Guidelines for the Final Project

Examples of papers written based on the final project of this course/section:

- Hossain MB**, Kopec J, Atiquzzaman M**, Karim ME. The association between rheumatoid arthritis and cardiovascular disease among adults in the United States during 1999-2018, and age-related effect modification in relative and absolute scales. Annals of Epidemiology. Published online: March, 2022 **https://doi.org/10.1016/j.annepidem.2022.03.005 (https://doi.org/10.1016/j.annepidem.2022.03.005)**
- Yusuf F**, Tang T, Karim ME "The association between Diabetes and Excessive Daytime Sleepiness among American adults aged 20-79 years: Findings from the 2015-2018 National Health and Nutrition Examination Surveys", Annals of Epidemiology, Published online Feb 1, 2022. **https://doi.org/10.1016/j.annepidem.2022.01.002 (https://doi.org/10.1016/j.annepidem.2022.01.002)**
- Delgado-Ron JA**, López-Jaramillo P, Karim ME (2020) Sodium intake and high blood pressure among adults on caloric restriction: a multi-year cross-sectional analysis of the U.S. Population, 2007-2018, Journal of Human Hypertension, Published: 27 September 2021, **https://doi.org/10.1038/s41371-021-00614-4 (https://doi.org/10.1038/s41371-021-00614-4)**
- Iyamu IO**, Oladele EA, Eboireme E, Karim ME. Is regular access to internet services associated with comprehensive correct HIV/AIDS knowledge among people aged 15-49 years in Nigeria? Findings from the 2018 Demographic Health Survey. *Journal of Consumer Health on the Internet*, Volume 25 Issue 3, Published online: 09 Sep 2021, 2021; doi: **https://doi.org/10.1080/15398285.2021.1943634 (https://doi.org/10.1080/15398285.2021.1943634)**
- Rossa-Roccor V**, Karim ME (2021) Are US adults with low-exposure to methylmercury at increased risk for depression?: A study based on 2011-2016 National Health and Nutrition Examination Surveys. *International Archives of Occupational and Environmental Health*, 94, pages419–431, **https://doi.org/10.1007/s00420-020-01592-9 (https://doi.org/10.1007/s00420-020-01592-9)**
- Nikiforuk A**, Karim ME, Jassem A, Patrick D (2020) Influence of Chronic Hepatitis C Infection on the Monocyte-to-Platelet Ratio a Complete Blood Count Biomarker in the United States: A Cross-Sectional Study of the National Health and Nutrition Examination Survey (NHANES) from 2009-2016. *BMC Public Health*, (2021) 21:1388, doi: **https://doi.org/10.1186/s12889-021-11267-w (https://doi.org/10.1186/s12889-021-11267-w)**
- Jeong, D.**, Karim, M. E., Wong, S., Wilton, J., Butt, Z. A., Binka, M., ... & Janjua, N. Z. (2021). Impact of HCV infection and ethnicity on incident type 2 diabetes: findings from a large population-based cohort in British Columbia. *BMJ Open Diabetes Research and Care*, *9*(1), e002145, doi: **http://dx.doi.org/10.1136/bmjdrc-2021-002145 (http://dx.doi.org/10.1136/bmjdrc-2021-002145)**
- Nisingizwe MP**, Tuyisenge G, Hategeka C, Karim ME. (2020) Are perceived barriers to accessing health care associated with inadequate antenatal care visits among women of reproductive age in Rwanda?, *BMC Pregnancy and Childbirth*, 20(88): **https://doi.org/10.1186/s12884-020-2775-8 (https://doi.org/10.1186/s12884-020-2775-8)**
- McLeod, KE**; Karim, M.E. (2020) The relationship between mood disorders and experiencing an unmet healthcare need in Canada: Findings from the 2014 Canadian Community Health Survey. *Journal of Mental Health* Published online: 27 Aug 2020, **https://doi.org/10.1080/09638237.2020.1818192 (https://doi.org/10.1080/09638237.2020.1818192)**

