UNIVERSITY OF ILLINOIS CHICAGO

Tips for Managing and Analyzing Covid-19 Data

CCTS Bioinformatics and Biostatistics Cores

October 2020

The below guidance on working with Covid-19 data was developed by the CCTS Biostatistics and Bioinformatics Cores. For more information on this topic, including advice about how to apply it in your research, consider scheduling a consultation with CCTS.

Data Organization and Management

1. Data points need accurate timestamps.
2. Data should be recorded at its most precise measurement level, not grouped into ranges or otherwise "coarsened". When precise measurements are recorded, options for later grouping into ranges or other transformations are preserved. These steps are easily accomplished in database or statistical software later.
3. To the extent feasible standard variable definitions and units should be used. (OMOP Common Data Model Specifications Common Data model, COVID-19 Clinical Data Warehouse Data Dictionary, FDA, Vocabulary). [This needs a simple starting point.]
4. Example mock-up longitudinal data structure:

ID	COVID_STATUS	AGE	GENDER	drug_A	drug_B	date_measured	hopital_los	copd	Systolic	Diastolic
1	0	56	1	0	1	3/1/2020	6	1	120	67
1	0		1	0	1	3/2/2020			130	80
1	1		1	0	1	3/3/2020			110	
1	1		1	1	0	3/4/2020				79
1	1		1	1	0	3/5/2020			130	89
1	1		1	1	1	3/6/2020			120	78
2	0	47	0	1	1	3/24/2020	2	0	114	56
2	0		0	0	0	3/25/2020			115	67
3	1	70	1	1	0	3/26/2020	3	1	124	78
3	1		1	1	0	3/27/2020			124	79
3	1		1	1	0	3/28/2020			124	79

Figure 1: Example Data Structure

- Variable names (`ID`, `COVID_STATUS`, `AGE`, etc. in the above table) should be in the first row starting from letter or '_' (not a number) with no space within about 15 letters or shorter. Upper or lower cases will all be ok.
- Missing data should be in blank or in one unique code such as NA or 9999 that is never used as a real value anywhere within the dataset. The missing data code rule should be consistent or same across all variables in the dataset. 0 should be different from missing.

- Repeated measure database should have multiple rows per ID. One line/row per event (time point). If the variable is time variant, each time point's values should be put unless it is missing. Time invariant variables may have only one value at the first line per ID (or patient) and the following lines with blank. Or, the same value should be filled in for the time invariant variables.
- Example variables listed above:
  - `ID`: De-identified random number but unique for each patient. Remove MRN and other patient information that is not needed for the analysis.
  - `COVID_STATUS`:1=yes, 0=no
  - `AGE`: Continuous age in integer is preferred.
  - `GENDER`: 0, 1 representing either female or male for each with other categories if applicable. The definitions for numeric codes should be put in a separate list.
  - `drug_A`: Either 1 or 0 for 1=yes, 0=no, or put the dose.
  - `date_measured`: The date format should be consistent/same through the database such as MM/DD/YYYY.
  - `hopital_los`: Hospital Length of Stay. Integer values. If it is one value
  - `copd`: Related disease history. Status 1 or 0 for yes, no respectively.
  - `Systolic`: BP systolic. Continuous value in one number only.

Statistical Issues

1. Small sample sizes pose challenges to analysis. Epidemiological aspects of the disease such as means of infection and transmission are difficult to study in small samples. A small sample of patients treated in one or a few facilities is not representative of larger populations and thus general claims should be advanced cautiously. Small samples limit the statistical power to detect effects (trends, mean differences).
2. Time intervals between observations will often be irregular (as seen in timestamps), which means that additional care is needed when describing trends. (Running averages, smoothing and regression splines, orthogonal polynomials)
3. A series of observations on the same variable from the same patient are serially correlated. Longitudinal data analysis techniques that take those correlations into account are needed. (Linear mixed models, generalized LMM, GEE, etc.)
4. Many series of clinical records and laboratory assays will likely form the data set. A strategy to study the synchrony among series is needed. (joint trends, correlations)
5. Period effects should be considered in the analysis. That is, over the months the strain of Covid-19 virus may have changed, treatments may have changed, and so on. The time period during which a patient was treated may alter outcomes (early months vs. later months). (Refer to age-period-cohort analysis; likely emphasis on age and period.)
6. The number of patients in a data set is N. The number of variables included in an analysis is P. Covid-19 data sets may have large P relative to N (and P > N is a special problem). Strategies to reduce P may be needed, including variable selection and combining variables (principal components, factor analysis).
7. Comparisons between groups of patients must be carefully considered. Groups might be defined in terms of patient characteristics (age, race/ethnicity, gender, comorbidities), or in terms of features of the course of illness (e.g., slow vs. rapid progression of symptoms). The fact that two groups differ may not be causal. For example, racial and ethnic groups may have different living conditions, and the living conditions affect infection rates. Special statistical techniques are needed to attribute causality, especially in observational studies.
8. Multivariate analysis that considers multiple dependent variables at once may prove useful. Refer to principal components, factor analysis, canonical correlation, multivariate analysis of variance.

9. The analytical challenges mentioned above will often exceed typical experience of scientists and physicians. Consider consulting a professional statistician, such as those in the Biostatistics Core of the CCTS, to help define credible analytical strategies.

10. Some studies of medical interventions, such as administration of a particular drug, may need to examine time-to-response or duration of response. For other studies which are interested in the death risk of COVID-19 patients with severe symptoms, analysis for time-to-death data is crucial as well. In cases where COVID-19 patients had co-existing fatal diseases, such as cancer, competing risks would need to be addressed properly in the analysis. There are a number of biostatistical techniques, generally termed survival analysis, that focus on elapsed time and whether event of interest occurs or reoccurs as the outcome. These techniques are relatively technical and best employed in collaboration with a biostatistician. Application of such techniques depends on accurate timestamping of observations.