

What affects membership's assessment?

– Interpretable Models for Satisfaction with Different Aspects of HNC

Consultant: Ruiping Ke

Client: Charlene Cowell, Hemophilia of North Carolina's new Executive Director

Professor: Dr. Steve Marron, Dr. Perry Haaland and Dr. Vidyadhar Kulkarni

April 1st, 2021

Abstract

This report created statistical models and framework to find out and describe the characteristics of respondents who may not be satisfied with the service provided by HNC(Hemophilia of North Carolina), and then we can find out where could make the HNC program do better. Using Categorical Regression, Lasso Models (Variable Selection), Ordinal Regression and Kruskal-Wallis test, we identified several key factors that significantly impact the membership's satisfaction to improve and provide impactful services that meet the diverse needs of membership. Different bleeding disorders, races and ages of the survey participants are or the person they are connected to may cause their views on HNC and their satisfaction in various aspects to be different. We discover that whether the member is an Asian /Asian American(race) or not has a statistically significant impact on the overall satisfaction of HNC. If the survey participants are **not** Asian /Asian American, they are **more** likely to be satisfied with HNC overall. In the future work, HNC could consider how to take care of people in different situations to make them be more satisfied.

Executive Summary

An overview of the major findings with an indication of where details can be found in this report is as follows:

- Respondents who have another family member of someone with a **bleeding disorder**, who are **not** "Platelet Disorder" bleeding disorder and who have a **higher** agreement with the statement **HNC is their top source of support for living with a bleeding disorder** tend to think the event "**Educational Dinner**" is more important. (Part II(2) Q16_3 in Results)
- The type of **bleeding disorder** of the respondents are or the person they are connected to is affected by and how they feel have a statistically significant impact on the connection strength, recommendation willingness and satisfaction in all aspects related to HNC. (Part I in Results and Part II in Results)
- There are **not** significant differences between the reclassified bleeding disorders groups(group a: A, B, inhibitor (current or tolerized), group b: von Willebrand disease and group c:all others) in their feelings about connection with HNC, their

willingness to recommend to others, the degree of affirmation and overall satisfaction with HNC in all aspects.(Part III in Results)

- *Race and age have a statistically significant impact on respondents' evaluation for HNC, their connection to HNC and their willingness to recommend HNC.((Part IV in Results)*

Suggestions of future work are discussed in the Discussion/Future Directions section at the end of this report.

Background

Hemophilia is an inherited bleeding disorder in which a patient's blood doesn't clot normally because it lacks sufficient blood-clotting proteins (clotting factors). There are two major forms of the disease: hemophilia A (factor VIII deficiency) and hemophilia B (factor IX deficiency).

HNC(Hemophilia of North Carolina) is a nonprofit organization dedicated to improving the quality of life of people affected by bleeding disorders through advocacy, education, promotion of research, and delivery of supportive programs and services. Their vision is for all pelp affected by bleeding disorders to achieve their full potential without barriers or limitations. HNC strives to provide impactful programs and services that meet the diverse needs of the membership. Finding out what is helpful to the membership, what HNC does well, and what HNC can improve upon will help HNC continue growing the programs and services for membership as they plan for the future.

Data

Our dataset was from the Hemophilia of North Carolina (HNC) Needs Assessment. The dataset was collected in December, 2020, containing 116 observations and 101 variables. Respondents' satisfaction is reflected by the feedback about how connected they feel with HNC, how likely they would recommend HNC to friends/family/other people affected by a bleeding disorder, and how does HNC meet their needs.

