CORAL
DIVISION OF
CHRONIC DISEASE RESEARCH ACROSS THE LIFECOURSE

PROJECT Viva

Guidelines for Analyzing and Reporting Project Viva Data

Over the years we have found ourselves making the same suggestions over and over to trainees and co-investigators proposing analyses using Project Viva data. We therefore compiled several of these suggestions into this document, which we hope you will use as a reference as you are preparing your analysis plans, analyses, and manuscripts. We jokingly refer to this document as our "manifesto" but in reality, it is meant to be a working document – so please feel free to question anything that is here or suggest edits and additions. Thanks!

For reporting:

- Remember that Viva is not an acronym (it's Viva, not VIVA).
- Please always use active voice in your writing. Active voice generally requires fewer words, and more clearly communicates who did what.
- Use past tense when reporting what you already did (e.g. in a manuscript).
- Avoid labelling people with their diseases; use people-first language. For example, instead of "among obese asthmatics", write "among children with obesity and asthma."

For analyses:

- Usually we are evaluating strengths and directions of associations, rather than binary yes/no or present/absent.
    - Thus, we prefer aims to examine "the extent to which" an exposure is associated with an outcome, rather than "testing whether or not" an exposure is associated with an outcome.
    - In parallel, when reporting results, we prefer to avoid an excessive focus on p-values (especially in isolation) and thus an inevitable arbitrary division of results into "significantly different" or "not significantly different." Instead, we focus on effect estimates, together with corresponding 95% confidence intervals to interpret study findings and place them within a biological context.[1]
    - Similarly, when interpreting confidence intervals, consider the precision and magnitude of potential effects contained within the CI, not just whether or not it includes or excludes the null.
    - P-values are, however, useful in GWAS or EWAS analyses (typically adjusted for multiple testing) to indicate significant hits, when we are looking to see if interactions are significant (see below), or when making other decisions. There may be other circumstances when p-values are helpful, but please be thoughtful about whether they are needed or not.
- Use of race/ethnicity in analyses. Many exposures and outcomes vary by race/ethnicity, but it is well understood that these differences are the result of social and structural racism rather than any fixed or biological differences between identify groups. We encourage investigators to learn and apply best practices related to the use of race/ethnicity in their analyses (see e.g. Kaplan et al., 2003[2]).
- When choosing covariates to include in analyses as potential confounders, first define your causal question (see below). Once you define your question, use a causal diagram (e.g., directed acyclic graph, DAG) to select confounders based on *a priori* assumptions/knowledge about underlying relationships between measured and unmeasured variables. For those unfamiliar with the use of causal diagrams for encoding assumptions about confounding for different types of causal questions, here are high-level guidelines for selecting potential confounders for common types of causal questions considered in Project Viva:

[1] *Greenland et al., Statistical tests, P values, confidence interval, and power: a guide to misinterpretations, Eur J Epidemiol, 2016; Ranstam et al., Why the P-value culture is bad and confidence intervals a better alternative, Osteoarthritis and Cartilage, 2012. Sterne JAC and Davey Smith G. Sifting the evidence—what's wrong with significance tests? BMJ 2001;322:226–31*

[2] Kaplan JB, et al. Use of race and ethnicity in biomedical publication. JAMA. 2003. PMID: 12771118

