Final Report

Zihuan Qiao, Anna Cook, Yinfeng Zhou, Zixuan Liu, Jiaheng Li

3/25/2021

Introduction

Project Background

The client is researching rare disease advocacy organizations. In particular, she is interested in comparing organizations for rare vs. ultra-rare diseases along several different factors including funding, patient outreach, research, etc. in order to assess the organizations' need for support. The main research questions is, how do the prevalence of diseases (rare vs. ultra-rare) and the size and age of the organization, affect the outcomes of the advocacy organization? (priorities, funding, patient outreach, research, etc.)? For example, the client may expect to find that organizations for less rare diseases may have more resources, better patient outreach, etc. while more rare diseases may have fewer resources available and may need more support. In addition to this main research focus, the client is interested in exploring the data more generally and looking for patterns that may arise. The client is seeking advice from the MSSP team about how to effectively recode variables as necessary, conduct an initial exploratory data analysis, and determine an appropriate regression model in order to make the comparisons of interest.

Variable Description and Data Processing

The client has collected survey data from 217 different organizations' leaders or representatives, located in various locations worldwide, with one response per organization. The survey includes questions related to demographic data, budget/funding, disease prevalence, research efforts, etc.

There are three independent and three dependent variables which our analyses will focus on. The independent variables are organization size, organization age, and disease frequency (1 case per x births). We will refer to these as Size, Age, and Frequency, respecively. The dependent variables are organizations' top priority, whether research efforts are handled internally or externally, and whether clinical trials are in progress for a therapy/treatment or not. We will refer to these as Priority, Research, and FDA Therapy, respectively.

The first step in processing the data was to bucket the independent variables into discrete levels based on the client's literature review and judgments. Size has three levels: Small (0-300 members), Medium (300-1000 members), and Large (1000+). Age has two levels: Older (10+ years) and Younger ($<$ 10 years). Frequency has three levels: Rare (more frequent than 1 in 200,000 births), Ultra-Rare (less frequent than 1 in 200,000 births), and Unknown, although we are only focused on the Rare and Ultra-Rare levels for the sake of the analysis, so any responses of "Unknown" were removed.

EDA

In order to get a general sense of the dataset, we began with an exploratory data analysis (EDA). First, we plotted histograms of the independent variables (Size, Age, and Frequency). The histograms are shown in Figure 1. From these plots, we can see that there is some imbalance in the data. First, it appears that the organizations for ultra-rare diseases tend to be smaller and younger, whereas the rare disease organizations are more uniformly represented across the different size and age categories. Second, ...

Next, we plotted histograms of the dependent variables of interest (Priority, Research, and FDA Therapy). From these plots, we can see further imbalance in the data (see Figures 2-4). When looking at the priority

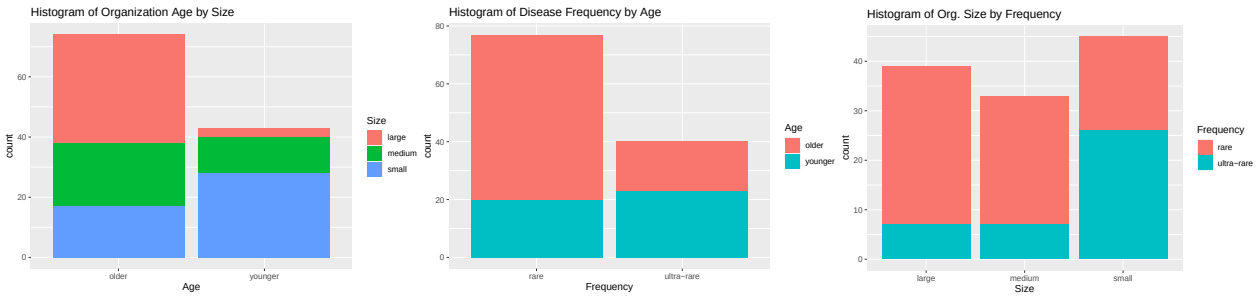

Figure 1: Histograms of each combination of independent variables.

variable (See Figure 4), we can see that many of the priorities are not equally represented by organizations in each of the frequency/age groups. The same holds true for the research and FDA therapy groups. This will be an important caveat to consider when interpreting results of the final analysis.

