**Statistical Critique of a cross-sectional study**

*Que, J., Shi, L., Deng, J., Liu, J., Zhang, L., Wu, S., Gong, Y., Huang, W., Yuan, K., Yan, W., Sun, Y., Ran, M., Bao, Y., & Lu, L. (2020). Psychological impact of the COVID-19 pandemic on healthcare workers: a cross-sectional study in China. General psychiatry, 33(3), e100259. https://doi.org/10.1136/gpsych-2020-100259*

Instructions: Please *briefly* respond to each question below (1-2 sentence responses). If you have any questions, please contact the course instructor well before the deadline. Group work among peers is acceptable (and encouraged) but not required. However, the individual student is responsible for submitting their own work. For purposes of this assignment, please follow this outline exactly. Your submissions must remain in outline form. This assignment is worth 10 points. Each question has a point value identified in parentheses.

**Sampling**: When investigators use a *probability* sampling scheme, they can increase the generalizability of their findings (assuming internal validity). However, designing and analyzing a study with a *probability* sampling scheme requires more funding, expertise and time. This barrier often results in the use of *non-probability* sampling schemes. However, just because a non-probability scheme is used, it does not automatically make a study ungeneralizable but it may limit generalizability to participants meeting eligibility criteria.

The authors of this study sampled using a *non-probability technique* (via the investigators social networks on WeChat). This could be a considered a *snowball sampling* technique (where one respondent is linked to subsequent respondents). We did not discuss this in-depth during class, however when you have a correlation in your sampling units (one participant is somehow linked to 1 or more additional participants) the analysis needs to take this into consideration. Unfortunately, the investigators did not recognize this and therefore their analysis did not account for this.

1. **(1 pt) If you were to re-design this study, what is a simple improvement you could make to the non-probability sampling scheme that would remove this snowball sampling?**

**Measurement of Variables**: In epidemiology, we discussed the importance of measurement. The underlying definitions are important if we are to estimate a concept accurately. In statistics, we take this measurement and attempt to make inferences from it. However, the choice of our inferential statistics is based on the scale of measurement (e.g. NOIR). For each of the below variables, please describe a) the source definition of the variable, b) how they are measured in the survey (e.g. categorical vs continuous), c) how are they analyzed (e.g. categorical vs continuous) and d) are they analyzed as an outcome or predictor (exposure)? Two examples are below:

Depression:

1. Depression was measured using the PHQ-9 which evaluates depressive symptoms within the last 2 weeks.
2. Depression was initially measured as a continuous variable in the survey. Specifically this continuous variable would be a ratio scale (it has the ability to be zero)
3. Depression analyzed as a categorical variable by dichotomizing at (<10: not depressed vs >=10: depressed). Specifically, this would be on the nominal scale and more importantly, this is a binary categorical variable (this has implications for your choice of statistical test).
4. Depression was analyzed as an outcome variable.

Regular exercise:

1. Exercise was first asked as: “Did you engage in physical exercise for 150min or more per week?”
2. This was first measured as a categorical (yes/no) variable. This would also be considered nominal or binary.
3. Since this was only measured as a categorical (binary, yes/no) variable, it cannot be reduced any further.
4. This was analyzed as a predictor/exposure variable.
5. **(1 pt) Attention to pandemic information:**

Understanding how variables are measured will help you understand the statistical hypothesis test and procedure. For example, from the above description, we can define the null and alternative hypotheses for the comparison between regular exercise and depression.

* The null hypothesis is that there is **no** association between exercise and depression.
* The alternative hypothesis is that there **is** an association between exercise and depression.
* Since the test is **two-sided**, they do not pose a direction (e.g. exercise could be associated with more or less depression). Note that the analysis plan does not specify one vs. two sided, but this is the default when it is not specified.
* Since the outcome is binary (depression yes/no) and the “exposure” is binary (regular exercise yes/no) we can choose from a variety of statistical tests: **chi-square, Fisher exact or logistic regression**. Review your handout in Moodle for “Choosing a Statistical Test” for more information.

**Research Data**: While you will unlikely ever recreate this study, it is reasonable that you will need to create a research database. Below is an empty dataset. Using guidelines on database development from class, fill this out with fake data for 5 pretend respondents. The data elements you will collect are: age, depression and exercise but all 5 columns will need to be filled out (HINTS: what is a primary key? What source data do you need to collect to classify someone as depressed?).

1. **(1 pt) What will the structure of this dataset look like?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. **(1 pt) Develop a codebook for the 5 variables in the above dataset.**

**Analysis & Presentation of Results**: Analysis plans should be aligned with the study objectives. From the introduction, the authors state: “This cross-sectional study investigated the prevalence of psychological problems in different healthcare workers during the COVID-19 pandemic in China and explored the demographics and COVID-19–related and work-related factors that are associated with various psychological problems.”