- Closson K**, Karim ME, Sadarangani M, Naus M, Ogilvie GS, Donken R (2020) Impact of human papillomavirus vaccine status on sexual-activity related outcomes among young women recommended for vaccination in the United States, *Vaccine*, 38 (52): Pages 8396-8404, 14 December 2020, **https://doi.org/10.1016/j.vaccine.2020.10.033** **(https://doi.org/10.1016/j.vaccine.2020.10.033)**
- Basham, C**; Karim, M.E. (2019) Multimorbidity prevalence in Canada: A Comparison of Northern Territories with Provinces, 2013/14. *International Journal of Circumpolar Health*, 78:1, 1607703, **https://doi.org/10.1080/22423982.2019.1607703** **(https://doi.org/10.1080/22423982.2019.1607703)**
- Nethery, E.**, Schummers, L., Maginley, K. S., Dunn, S., & Norman, W. V. (2019). Household income and contraceptive methods among female youth: a cross-sectional study using the Canadian Community Health Survey (2009–2010 and 2013–2014). *CMAJ open*, 7(4), E646. **https://doi.org/10.9778/cmajo.20190087** **(https://doi.org/10.9778/cmajo.20190087)**

NOTE: You will note that some of these papers' formats / writing style deviate from the outlined structure provided in this current course. That is usually because the authors tried to adhere to the journal's submission format, or changes occurred during the process of addressing the reviewers' or editors' comments or because course instructions evolved over time. Please check for the exact rubrics provided for your assignment for clear guidance of what is expected from your assignments.

How to choose a topic for the 604 final projects?

- Figure out what interests you in general.
- Better to choose a topic/disease area about which you or your support network have some knowledge.
- Then look for something a bit more specific from the data dictionary (for **public data**, not the microdata!) to figure out what variables are available that will enable you to answer that research question
- If not, reiterate!
- The final paper that you write for this course/section **can not be one of your Ph.D. thesis chapters**.

Some general considerations while choosing a research topic (for 604 final projects):

1. Be clear about your **study objective/hypothesis** (pick 1 primary objective).
2. One of the relevant factors for choosing an 'aim' for this class project is **time**. By the end of Sept, students are supposed to create the analytic data and run at least the first set of the primary analysis (e.g., linear or logistic regression based on their research question) as well as the primary set of diagnostic analysis for their regression. When you pick a research topic, please keep this timeline in mind.
3. In this course, we highly recommend using **openly accessible survey data** (that requires **no ethics application**; with a well-documented probabilistic sampling design; non-probabilistic surveys are strongly discouraged) that provides detailed survey feature information (such as strata, cluster and survey weights). One popular data source is NHANES (CCHS public data lacks some of this info). Since strata and cluster information are essential for standard error calculation (that we will learn later in the course), any data without such info would be problematic in terms of inference and generalization.
4. At the very least, 3 cycles of data are a requirement for this course, for those using cross-sectional survey data. More is encouraged. Students from this class previously used up to ~20 cycles of data (usually necessary for questions with rare exposure or outcome).
5. Be sure to check that the **data dictionary** you are consulting is from a **public-use data**, not for the master data (for which you will need additional permission).

6. Make sure the **variables relevant to your aims** (exposure, outcome and important potential confounders) are **present in your data**.
7. Be sure to pick **one outcome of main interest** relating to your hypothesis. If there is another related outcome variable in your dataset that closely resembles your outcome variable, consider that for sensitivity analysis.
8. Be sure to pick **one exposure** of main interest relating to your hypothesis.
9. Try choosing a **binary outcome and a binary exposure**, if possible. Within the **survey data analysis context**, running multinomial regression, ordered categorical regression or survival analysis (and associated diagnostics or sensitivity analyses) can be hard sometimes, in part due to the unavailability of properly developed software.
10. The definition of the "target population" and a viable way to link your sample to that population is important (usually more problematic for non-probability sampling data). I need an explanation of how the sampling was done, what is the target population you are aiming for? The population component you write in your PICOT should be your first step in thinking about the target population.
    - Firstly, how can you infer or project about that or any well-defined population from needs to be clarified. Is this a provincewide data? Are you claiming that this data has captured nearly all such individuals from the target population under some well-defined inclusion/exclusion criteria? Or a sample of them where you know what is the probability of each person getting selected, so that you can infer what can happen at the population level?
    - Secondly, when you make statistical inferences about such a population, what distributional assumptions do you make if you do not know the probability of each participant being independent or clustered?
    If you can't answer these questions, then the study could be considered as a purely descriptive study, and usually outside of the scope of a final project. Take a look at one or more publications that happened using this same data (if any) to get a better understanding of the setup.
11. While selecting an exposure group, the best strategy is to think about whether the exposure is modifiable or not. If not (e.g., one example is 'race' for a participant), then you may have trouble justifying the confounder list. Most of the list of associated variables may end up being mediators instead.
12. Be sure to **thoroughly search the literature**, whether this same question has already been answered in the same data/context/time-period. If the question is already answered in recent literature, you should try to pick another question; or pick a different but interesting angle (answering a somewhat different question).
13. I encourage you to think harder about whether you have considered <u>all of the covariates that could be useful in explaining the relationships</u> (outcome vs. exposure, or outcome only or exposure only predictors, with the understanding of what are potentially IV and mediators) they are considering. On your part, that may require some literature search. However, CCHS/NHANES/similar public datasets have a limited variable list, and the general recommendation would be to include all possible potential confounders that are 'relevant' for your hypothesis, at least at the initial stage of the analysis, even if you think some of them are weak confounders/risk factors. However, to be clear: that does not mean including a long list of irrelevant variables; they have to be somewhat relevant in affecting the relationship of interest. Later, you can weed out weaker/unhelpful confounders once you proceed with your proper analysis. A reviewer will always look for an opportunity to question about whether you have considered all the available confounders in your dataset. Examples of some of the variables that are frequently useful in explaining many of the research topics (may not be relevant for all questions):
    - age, sex, race/ethnicity, birth country, marital status,
    - income/employment status, SES, education, working status/student status