There are 47 questions in the questionnaire in total. The questions we are interested in this report are Q2, Q4, Q16, Q17, Q18, Q19, Q20, Q24, Q25, Q32, Q33, Q34, Q36, and Q37, as listed in **Table 1**. For Question 2, everyone responded, so the processing method we used in this report for the missing value is to delete all the individuals that have not answered any of the multiple-choice questions except Q2. In this way, we kept 96 samples in the data. However, for the investigation of different questions, we also deleted few(1-2) observations when the response in the specific model is NA, meaning there is no response for the question. We deleted them because we cannot predict when there is no response.

| Questions List | |
|------------------|---|
| Questions Number | The Description of the Corresponding Question |
| Q2 | The connection to the bleeding disorders community |
| Q4 | The type of bleeding disorder |
| Q16 | How important the corresponding event is listed in Q16 |
| Q17 | How important the corresponding program is listed in Q17 |
| Q18 | How important the corresponding service is listed in Q18 |
| Q19 | How important the current needs of the bleeding disorders community is listed in Q19 |
| Q20 | What are the respondent' current needs? |
| Q24 | How connected do respondent feel with HNC? |
| Q25 | How likely would respondent be to recommend HNC to friends/family/other people affected by a bleeding disorder? |
| Q32 | Let us know how you feel about the following statements related to HNC? |
| Q33 | Overall, how does HNC meet your needs? |
| Q34 | Race/Ethnicity |
| Q36 | Gender |
| Q37 | Age range of the person taking the survey |

Table 1: The description of the corresponding Questions that we are interested in

For questions that express survey participants' satisfaction on the scale of 1-5, 1 means very important/very high/strongly agree/extremely connected/extremely likely/extremely well, while 5 means not at all. So, it indicates the smaller the number, the higher the level of satisfaction.

Results

There are four parts of the result:

(I) Predict Q24, Q25, Q32 and Q33 separately by using control variables Q34, Q36 and Q37

Because there are 10 predictors, we used Lasso Regression to do the variable selection first, and then we did the Ordinal Regression by using the selected predictors.

Taking the prediction of Q24 for an example, we used the parameter lambda which is equal to specify a grid of adjustment parameters, and lasso will output the results corresponding to these adjustment parameters.

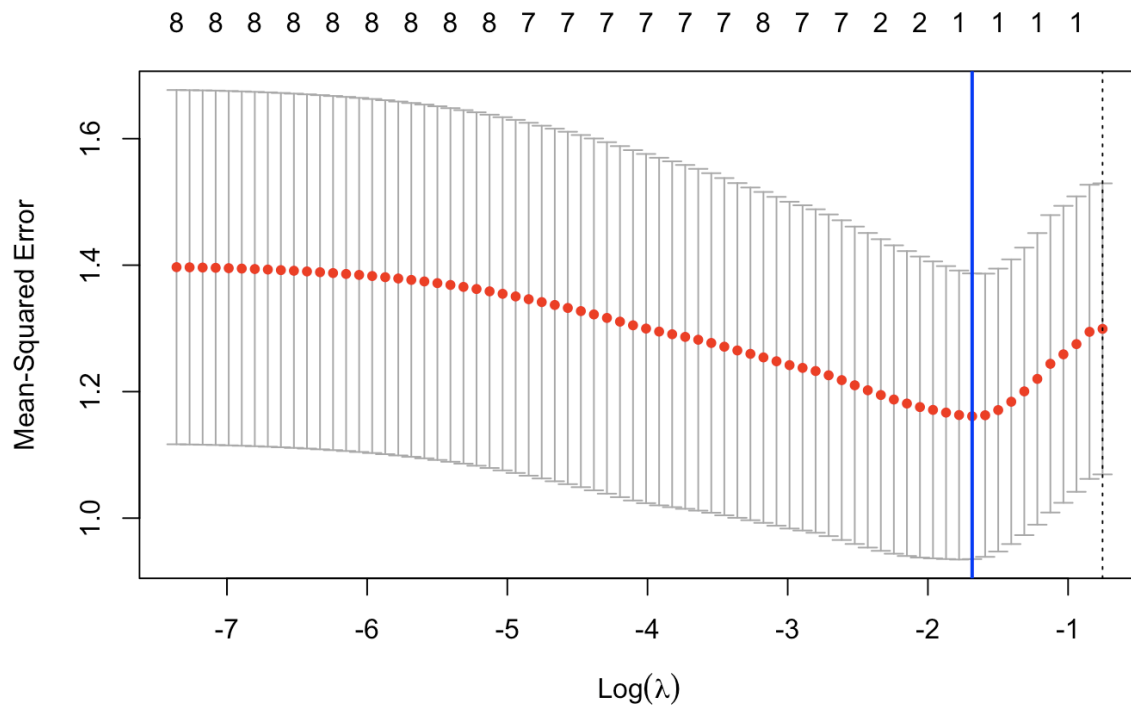


Figure 1: The way to find the optimal adjustment parameters λ

This graph depicts the optimization of the parameter λ and the blue vertical line shown in Figure 1 indicates the minimal mean-square error is achieved that $\log(\lambda)$ equals the values there.