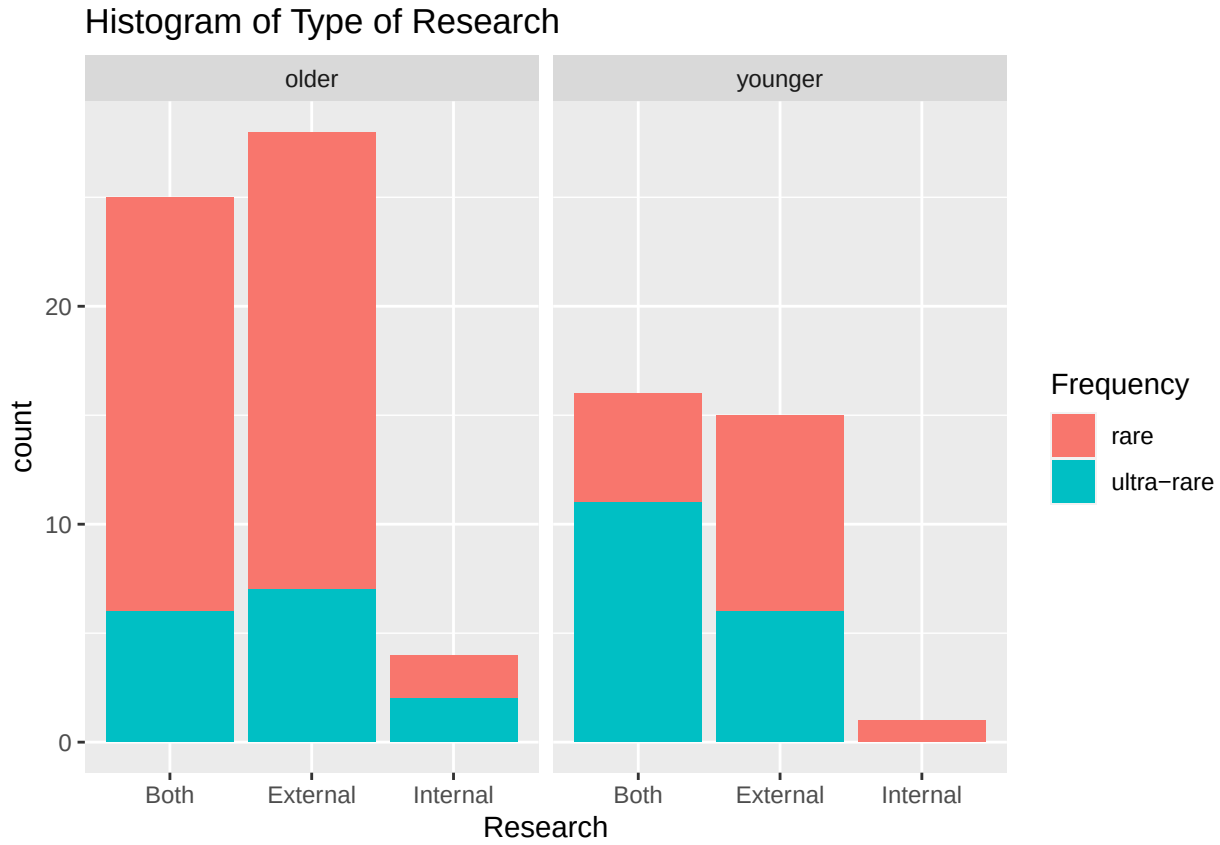

Figure 2: Histogram of internal research, external, or both, broken down by age and frequency.

