NCSU-Duke SIBS Data Hack-a-Thon 2022

Project Description

The NCSU-Duke SIBS Data Hack-a-Thon is a data analysis assignment where SIBS participants work in assigned teams to tackle a large, complex data set with the goal of addressing scientific research questions.

Background and Data Source

Acute myocardial infarction (MI), more commonly known as a heart attack, is a condition that occurs when one or more areas of the heart muscle does not get enough oxygen. This happens when blood flow to the heart is abruptly cut off, typically caused by a blockage in one or more of the coronary arteries. MI results in heart tissue necrosis and is often life-threatening. For an overview of the clinical definition and management of MI, please see White and Chew, 2008.

Treating and preventing MI is one of the most challenging problems of modern medicine, and it has been the focus on much medical research over the past 50 years. Acute myocardial infarction is associated with high mortality in the first year after it. The incidence of MI remains high in all countries. This is especially true for urban populations of highly developed countries, which are often exposed to chronic stress factors, irregular and not always balanced nutrition. In the United States, for example, more than a million people suffer from MI every year, and between 200,000 and 300,000 of them die from acute MI before arriving at the hospital. The disease course in patients with MI is different. MI can occur without complications or with complications that do not worsen the long-term prognosis. At the same time, about half of patients in the acute and subacute periods have complications that lead to worsening of the disease and even death. Even an experienced specialist cannot always foresee the development of these complications. In this regard, predicting complications of MI so necessary preventive measures can be carried out in a timely manner is an important task (Golovenkin, et al; 2020).

The data you will working for this project is a data base developed by Golovenkin, et al (2020) to provide researchers access to an MI dataset with real-life complexity. The database contains data collected on MI complications from patients admitted to admitted to Krasnoyarsk Interdistrict Clinical Hospital №20 for acute MI in Russia between 1992 and 1995.

The database contains the records for 1700 patients. Each record contains 111 covariates that contain the patient's demographic information, medical history, and clinical features of their MI. Most covariates were measured at time of hospital admission; however, 3 were measured during the first day of hospitalization, 3 on the second day of hospitalization, and 3 on the third day of hospitalization. Each record contains 12 outcome variables, complications of MI. The reference literature does not state when the outcome variables were measured, but for the purposes of this project you may assume that all outcomes were measure after the third day of each patient's hospital admission. The database contains 7.6% of missing values.

You can download the data from here:
https://leicester.figshare.com/articles/dataset/Myocardial_infarction_complications_Database/12045261?file=23581310

You can download a data dictionary here:
https://leicester.figshare.com/articles/dataset/Myocardial_infarction_complications_Database/12045261?file=22803572

You can download a data summary here:
https://leicester.figshare.com/articles/dataset/Myocardial_infarction_complications_Database/12045261?file=22803695

Once you have downloaded the data, you can use the following commands to import it into R:

```
mi_comp = read.csv('<Your File Path!>/Myocardial infarction
complications Database.csv')
```

Once you have downloaded the data, you can use the following commands to import it into SAS:

```
proc import file="<Your File Path!>/Myocardial infarction
complications Database.csv"
    out=work.mi_comp
    dbms=csv;
    run;
```

Research Questions

Your primary goal is to identify risk factors for one of the following MI complications: atrial fibrillation, pulmonary edema, or relapse of the MI. That is, you need to determine which covariates have statistically significant effects on the occurrence of these outcomes. Here are some questions you may like to answer:

1. Which covariates measured at hospital admission are correlated with a complication (i.e., which covariates may be useful for clinicians to monitor when treating MI patients)? Which are associated with an increased risk of complication? Which are associated with a decreased risk of complication?
2. Are covariates measured during the hospital stay correlated with a complication? Do they provide information about the odds of a complication independent of covariates measured at the time of admission (i.e., should clinicians treating MI patients measure these covariates during the hospital stay)?
3. Can these covariates be used to build a model that predicts the odds of a complication?

Your team may consider a research questions of your own devising in addition to or instead of these. As is always the case with real world data, there are some complications will need to address and account for when completing your project. Here, you will need to decide how to handle the missing data values that occur for some variables.

Remember that you are working with real data, trying to find answers to real questions about the relationship between clinical factors and the risk of MI complications. The “correct answers” to these problems are complicated, varied, and for the most part not unanimous. Don’t limit yourselves by trying to find one right answer but rather provide data-analysis-based evidence in favor of your answer to the best of your abilities, admitting the limitations when necessary.

What Does Your Team Have to Do?

Each team is responsible for using statistical methods discussed during the SIBS program address at least one research question. It is up to each team decide which question(s) to address and which statistical models and associated methods are appropriate for addressing those question(s). Each team is also responsible for creating a 15-minute presentation describing the research question(s) they chose to address, their statistical approach, and the results of their team’s analysis.

Each team’s presentation should contain the following components:

- Background on the study and the research question(s) addressed in the analysis
- Descriptive analysis, including data visualizations, of the cohort being studied
- Inferential analysis, including a justification of why the selected statistical methods are appropriate for addressing the selected research questions(s)
- Discussion of the findings and their limitations based on the data source

Two presentations will be selected to win prizes in the following categories:

- Best Insight—highlighting the presentation that best communicates the story of the data, supported by sound statistical analyses
- Judges Choice, which could highlight a presentation with the best visualization, best surprise element, or best use of statistical modeling, for example

Resources

As mentioned above, you will likely want to use the R packages and SAS procs covered in the SIBS lectures and labs to help with your analyses, but there are many other techniques to analyze this data and answer the research question that we may not have covered. Feel free to research such techniques and make use of them. Further, you may want to reach out to the clinicians at DCRI to get expert opinions on important factors in the outcomes of acute MI, whether or not these factors appear in this dataset.

Finally, there are a few important resources available for your perusal on the Moodle page, under the heading “Project Information and Resources,” all of which can be found in the “2022 Hack-a-Thon” folder. First, you will find two presentations from the group of students who participated in SIBS in the Summer of 2019. Take a look at these two projects to get ideas about how you may—or may not—want to present your work at the end of the program. Another resource you will find in this folder is a document, titled “How to Get Started” that will help you begin data cleaning, exploratory analysis and some rudimentary inferential analysis as well. This is meant to be a guide for getting started on the Hack-a-Thon, using the data from 2019. The code will be easily generalizable to the data for this year’s project.

As always, you can also reach out to the SIBS mentors and professors for guidance and support as your team progresses through the project. The morning and afternoons session on Tuesday July 19 and the morning session on Wednesday July 20 will be dedicated to finalizing all analysis tasks and preparing the team presentation. During the afternoon session on Wednesday July 20 each team will practice their presentation in front of the SIBS mentors ahead of the final presentations on Thursday July 21.

References

Golovenkin SE, Gorban A, Mirkes E, Shulman VA, Rossiev DA, Shesternya PA, Nikulina SY, Orlova YV, Dorrer MG. (2020): Myocardial infarction complications Database. University of Leicester. Dataset. https://doi.org/10.25392/leicester.data.12045261.v3

White HD, Chew DP. Acute myocardial infarction. The Lancet. 2008 Aug 16;372(9638):570-84.