  - Total effect of a time-fixed exposure: consider all suspected common causes of exposure and outcome. Do not include covariates that are themselves affected *by* the exposure. Keep in mind that if addition of a covariate changes the effect estimate for the exposure-outcome relationship, this does not mean that the covariate is a confounder. For example, adjusting for a covariate affected by an exposure may change the estimate because of "collider bias." If there is uncertainty about whether a covariate is a confounder or an effect of exposure, conduct sensitivity analysis adjusted and not adjusted for this variable and explicitly discuss this consideration in the paper.
  - After selection of potential confounders based on the above *a priori* approach, you may have too many covariates given your sample size. If so, you can remove covariates based on *a priori* assumptions about which covariates are the strongest confounders. You can then sequentially add back additional sets of covariates to see if the conclusions change meaningfully.
  - Sometimes investigators choose to use automated model selection approaches, particularly in the absence of *a priori* knowledge. We discourage this approach. Be aware that confidence intervals and p-values do not retain their usual properties after this sort of covariate selection process.
  - If one of your aims focuses on mediation, or if you plan to examine time-varying exposures, please consult with Jessica Young, Project Viva biostatistician, for more information about how to approach these questions. (also see below)
  - Consider whether variables might be effect modifiers and if so please include this consideration in your DAG/analysis plan. If you find evidence that a variable is an effect modifier (e.g. different effect estimates in stratified analyses, or significant interaction terms), make sure you report results in light of this modification. Because of accumulating evidence that many early life exposures have sex-specific associations with outcomes, we suggest that investigators plan to investigate effect modification by sex, at a minimum. (more on this topic below)
- If the goal of your analysis is prediction, then there is no "exposure" or "confounders" (see Conroy and Murray[3]). These concepts don't apply in that case, they apply only when the goal is to estimate a causal effect. The exposure is the variable you would hypothetically intervene on, if you could, to determine the effect on the outcome. Confounders are what you must control to get at that exposure effect in observational data where exposure was not randomized. You know you are doing causal inference (whether implicitly or explicitly) when you try to adjust for confounders and you report a contrast in the outcome risk/mean based on different levels of exposure (or maybe jointly of several exposures together). A typical way to report this is via the estimated coefficient on exposure from a regression model but there are other ways (e.g. weighted methods or matching). By contrast, when the goal is prediction, you are trying to quantify the chance of a future outcome based on a set of current observed covariates (e.g. what is the future risk of obesity among infants who have mothers with a particular set of characteristics). Sensitivity and specificity are measures that can be reported for your prediction approach when the outcome is binary. Standard regression may be used or more sophisticated approaches for coming up with a prediction algorithm.[4]
- When articulating a causal question, it is useful to think about the trial (intervention) you would ideally run to answer that question.[5,6] That may be difficult when the exposure is selected such that we don't know exactly how to implement that trial (e.g. we state interest in an "effect of obesity" on a later outcome). In this case, you might consider changing your goal to prediction (e.g. rather than trying to estimate an "effect of obesity on X outcome," you might ask is "obesity predictive of X outcome?"). If using regression, adjustment for covariates is

[3] Conroy S, Murray EJ. Let the question determine the methods: descriptive epidemiology done right. British Journal of Cancer 2020. https://doi.org/10.1038/s41416-020-1019-z.

[4] Sherri Rose. Mortality Risk Score Prediction in an Elderly Population Using Machine Learning. *American Journal of Epidemiology*, Volume 177, Issue 5, 1 March 2013, Pages 443–452.

[5] Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64.

[6] Chiu Y, Rifas-Shiman SL, Kleinman K, Oken E, Young JG. Effects of intergenerational exposure interventions on adolescent outcomes: an application of inverse probability weighting to longitudinal pre-birth cohort data. *Paediatric and Perinatal Epidemiology*. 2020 May;34(3):366-375.

not needed although other variables might be included to answer the question of whether, in this example, obesity remains predictive once fixing those other variables. Because there is no need to include these other variables (they are not needed to control confounding as this is not a relevant concept for prediction as per above), a small set can be considered. Stratified analyses are an alternative to a regression model (which may be mis specified). E.g. we might ask "is obesity predictive of the outcome among girls, or among boys?" When the analysis is motivated and structured for the goal of prediction, the results may inform where intervention resources should go (e.g. let's prioritize resources for implementable interventions that may mitigate the outcome in children with obesity because they are at high risk for outcome X) rather than informing what would happen if we (somehow) eliminated obesity.

- Often, one or more variables will have some missing values.
    - In general, our policy in Project Viva is to use multiple imputation to address missingness. We typically conduct the imputation among all 2128 Viva births, including all exposures, outcomes, and covariates in the models as well as any additional variables that may be helpful in predicting the missing values. Then, we conduct the analyses among those who were 'eligible' for having the exposure
    - In some cases, when very few values are missing (e.g. <5% missingness,) or in extremely high dimensional analyses such as EWAS, it may be appropriate to use a complete case analysis approach.
- Consider the number of comparisons you are making and whether you need to account for multiple testing. However, because often we are examining multiple correlated outcomes that are not independent of each other, we generally (except with GWAS/EWAS) do not use overly conservative methods such as Bonferroni correction. Sometimes we do calculate false discovery rate (FDR) adjusted CI's or p values, based on the number of tests, or the number of classes of outcomes. More often, since we often have correlated outcomes (e.g. weight, BMI, %body fat, and FMI), we look at patterns of associations and are conservative in overcalling isolated significant findings.
- Remember to check model assumptions early in the analysis process. For example, with a continuous exposure and outcome, check for a linear relationship, don't assume it; remember look for normality and homoscedasticity of model residuals before moving immediately to transform an exposure or outcome that is not normally distributed, as native units are often more readily interpretable.