As the last step in the EDA, we conducted a series of chi-squared tests to check for independence among the predictors of interest. The results of the tests are displayed in Figure ___. For all three pairs of predictors, the test shows p-values < 0.05, indicating that the variables are not independent at an alpha level of 0.05. Including correlated predictors in a regression model is problematic, so we must take this into account when fitting the models in the next step of the analysis.

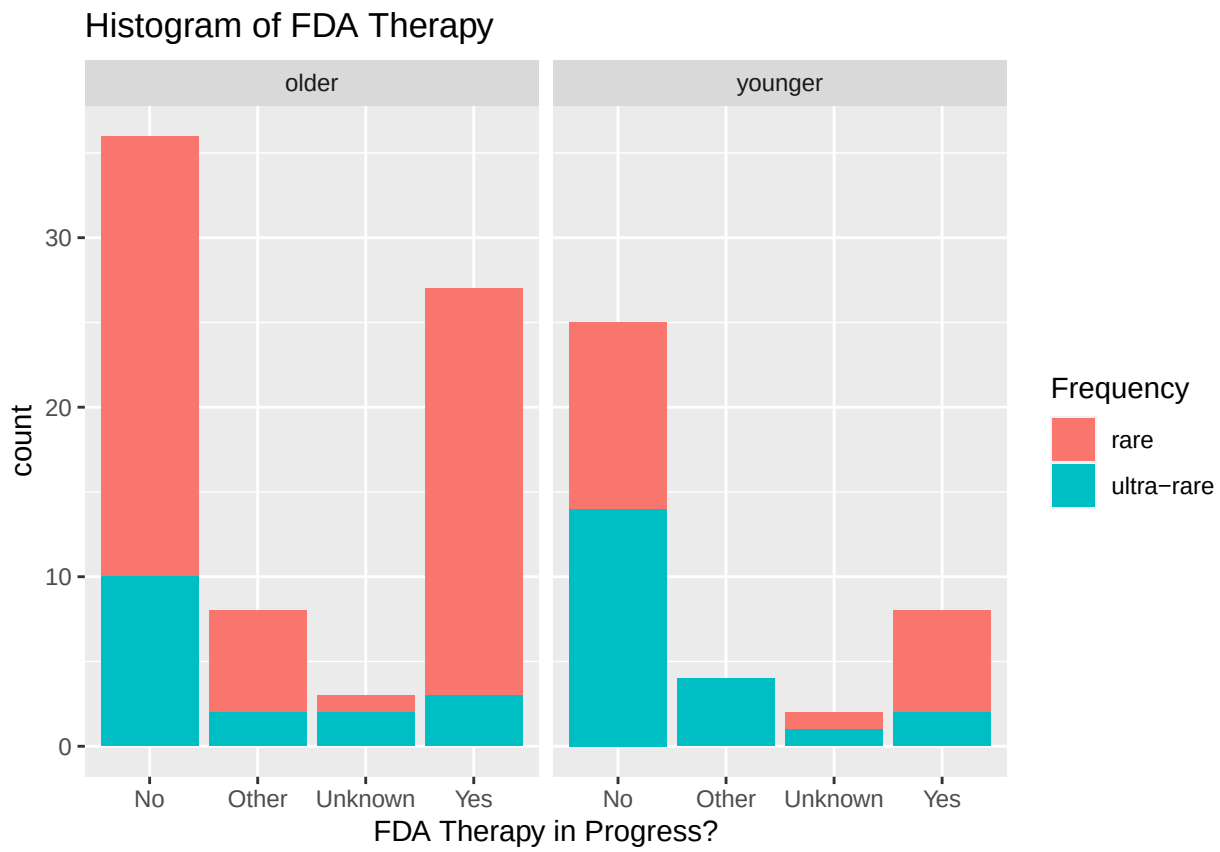

Figure 3: Histogram of whether FDA clinical trials are in progress or not, broken down by age and frequency.