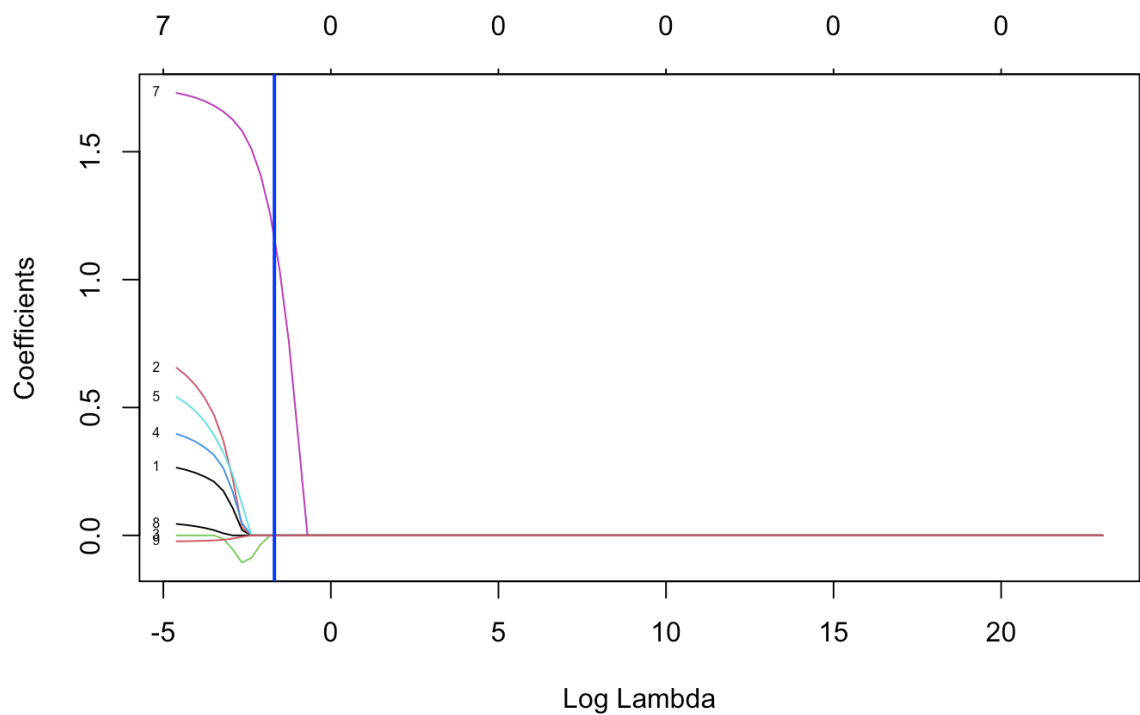


Figure 2: The change trajectory of coefficients with different parameters λ

In Figure 2, each curve represents the change trajectory of each independent variable coefficient. Same as shown in Figure 1, the blue vertical line shown in Figure 2 indicates the minimal multinomial deviation is achieved that $\log(\lambda)$ equals the values there. The number of the corresponding curve--1 means Black or African American, 2 means Asian or Asian American, 3 means White or Caucasian, 4 means Hispanic or Latino, 5 means American Indian or Alaska Native, 6 means Native Hawaiian or other Pacific Islander, 7 means Multiracial, 8 means gender, and 9 means age. We can see that the independent variable 7(Multiracial) represented by the purple line has a non-zero coefficient when the λ value is large, and then becomes larger as the λ value becomes smaller.

