Critical appraisal of a protocol for a cohort study on the incidence of psychosis on (condition redacted) patients

DOCUMENT: SAR-2023-009-LM-v01

From: Felipe Figueiredo **To:** Undisclosed MD researcher (US)

2023-02-09

TABLE OF CONTENTS

Critical appraisal of a protocol for a cohort study on the incidence of psychosis on (condition redacted) patients

Document version

Version	Alterations
01	Initial version

1 ABBREVIATIONS

- MOCA: Montreal Cognitive Assessment Test
- HR: hazards ratio
- OR: odds ratio
- RR: risks ratio

2 CONTEXT

This technical report reviews and enhances the protocol for a clinical research study being prepared for submission to a research grant. The study is well designed and protocol is well written.

2.1 Objectives

- To review the study design and statistical methods of the protocol;
- To review and provide suggestions to the training plan of the researcher on statistical and machine learning methods.

3 EXECUTIVE SUMMARY

- Comments and reviews were provided to enhance the writing of the study protocol;
- Minor adjustments are recommended to the study design;
- Comments and reviews were provided on the methods planned for each research question;
- An example of re-writing the statistical methods as a single paragraph is offered;
- Suggestions for topics are provided to complement the training plan mapped.

4 REVIEW

4.1 Methods of the project

4.1.1 Overall structure of the document

The protocol is well written, but the effort is not helped by the complexity of the subject matter involved. There are two distinct research questions under investigation that are treated in similar ways, but the structure does not help to ascertain this fact leaving the reader with the added effort of scrolling the document to compare how the approaches for each question converge and how they differ.

It can be recommended that the two research questions are restated at the start of the methods section noting what methodological choices will be used for both of them, and only then describing how the methods differ in order to accommodate specific requirements of each question.

Another observation can be made on the absence of clearly defined endpoints. It can be presumed from the protocol that the occurrence of psychosis would be considered an endpoint, thus removing study participants that reach this event from the cohort. Other possibilities that could be considered would be death, loss to follow up or migration out of the study area.

4.1.2 Study design

4.1.2.1 Outcomes and exposures

```
(redacted paragraph from protocol)
```

The primary outcome of the study is the occurrence of new psychosis diagnosis. The main predictor under consideration for this outcome is the number of illusory responses and this independent variable will be used as an exposure when statistical modeling is considered at the analysis phase.

```
(redacted paragraph from protocol)
```

There is a second analysis in which the main predictor for this outcome is the plasma concentration of (biomarker redacted). Since a multivariate analysis is planned for this project it stands to reason that these two predictors were planned to be evaluated in the same analysis. Furthermore, having a single and consistent methodology simplifies the methods of the protocol and facilitates understanding by the readership.

This is further discussed in section 3.1.3.1 and an example is included for completeness.

```
(redacted paragraph from protocol)
```

The secondary outcome is the score from the MOCA test. There is no analysis planned to evaluate this outcome in the protocol.

Either an analysis method is missing, or if one is not planned to be performed then the outcome should be removed from the protocol. An option is offered in section 3.1.3.1.

4.1.2.2 Participant recruitment design

```
(redacted paragraph from protocol)
```

This is a complex design that in some sense will increase precision on estimates for the first cohort that will be tracked for the longest period, while also producing less precision on the estimates for the last cohort to be recruited.

It makes sense to collect more data on the first cohort, as it will enable a sub-group or sensitivity analysis that could further investigate the effect on the full study population. Such sensitivity analysis could compare the estimates of the more detailed information from the first cohort with the best population estimate from the full data set. It is, however, unclear how the added complexity of collecting regressively less data from the newer cohorts will benefit the predictions from the study, compared to a more straightforward design that monitors all cohorts for a single year.

It is recommended that one sentence is added to the appropriate section to justify why this design was chosen.

4.1.3 Statistical analysis methods

4.1.3.1 Single statistical methodology

Both paragraphs of “Experimental design and statistical analysis” contribute similar information to the reader, and it would reasonable to consider merging them into a single paragraph.

This simplification could in turn be extended to the actual methods: since the logistic regression is being used as the main analytic tool of exploring the relationship between variables in the data, one multivariate model could be developed to answer each question, per outcome of interest. The psychosis outcome could be investigated with the logistic regression while the MOCA test could be evaluated with the linear regression.

In the interest in keeping the original plan, each of the models described above could be accompanied by univariate versions with only the main predictor (the image test). This simplified version would result in a crude estimate of the outcome, and this approach is often used in the medical literature to serve as a benchmark for interpreting the contribution of other covariates in the observed effect.

What follows is an example of how such paragraph could be written, based on the existing text.