Given this study goal, the first step should be a *descriptive analysis* of each of these pieces: 1) types of healthcare workers, 2) demographic factors, 3) COVID-19 related factors, 4) work-related factors and 5) psychological outcomes. Some of these are measured categorically and some are continuous.

1. **(1 pt) What descriptive statistics would be used for categorical vs continuous variables? (HINT: There should be 3 ways you can summarize this data. One for categorical data and two options for continuous data depending on normality of the distribution).**

In Table 1, we can see descriptive epidemiology at work in regards to: 1) types of healthcare workers and 2) demographic factors. In Table 2, we can see descriptive epidemiology related to 5) psychological outcomes. Unfortunately, nowhere do we see descriptive epidemiology for: 3) COVID-19 related factors, 4) work-related factors. This is a failing of their analytic approach.

The second step you would expect to see is how these variables are related to each other. For example, to assess the prevalence of psychological problems in different healthcare workers, you would need to cross-tabulate these two variables (e.g. a 6 by 3 table summarizing the n(%) for each type of healthcare workers and level of anxiety). This is presented for each psychological problem in Table 2 (see anxiety example below).



From this table, you can see that nurses and technicians had a similar prevalence of moderate-severe anxiety (~15% for each) and residents had the lowest (~9%). You can also see the difference in the sample sizes for each provider group which may influence your interpretation.

The authors state that they “explored the demographics and COVID-19–related and work-related factors that are associated with various psychological problems.” This is presented in Table 3.

1. **(1 pt) In table 3, what statistical method did they use to estimate how these are associated? What is another statistical technique we discussed in class they could have used evaluate these types of variables?**

In Table 3 they report OR’s and 95% CI’s for each demographic factor, COVID factor and work factor. If they added the n(%) for each characteristic/outcome relationship, your understanding of these associations would be improved. The OR’s provided in this table only give you a *relative measure of association*, whereas having the n(%) would give you the *absolute measure of association*.

1. **(0.50 pt) Explain why understanding the difference between relative and absolute measures of effect is important.**

A common third step after unadjusted (or “crude”) logistic regression is to use a *multivariable* approach. There are two different reasons to do a multivariable model: a) evaluating associations between two or more variables (often causal associations…is x associated to y) in which they want to adjust for confounding or b) prediction of an outcome. Although not a true prediction model, this study is more closely aligned with the prediction goal. My assumption is that the goal of this multivariable model was to see what factors (demographic, COVID or work) were the most strongly associated with each outcome (more on this later).

They state that they used multiple backward logistic regression to do this. Just so you are aware: multivariable logistic regression, multivariable model, adjusted logistic regression are all similar ways to describe the same statistical procedure. They included variables into the multivariable approach if there was a p-value <0.05 in the unadjusted approach. Interestingly they do not present p-values but the analysis plan hinges on p<0.05. Instead, they put an asterix (or 2 or 3) next to a variable that had a p-value <0.05.

1. **(1 pt) How can you tell (other than the asterix) if an association is significant at the 0.05 level.**

**The Problem with Statistical Significance**: The literature is pretty clear that dichotomous thinking (e.g. significant (p<0.05) vs not significant (p>0.05)) is generally not a good idea and making modelling choices based on this is not best practice. This is at the center of many discussions currently going on in the statistical community. Experienced statisticians could debate several more appropriate and defensible ways to build a multivariable model with this data, however that is not the focus of this class. What I do hope is that you will gain enough fundamental understanding of statistics for some of these red flags to pop up (in this case a poor choice of multivariable model building – or just knowing that selecting variables purely on statistical significance is usually not a good choice). To illustrate this, let’s consider a few scenarios below:

* Most studies set alpha at 0.05 (e.g. setting the Type I error rate to 5%). Since there are (14 x 4) 56 hypothesis tests that are being done in Table 3 (if I counted right!), we would expect 5% of these (2-3 hypothesis tests) to be significant just by chance. Since approximately 23 variables are significant (have 1, 2 or 3 asterixis next to the variable), we run into the problem of determining which ones are truly significant. We have no way of determining which are true positives, false positives (type I error), true negatives or false negatives (type II error). We would need a much larger (and representative) sample as well as repeated research studies showing similar effects. The guidance from the Hill Criteria would need to be applied to begin thinking about true relationships.
* Current guidelines suggest that if you choose to present a p-value, you should present the actual value (not the threshold or an asterix) so the reader can interpret this piece of information along with implications of the study design, biologic plausibility, bias or other relevant information. In other words, p-values are just one piece of information, not the only piece of information used to make a decision. For example, if they had provided actual p-values and the association between exercise and anxiety was p=0.06 many researchers would likely have advocated to include it in their multivariable model as it is pretty common knowledge that regular exercise is beneficial for managing psychological symptoms.