A dataset can be repeatedly split into a training dataset and a validation dataset: this is known as **cross-validation**. The **training set** is a dataset of examples used during the learning process and is used to fit the parameters (e.g., weights) of, for example, a classifier. The **testing set** is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place. A better fitting of the training dataset, as opposed to the test dataset, usually points to overfitting. We used the training set for cross-validation and the `cv.glmnet()` function is used to perform cross-validation. In this way, the optimal adjustment parameter 0.186 is obtained, as shown in Figure 1. By using the optimal adjustment parameters, we predicted the test set, and got the predicted mean square error: 1.533 .

| Regression Coefficient Table | | |
|---|----------|---------|
| Variable Names | Estimate | P-value |
| Black or African American | 0.24 | 0.429 |
| Asian or Asian American | 0.90 | 0.437 |
| Hispanic or Latino | 0.77 | 0.079 |
| American Indian or Alaska Native | 0.46 | 0.691 |
| Native Hawaiian or other Pacific Islander | -1.997 | 0.133 |
| Multiracial | 1.67 | 0.016 |
| Gender | 0.11 | 0.580 |
| Age range | -0.11 | 0.177 |

Table 2: The Regression Coefficient Table for response Q24

The regression coefficient table lists the estimated partial regression coefficients of the output model. The table shows the estimated coefficients and the results of the p-values for each partial regression coefficient. P-values and coefficients in regression analysis work together to tell us which relationships in the model are statistically significant and the nature of those relationships. The coefficients describe the mathematical relationship between each independent variable and the dependent variable. The p-values for the coefficients indicate whether these relationships are statistically significant.

For Q24 (How connected do you feel with HNC?), survey participants are more likely to feel extremely connected with HNC when they are not Hemophilia B (factor 9 deficiency) bleeding disorder; they tend to feel very connected when they are Platelet bleeding disorder; while people with Inhibitor(current or tolerized) bleeding disorder are more likely to feel fairly connected.

Q25: For Q25 (How likely would you be to recommend HNC to friends/family/other people affected by a bleeding disorder?), survey participants tend to be extremely likely and very likely to recommend HNC when they have Rare factor deficiency (factor 1, 2, 5, 7, 10, 12, or 13 deficiency) bleeding disorder.

For Q32_3(How do you feel about the statement “A leader in advocacy” related to HNC?), the age range of the person taking the survey(Q37) has a statistically significant effect on respondent’s feeling about the statement “A leader in advocacy” related to HNC. The lower the age of people, the higher their agreement with this statement.

For Q32_5 (How do you feel about the statement “My top source of support for living with a bleeding disorder ” related to HNC?), whether the member is multiracial(race) or not has a statistically significant impact on the level of agreement with this statement to HNC. If the survey participants are Hispanic or Latino, they are **less** likely to agree with this statement.

For Q33 (Overall, how does HNC meet your needs?), whether the member is an Asian /Asian American(race) or not has a statistically significant impact on the level of agreement with this statement to HNC. If the survey participants are **not** Asian /Asian American, they are **more** likely to agree with this statement.

In conclusion, race and age have a statistically significant impact on the evaluation for HNC in many aspects, especially race. In the future work, HNC may consider how to take care of people of different races to make them more satisfied.

(II)Predict responses Q16, Q17, Q18, Q19, and Q20 with predictors Q2, Q4, Q24, Q25, Q32, Q34, Q36, and Q37

(1)Check the distribution of Q16, Q17, Q18, Q19, and Q20

There are 9 questions, including 32 predictors, so we used Lasso regression in this part to accomplish variable selection. We chose some of them to display in this report, saying which results are similar to this. The symbols used in this part, taking Q16_1 for an example, means the third question in Q16.

First, we checked the distribution for different parts of Q16, Q17, Q18, Q19, and Q20 separately and then decided which methods we would like to use.(Using Logistic Regression if all are strongly agree or agree and Categorical Regression if the categories of responses are more than two types). To be more specific, taking the distribution of Q16 for an example, we found that the feedback for the importance of “Adult Retreat”, “Holiday Celebration”, “Retiro de la Unión Latina” and “Teen Retreat” are mostly concentrated on “Very

important”, ”Important” and “Neither important or not important”. However, different from other events mentioned in Q16, there are still about 10 individuals who think these events are” Not important” and “Not at all important”. So, we used Categorical Regression to analyze these responses. After checking the distribution, we found that the distributions of Q16, Q17, Q18, Q19 and Q20 all have at least 3 categories. Taking Q16_1 for an example, the corresponding histogram is shown in Figure 3. So, the distribution of the answers corresponding to questions listed above suggests that it is reasonable to do the Categorical Regression.