We will determine whether illusory responses on the NPT performed 1-3 months pre-DBS and elevated plasma (biomarker redacted) collected in the same period predict early-onset post-op psychosis (i.e. day of surgery) and late-onset post-op psychosis (i.e. 6 months and 1 year, then annual) (defined above, Psychosis outcome measure). We will estimate the effect of the number of illusory responses and the (biomarker redacted) level on psychosis using logistic regression, adjusting for age, sex, disease duration and enrollment year (Year 1, 2 and 3) as a categorical co-variate. We will also estimate the effect of the number of illusory responses on the MOCA test using linear regression, using the same approach and covariates of the psychosis model. In each case we will also conduct simpler, univariate models with only the number of illusory responses as a crude estimate of the overall effect, to aid in the interpretation of the full model. If loss to follow-up and drop outs are sufficiently large so as to reduce study power, we plan to use forward-backward step-wise selection to reduce model complexity by only including significant terms, using p<0.20 as the threshold for inclusion. Both the NPT and (biomarker redacted) will be kept in the step-wise process as fixed parameters in the model selection. In the presence of a sufficient number of observed psychoses, we also plan to assess whether the NPT and (biomarker redacted) interact with each other. Furthermore, we will repeat the same analysis with respect to the association with early onset or late onset of post-op psychosis. The predictive accuracy of the logistic model will be evaluated using the receiving operating characteristics (ROC) curve and the area under the ROC curve. The sensitivity and specificity for selected cutoff value will be estimated, and their exact 95% confidence interval will be constructed. We will also calculate the area under the ROC curve (AUC) in subgroup of patients according to their enrollment year separately.

The paragraph above differs from the two original paragraphs in four main ways:

1. it removes the significance tests (Fisher's test in the case of NPT and t-test/Mann-Whitney in the case of (biomarker redacted)) since they provide redundant information already discoverable by the regression models chosen;
2. It adds a second, simpler model to aid interpretability of the results from the main model;
3. It adds an analysis approach for the second outcome, the MOCA test; while still approaching both questions with the same consistent methodology;
4. It offers a contingency to be used if study power turns out to be insufficient for the desired full model.

An additional observation can be made that the paragraph states that 6 months post-op is considered late onset psychosis but there is no mention on how the early onset will be treated. It is reasonable to presume this will add more levels to the categorical variable measuring time under observation (Year, was classified as 1 to 3, but does not include neither the 6 months mark nor the early onset).

Finally it must be noted that the modeling strategy provided above voids the sample calculation described in the protocol. This arises from the fact that this calculation was performed based on the significance tests planned to be done prior to the regression models. See the next section for a discussion on this topic.

This consultancy often adopts this approach of using the crude estimate as a reference for interpretation of the fuller model, with great success in readability and interpretability of the analyses. Some examples can be seen in the reports **SAR-2022-031-AH** and **SAR-2022-023-AD**, both using linear regression to estimate changes in continuous score outcomes and **SAR-2023-004-BH** that uses the Cox regression to calculate incidence.

4.1.3.2 Sample size calculation

```
(redacted paragraph from protocol)
```

The sample size calculation should have been done for the outcome (the dependent variable “occurrence of new psychosis cases” which is a binary variable), and not the main predictor of interest or any other independent variable. One way to analyze the data would be to calculate the Risk Ratio or the Rate Ratio, if a univariate analysis is considered. More sophisticated analyses methods include the logistic regression if a multivariate analysis is desired (see below).

It is recommended that a new power calculation is performed on the dependent variable. If the sample size is determined to be 50 participants at each year, then the power calculation can be used to determine the maximum detectable effect size for that sample size at 80% power.

The modeling approach described in the previous section considers two exposures and four covariates, for a total of six variables before considering interaction terms. This makes the 150 sample size planned to be collected more than enough for the logistic regression, when the rule of thumb of at least 5–10 observations per variable is used. This rule of thumb could be used instead of an explicit sample size calculation.

4.1.3.3 Estimation of period prevalence

The choice for the logistic regression is adequate for the purpose of estimating the period prevalence of the outcome, and it estimates the odds-ratio of changing status from at-risk to psychosis when there is a large enough change in the values of the predictors.

On the other hand, the research question of the study seems to point to an interest in the incidence of new cases of psychosis. The OR and the RR are approximately close under some conditions, in which case one can be used as a proxy to the other.

An alternative would be if the researcher considered calculating the incidence directly with the RR or the HR instead of the OR. Within the regression modeling approach the RR can be modeled with the Poisson regression; the HR of the time-to-event can be modeled with the Cox regression.

4.1.3.4 Unnecessary categorization of a numeric variable

The number of illusory responses is planned to be analyzed as a categorical variable with levels 0, 1, 2, 3 and >3. The use of the “>3” cut-point is discouraged, as noted by Douglas Altman and co-authors in many occasions (Altman, 2005; Altman & Royston, 2006; Naggara et al, 2011).