- history of chronic conditions (mostly the ones relevant for outcome and exposure of choice, often in the form of co-morbidity or multimorbidity, or individually), perceived health status, mental health (e.g., depression), access to healthcare, hospitalization, type of medication used, has basic health insurance coverage, has extended health insurance
- state/province of living (where available), living arrangement, sense of community belonging, urban/rural
- immigration status, length of time since immigration
- BMI/obesity, physical activity, diet, other lab-related measurements
- alcohol use, smoking, substance use, cannabis use, diet
- unmet healthcare needs/access to doctors/healthcare utilization, general health status
- cycle/year

14. It is a very good idea to draw directed acyclic graphs (DAG) to think clearly about your analytic plan. This will also help you think clearly about potential effect modifiers, colliders, and mediators.
15. Often you will see multiple variables are related to the variable that you are looking for. In that case, it is a good idea to think about which variable will actually serve you better. Sometimes the other related variables can be useful for sensitivity analyses, and hence may be worth saving.
16. Consider eligibility criteria carefully (inclusion and exclusion) to create the analytic data based on your chosen research question. Usually a flow diagram is helpful in keeping track of who is included vs excluded and that helps make the generalization later.
17. Once you have a basic idea about exposure, outcome, relevant covariates, and weight/survey variables, quickly create a **complete-case analytic data** removing the invalid responses ('don't know', 'not applicable', 'refused', etc.). Check the sample size of this analytic data.
18. Save the software codes (e.g., R) that were used to create this analytic data, and document them properly (e.g., with comments of what was done in each stage), so that you can modify the **analytic data** easily if later you wanted to keep the invalid responses (e.g., to impute them through a **missing data analysis**) or add more variables in the analytic data. In most scenarios, complete case analyses are not justifiable for the final version of the project, and hence are highly discouraged.
19. Before finalizing your aim for the class project, create a *Table 1*; run a couple of **bivariate analyses on the created analytic data** (e.g., cross-tabulation of outcome-exposure; as well as outcome and some of the most important covariates. Also, run adjusted and unadjusted regressions to see if the categorizations need changing/merging). This will help you understand if there are any sample size issues in your data.
20. Choosing a topic of **predictive modelling is not allowed**. The model has to include one primary exposure variable of interest.
21. If one of the variables you are considering is coming from an **optional component** (particularly when choosing CCHS), be extremely cautious about the sample size. If, your outcome or exposure variable is coming from an optional component (which is highly discouraged), make a cross-table of outcome-exposure variables and see if the resulting sample sizes are adequate for your analysis.
22. Once SAP is submitted and approved, students are strongly encouraged to follow the aim/research topic outlined in that SAP. Not getting statistical significance for the exposure variable can't be a reason to change the research topic.
23. Just so that the expectations are clear: the instructor/TAs will not be helping any students with how to analyze their specific dataset (or exact R coding of the analysis, or data manipulation) that is part of their "final project" (for which the students will be graded based on their ability to implement/interpret methods that were taught in the class/lab). The instructor/TA can, however, help students with general advice on what general directions they can take to solve the problem.

24. The general expectation for the final project is that the students will be able to incorporate some of the advanced methods taught in the class/labs (e.g., missing data, propensity score, machine learning - whichever are appropriate/helpful for the project), preferably as sensitivity analyses. The labs/assignments so far are carefully designed such that students can understand the basics of the implementation procedure, but it is also expected that they can expand based on their needs/comfortability with the analysis/their ability to interpret in their final project's dataset.