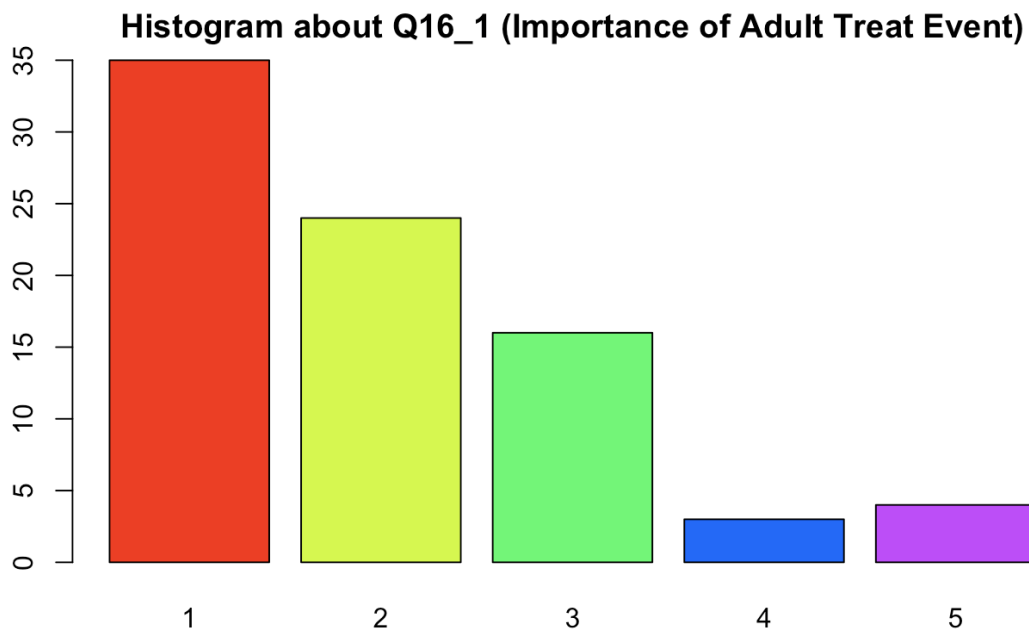


Figure 3: The histogram about Q16_1

In Figure 3, x axis means the participants' satisfaction on the scale of 1-5, while y axis represents the number of participants for corresponding feedback

(2)The corresponding Classification Regression and Lasso Regression for the answer to each question

The symbols used in this part, taking Q16_3 for an example, means the third question in Q16. The details about the figure of Lasso Regression is shown in part(I).

Due to limited report length, the text in the report selects only representative issues for description. The relationship between other issues and other independent variables is listed here in a table.(All predictors listed have statistically significant effects on corresponding responses.)

| Predictor \ Response | 1 | 2 | 3 | 6 | 11 | 12 | 13 | 14 | 15 | 17 | 18 | 19 | 21 | 23 | 25 | 27 | 28 | 29 | 31 |
|----------------------|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---------|
| Q16_5 | | | | | | N | | | | | | | | | | | | | Male(P) |
| Q16_6 | P | | | N | | N | | | N | | | | P | | | | | | |
| Q16_7 | | | | | | | | | | | | | | | | | | | |
| Q18_2 | | | | | P | N | | | | | | | | | | | | | |
| Q18_5 | | | | | | | | | | | P | | | | | | | | |
| Q19_1 | | N | N | | | | | | | P | | | P | | | | | | |
| Q19_3 | | | | | | N | N | | | | P | | | P | | | | | |
| Q19_5 | | | | | | N | N | | | | | | | | | | | | |
| Q19_6 | | | | | | | | | | | | | | P | | | | | |
| Q19_8 | | | | | | | | | | | | | P | | | N | | | |
| Q19_9 | | | | | | N | N | N | | | | | | P | | | | | |
| Q19_13 | | | | | | N | N | | P | P | P | | | P | | | N | N | |
| Q20_6 | P | | | N | | | | | | | | | P | | | N | | | |
| Q20_8 | | | | | | | | | | | | | | | | N | N | | |
| Q20_9 | | | | | | | | N | | | | | P | | | | | | |
| Q20_13 | | | | N | | | | | | | | | P | | | | | | |