1. **(1 pt) What if the authors tested handedness and found that left-handed people were statistically more likely (p=0.049) to have psychological problems. Would you include this in the multivariable model? Why?**

* As mentioned above, regular exercise is nearly always good for psychological symptoms. Interestingly, the authors only included this in their model for depression and overall psychological problems. Using this background information about the biologic association between exercise and the other outcomes (anxiety and insomnia) and interpreting the 95% CI, we can see that these effects, although not significant, were just over the null threshold (upper 95% CI for anxiety was 1.09 and insomnia was 1.10).

1. **(1 pt) Using Table 3, what is another variable that is even closer to this significance threshold, but is not included in the depression model? (HINT: it is included in every other model).**

* Instead of a more thoughtful approach to their model, the authors chose “significant” variables and ended up with a model suggesting that “attention to negative pandemic information” was one of the strongest factors associated with all of the outcomes. In Epi, we discussed that this variable is poorly defined and there are issues with temporality. Regardless, the authors conclude that “greater risk of psychological problems may be associated with receiving negative information about the pandemic.” If we reflect on the implications of this for the entire multivariable model, the authors put in a poorly conceived variable (and this variable is strongly associated with the outcome). This becomes a problem!! The reason is that each variable in this model is trying to explain the outcome. When we put in bogus variables, this can pull explanatory power away from other variables so we are left with an uninterpretable mess. Table 4 ends up becoming useless.

**A Note on Analytic Methods**: Over the course of your career you will read studies and see analytic methods or statistics that are unfamiliar to you. You will need to decide how much you need to learn to interpret this new statistic. For example, this study uses the R2 statistic in Table 3, which we did not cover in class. If you search Wikipedia for “logistic regression” and then search that page for “r2” you will see that this is considered a goodness of fit measure (how closely the predictors explain the outcome) but is used in linear (not logistic) regression. Scroll a little lower and you will see “pseudo R2” which elaborates on alternatives for logistic regression. Knowing that a good analysis plan will provide information on methods used, it is concerning that the analysis plan does not refer to how they are using/interpreting the R2 value. Since there are so many types of R2 statistics, they also should specify which one they used. Given this, we do not know why the authors included it in Table 3 and since our quick search suggested that R2 provides very little added meaning to this table (and logistic regression in general) the authors have not defended their use of it. Red flag!

**Power and Sample Size Considerations**: No *a priori* estimation of sample size or statistical power was conducted. (*In this case,* “*a priori” means that the power calculation was conducted prior to the start of the study to determine the sample size and/or magnitude of the difference that they could detect).* However, they ended up with a sample size of 2285. In these situations, power calculations are not always done (although many argue that they should be!). As an alternative, some advocate for a sample size that is large enough to provide reasonably precise estimates (narrow CI’s).

As an example of the impact sample size has on results, in Table 3 the OR for the association between drinking and anxiety is **1.58; 95% CI=0.89 to 2.79**. If you compare this to the OR for neutral type of pandemic information is **1.54; 95% CI=1.18 to 2.01**. Since both of these OR’s are nearly the same, some may wonder why one is significant and one isn’t. This is an example of why descriptive epidemiology is important and adding n% provides you with more information to interpret. From these estimates, can assume that the *cell sizes are smaller* for the drinking group compared to the neutral type group because the *CI for this drinking association is wider* than the CI for neutral type of pandemic information. If the cell sizes are smaller, this is likely an issue related to statistical power.

1. **(0.50 pt) Reviewing table 3, do you feel that the CI’s are reasonably precise? Do you think statistical power a concern? Justify your answer.**

**Putting it all together**: Last semester we discussed extensively the internal validity of this study and concluded that there were a lot of major issues with the study design in relation to information bias, selection bias and confounding. While the specific associations estimated are likely not accurate due to bias, this semester we focused on the statistics. We reviewed above why the multivariable model was useless.

Regardless, the estimates of the prevalence of psychological problems may be in the general neighborhood of what truly existed at the time of the survey. From a statistical perspective, I would have liked to see confidence intervals on these estimates of prevalent psychological problems. This would better describe the variability of psychological problems and a range of plausible values consistent with the data.

Overall, the only thing I feel like I can conclude is that psychological problems were prevalent in China around February 2020 with approximately 20% of the *respondents* reporting at least one condition (anxiety, depression, insomnia). I am still concerned about the sampling they used and who these would represent. Regardless, this could be used to generate hypotheses and inform future work.