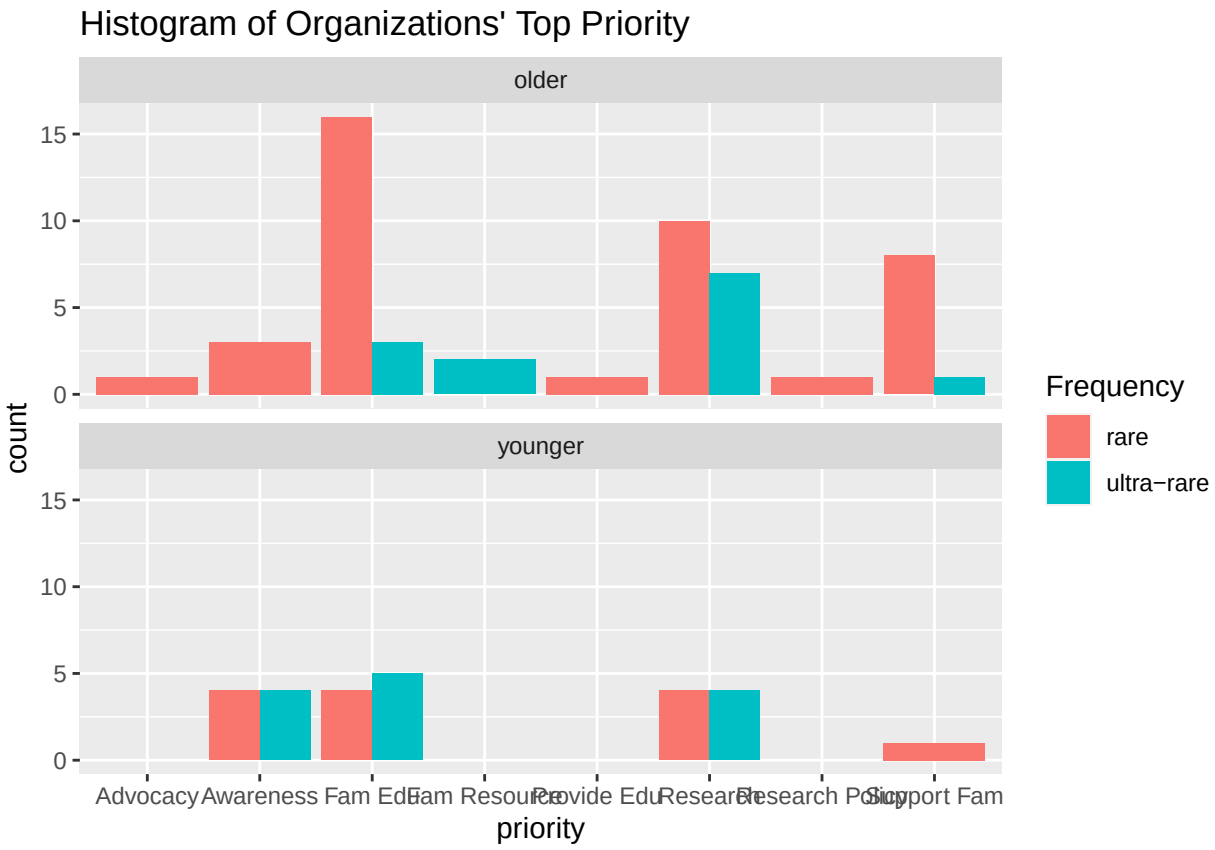

Figure 4: Histogram of organizations' top priorities, broken down by age and frequency.

Modeling

Because the variables of interest are categorical with more than two levels, the most appropriate regression model is a multinomial model. We fit a series of multinomial models, one for each response variable, using the "nnet" package in R. As mentioned previously, chi-squared tests showed correlation between the predictor variables, so instead of including more than one of these predictors in the model, we chose to

Top Priorities

Research Internal vs. External

FDA Therapy In Progress vs. Not In Progress

References

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Stefan Milton Bache and Hadley Wickham (2020). magrittr: A Forward-Pipe Operator for R. R package version 2.0.1. https://CRAN.R-project.org/package=magrittr

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Appendix