-Romero, M. *et al.* Association Between Alzheimer Disease and Cancer With Evaluation of Study Biases: A Systematic Review and Meta-analysis. *JAMA Netw. open* **3**, e2025515 (2020).

5. Behrens, M., Lendon, C. & Roe, C. A Common Biological Mechanism in Cancer and Alzheimers Disease? *Curr. Alzheimer Res.* **6**, 196–204 (2009).

6. Harris, R. A., Tindale, L. & Cumming, R. C. Age-dependent metabolic dysregulation in cancer and Alzheimer’s disease. *Biogerontology* **15**, 559–577 (2014).

7. Nudelman, K. N. H., McDonald, B. C., Lahiri, D. K. & Saykin, A. J. Biological Hallmarks of Cancer in Alzheimer’s Disease. *Mol. Neurobiol.* **56**, 7173–7187 (2019).

8. Papin, S. & Paganetti, P. Emerging evidences for an implication of the neurodegeneration-associated protein tau in cancer. *Brain Sci.* **10**, 1–24 (2020).

9. Driver, J. A. Inverse association between cancer and neurodegenerative disease: review of the epidemiologic and biological evidence. *Biogerontology* **15**, 547–557 (2014).

10. Olson, B. & Marks, D. L. Pretreatment cancer-related cognitive impairment—mechanisms and outlook. *Cancers (Basel).* **11**, 1–18 (2019).

11. Li, J. *et al.* Roles of peptidyl-prolyl isomerase Pin1 in disease pathogenesis. *Theranostics* **11**, 3348–3358 (2021).

12. Driver, J. A., Zhou, X. Z. & Lu, K. P. Pin1 dysregulation helps to explain the inverse association between cancer and Alzheimer’s disease. *Biochim. Biophys. Acta - Gen. Subj.* **1850**, 2069–2076 (2015).

13. Snyder, H. M. *et al.* Exploring the nexus of Alzheimer’s disease and related dementias with cancer and cancer therapies: A convening of the Alzheimer’s Association & Alzheimer’s Drug Discovery Foundation. *Alzheimer’s Dement.* **13**, 267–273 (2017).

14. Ganguli, M. Cancer and dementia. *Alzheimer Dis. Assoc. Disord.* **29**, 177–182 (2015).

15. Angelucci, F. & Hort, J. Prolyl isomerase Pin1 and neurotrophins: A loop that may determine the fate of cells in cancer and neurodegeneration. *Ther. Adv. Med. Oncol.* **9**, 59–62 (2017).

16. Hernán, M. A. & Robins, J. M. Causal Inference: What If (Harvard book) - Last version. (2019).

17. Musicco, M. *et al.* Inverse occurrence of cancer and Alzheimer disease: A population-based incidence study. *Neurology* **81**, 322–328 (2013).

18. Aiello Bowles, E. J. *et al.* Risk of Alzheimer’s disease or dementia following a cancer diagnosis. *PLoS One* **12**, 1–11 (2017).

19. Ording, A. G. *et al.* Cancer and risk of Alzheimer’s disease: Small association in a nationwide cohort study. *Alzheimer’s Dement.* **16**, 953–964 (2020).

20. Sun, M., Wang, Y., Sundquist, J., Sundquist, K. & Ji, J. The Association Between Cancer and Dementia: A National Cohort Study in Sweden. *Front. Oncol.* **10**, (2020).

21. Frain, L. *et al.* Association of cancer and Alzheimer’s disease risk in a national cohort of veterans. *Alzheimer’s Dement.* **13**, 1364–1370 (2017).

22. Schmidt, S. A. J., Ording, A. G., Horváth-Puhó, E., Sørensen, H. T. & Henderson, V. W. Non-melanoma skin cancer and risk of Alzheimer’s disease and all-cause dementia. *PLoS One* **12**, 1–13 (2017).

23. Freedman, D. M. *et al.* Associations between cancer and Alzheimer’s disease in a U.S. Medicare population. *Cancer Med.* **5**, 2965–2976 (2016).

24. Driver, J. A. *et al.* Inverse association between cancer and Alzheimer’s disease: Results from the Framingham Heart Study. *BMJ* **344**, 19 (2012).

25. Bolch, C. A. *et al.* Inverse probability of treatment-weighted competing risks analysis: An application on long-term risk of urinary adverse events after prostate cancer treatments. *BMC Med. Res. Methodol.* (2017) doi:10.1186/s12874-017-0367-8.

26. Mezencev, R. & Chernoff, Y. O. Risk of alzheimer’s disease in cancer patients: Analysis of mortality data from the US SEER population-based registries. *Cancers (Basel).* **12**, (2020).

27. Soriano, J. B. *et al.* Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir. Med.* **8**, 585–596 (2020).

28. Young, J. G., Stensrud, M. J., Tchetgen, E. J. T. & Hernán, M. A. A causal framework for classical statistical estimands in failure time settings with competing events. *Stat. Med.* **1**, 1–38 (2020).

29. Bennett, J. E. *et al.* NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4. *Lancet* **392**, 1072–1088 (2018).