<u>More on post-exposure variables: mediators and effect modifiers</u>

- **Mediators**
    - <u>A mediator is, by definition, a post-exposure variable that is possibly on the causal pathway from exposure to outcome.</u> Consider interest in the causal effect of an exposure (e.g. maternal smoking in pregnancy) on an outcome (e.g. infant growth). We might consider a post-exposure pregnancy complication (e.g. route of delivery) a possible mediator. <u>By definition, a mediator is affected by the exposure.</u>
    - We don't have to account for mediators in our analysis if we limit our target of analysis to total exposure effects; e.g., we don't need to estimate the mediating effect of route of delivery if what we care about is the total effect of smoking in pregnancy on infant growth. Total effects require conceptualizing some intervention on exposure (e.g. if we had assigned women to be either smokers or nonsmokers), as well as adjusting for common causes of exposure and outcome (i.e. confounders). You can estimate a total effect directly from trials, in which the exposure is assigned randomly (not ethical in our smoking example but we could at least imagine an unethical trial we might conduct). If we use observational data such as we have in Viva, we have to try to get at what we would have estimated had we been able to conduct such a trial. This is something we try to do via confounding adjustment.
    - Our goal when assessing mediators is to estimate direct and indirect effects of exposure on the outcome, i.e. effects through and outside of the mediator. Often, we imagine the putative mediator to be a candidate target for intervention – e.g. in our hypothetical smoking trial, we require all women in

both arms to delivery vaginally, or we require them all to deliver via cesarean. However, valid interpretation of the mediation estimation results as causal intervention effects with respect to some intervention on the mediator requires stronger assumptions about confounding. We must also adjust for all common causes of the mediator and outcome in that case and, for certain types of mediation questions, also all common causes of the exposure and mediator.

- **Effect Modifiers**
  - Effect modifiers are covariates that describe features of the study population within which we want to understand the causal effect of exposure on the outcome and, within levels of those covariates, this causal effect differs. We do not require conceptualizing interventions on effect modifiers, again these are covariates on which we simply wish to stratify the population. Generally, it is a bad idea to consider post-exposure variables as effect modifiers because stratifying on such covariates can create collider bias (see Jessica for references). When we restrict ourselves to considering effect modification by only pre-exposure covariates, just like with total effects, we need to account only for confounding by common causes of the exposure and outcome. In our example above of prenatal smoking and infant weight gain, we might stratify by maternal race/ethnicity, but we should not stratify by mode of infant feeding, or any other post-exposure variable. If you think about our hypothetical trial in which we randomly assign women to be smokers or non-smokers, we would stratify only on characteristics that are present at the time of randomization.
  - Note that, by allowing a pre-exposure covariate to be an effect modifier in our analysis, we are always stratifying on it in some way (whether via explicit stratification or via inclusion of main and interaction terms in our regression model, which is a model-based type of stratification). By this stratification we are also dealing with any potential confounding by that covariate as well. A baseline covariate may be both an effect modifier and a confounder, one of these, or neither. By contrast, a mediator of an exposure effect cannot be a confounder of that effect and generally cannot be an effect modifier of that effect because we should restrict possible effect modifiers to pre-exposure variables.
  - There are some cases in which we might be unclear about the temporal order of exposure and a covariate, particularly in cohort studies in which measurement of exposure and covariates are taken at the same time. For example, in Viva, if our exposure is prenatal smoking, should gestational weight gain, which is measured over the same period, be considered as a confounder or effect modifier or a possible mediator? In these cases, sensitivity analyses might be warranted where gestational weight gain is treated in different ways (e.g. in one analysis as an effect modifier and in another as a possible mediator). But the assumptions and interpretation across those sensitivity analyses will be quite different as per above and should be considered explicitly.
  - An unstratified outcome regression that includes only main terms for exposure and possible confounders is relying on the assumption of no effect modification. When this is used for the main analysis, a secondary analysis that allows effect modification will contradict the assumptions of the main analysis. To avoid this contradiction, the main analysis can instead allow effect modification. There are two ways to allow effect modification: one is by doing separate analyses within each stratum of the assumed effect modifier and the other is to estimate effects in a single model allowing interaction terms between exposure and the assumed effect modifier. The former is generally more robust to model misspecification while less precise. The latter is generally more prone to model misspecification while more precise.

- **Acknowledgements of Viva Funding Support in Publications**

  Please see the Project Viva list of grants to review when to cite specific grants.

  NOTE: Feel free to err on the side of over-acknowledging grants rather than missing an acknowledgment. There is no harm in acknowledging an extra or closed grant but if you acknowledge an active NIH grant, it must be compliant with a **PMCID**. If unsure of how to make a publication compliant, please contact Molly Ahern at Molly_Ahern@HarvardPilgrim.org.

Guidelines developed by:

Marie-France Hivert, MD SM
Ken Kleinman, ScD
Emily Oken, MD MPH
Karen Switkowski, PhD MPH
Jessica Young, PhD

Contact: project_viva@harvardpilgrim.org

1/21/21