4.2 Training plan

4.2.1 Statistical training

There is mention to a course named **Intermediate Biostatistics: Regression, Prediction, Survival Analysis**. Although the course syllabus is not available, from its name one can be presume that it covers (at least):

- Linear regression;
- Logistic regression;
- Cox regression;

and that each of the topics above include an introductory treatment of both residual analysis and performance metrics (like R-squared and pseudo R-squared and the c-statistic for predictions from the logistic regression).

If this is the case, it can be recommended to look up extra coursework that covers the Poisson regression. This model can be useful for estimating incidence measures such as the Risk Ratio and the Rate Ratio, both relevant for epidemiology and clinical research.

Other topics that would apply to most regression models include the analysis of two types of extreme values:

- Leverage;
- Influence.

An introductory treatment of these topics would include the C-hat statistic and Cook's distance.

Additionally, since the goal seems to include obtaining proficiency in "methods for analyzing longitudinal data", the researcher may consider auditing a more advanced course that covers multilevel models (sometimes called hierarchical or mixed effects models), as these skills are important for dealing with such data.

Technical skills

No data analyst skill set will be complete without solid practical data manipulation techniques. When doing the Introduction to R, the researcher will likely be introduced to techniques on how to treat dataset and prepare them for analysis. It is strongly recommended that some focus is employed in learning the **tidyverse** way of doing data processing, which differs from the **base R** way of doing these operations.

For report writing, the researcher might benefit from learning about **RMarkdown** files, which uses enables embedding and formatting of results directly from the R environment to a document.

4.2.2 Machine learning training

There is mention to a course named **Machine Learning**. Although the course syllabus is not available, from the text ("*classification/clustering/forecast*") one can be presume that it covers (at least) an overview of the main algorithms of the field, such as:

- Supervised machine learning
  - Regression algorithms
  - Classification algorithms
- Unsupervised machine learning
  - Distance-based clustering algorithms (e.g. k-means)
  - Hierarchical clustering

If Decision Trees is not included, it is recommended that some time is spent in learning at least the basics of both decision trees and decision forests methods (for classification and for regression purposes). This family of methods benefit from a high level of interpretability of the results.

Cautionary note

It can be cautioned that this field is extensive and many techniques may appear seductive in problem solving at large. It is very easy to loose one-self in the rabbit hole

of machine learning, so it would be recommended to keep specific goals in mind to guide focused skill building.

Observation

In the Machine Learning realm the methods are named “regression” and “classification” based on the type of the variable being predicted (numeric or categorical, respec.). This differs from statistical terminology that defines the logistic regression model under the regression umbrella, regardless of the type response variable. Specifically the logistic regression is part of the Generalized Linear Model family of techniques, with the Poisson model as another example of such model.

4.2.3 Epidemiology training

There is no mention of intermediate Epidemiology coursework in the training plan.

It must be noted the difference between obtaining proficiency and independence in new tools. Independence can arguably be understood as expertise, at least to some degree, but although it can be achieved, it is not expected that a clinical researcher with medical background excels at either Statistics or Machine Learning beyond the proficiency level. Epidemiology, on the other hand, is another matter.

It would be advisable that such training be considered (and arguably prioritized) since in clinical research the real autonomy of the researcher arises not only from the

An intermediate to advanced course would aim for a comprehensive reading of the most recognized textbook in the field (Greenland, Rothman, Lash, 2008). An introductory treatment of the subject could cover the abridged version, written by the one of the authors (Rothman, 2012) and can serve as an entry point to the field before embarking on the main textbook.

5 REFERENCES

- Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. BMJ, 332(7549), 1080.1. (https://doi.org/10.1136/bmj.332.7549.1080)
- Altman, D.G. (2005). Categorizing Continuous Variables. In Encyclopedia of Biostatistics (eds P. Armitage and T. Colton). (https://doi.org/10.1002/0470011815.b2a10012)
- Naggara, O.; Raymond, J.; Guilbert, F.; Roy, D.; Weill, A.; Altman, D. G. (2011). Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms. American Journal of Neuroradiology, 32(3), 437–440. (https://doi.org/10.3174/ajnr.a2425)
- Greenland, S., Rothman, K. J., Lash, T. L. (2008). Modern Epidemiology. United Kingdom: Wolters Kluwer Health/Lippincott Williams & Wilkins.

- Rothman, K. J. (2012). Epidemiology: An Introduction. United Kingdom: OUP USA.
- **SAR-2022-031-AH** – Fatigue profiles under long term illnesses: prospective cohort study.
- **SAR-2022-023-AD** – Association between KOOS scores and OTC analgesic use in patients using knee-braces.
- **SAR-2023-004-BH** – Effect of socioeconomic status of neighborhoods in mortality mortality rates after brain injury: retrospective cohort.

6 APPENDIX

6.1 Availability

All documents from this consultation were included in the consultant's Portfolio.

The portfolio is available at:

https://philsf-biostat.github.io/SAR-2023-009-LM/