Table 3: The relationship between responses Q16, ,Q17, Q18, Q19, and Q20 with predictors indicated below this table

*In Table 3, N (marked with blue color) means there is a significant negative relationship with the predictor and response, while P (marked with yellow color) means there is a significant positive relationship with the predictor and response. If the predictor is a binary variable, P/positive means if the respondent has this characteristic, the individual will give the lower grade for corresponding question, meaning **higher** satisfaction/agreement/demand and vice versa. What needs special explanation is #31, P means the gender of the respondent is male, suggesting that if the respondent is male, he would be more satisfied with corresponding responses.*

The following are corresponding description to the numbers of the first row listed in Table 2:

- 1: Work in the bleeding disorders industry (Q2_1)
- 2: Work in a Hemophilia Treatment Center (HTC) or hematology clinic that provides medical services to people with bleeding disorders (Q2_2)
- 3: Person with a bleeding disorder (Q2_3)
- 6: Sibling of someone with a bleeding disorder (Q2_6)
- 11: Hemophilia B (Q4_3)
- 12: Platelet Disorder (Q4_5)
- 13: von Willebrand disease (Q4_6)
- 14: Inhibitor (Q4_8)
- 15: How likely would you be to recommend HNC to friends/family/other people affected by a bleeding disorder? (Q25)
- 17: How do you feel about the statement that "Puts members first" related to HNC? (Q32_1)
- 18: How do you feel about the statement that "Acts in the best interest of members" related to HNC? (Q32_2)

19: How do you feel about the statement that “My top source of education about bleeding disorders” related to HNC? (Q32_4)
 21: Black/African American (Q34_1)
 23: White/Caucasian (Q34_3)
 25: American Indian/Alaska Native (Q34_5)
 27: Native Hawaiian/other Pacific Islander (Q34_6)
 28: Multiracial (Q34_7)
 29: Age range of the person taking the survey (Q37)
 31: Gender (Q36)

The following are the representative issues for description.

Q16:

Q16_3: We found that for respondent’s perception about the importance of the event "Educational dinner " statistically significantly depends on whether they have another family member of someone with a bleeding disorder(Q2), whether they are connected to or affected by bleeding disorder “Platelet Disorder” (Q4) and how they feel about the statement “My top source of support for living with a bleeding disorder” related to HNC(Q32). Respondents who have another family member of someone with a **bleeding disorder**, who are **not** “Platelet Disorder” bleeding disorder and who have a **higher** agreement with the statement HNC is their top source of support for living with a bleeding disorder tend to think the event "Educational Dinner" is more important.

Q16_4: The respondent’s perception of the importance of the event "Family Festival Walk for Bleeding Disorders" statistically significantly depends on whether they are connected to affected by bleeding disorder “Inhibitor”(Q4), how strong are they willing to recommend HNC to people around them(Q24), and how they feel about the statement “Puts members first”(Q32) related to HNC. Respondents who are **not** affected by bleeding disorder “Inhibitor”, who are **more** willing to recommend HNC to others and who have a stronger feeling that HNC put members first are more likely to think the event "Family Festival Walk for Bleeding Disorders" are more important.

Q16_10: Whether the respondent’s connection to the bleeding disorder community is **Sibling** of someone with a bleeding disorder or not(Q2), how they feel about the statement “My top source of education about bleeding disorders”(Q32) related to HNC, and whether the survey participant is Native Hawaiian/other Pacific Islander(Q34) or not have the statistically significantly impact on their attention to the "Teen Retreat". Survey participants whose connection to the bleeding disorder is **not** Sibling, who agree **more** with the statement HNC is their top source of education about bleeding disorders, and who are **not** a Native Hawaiian/other Pacific Islander tend to pay more attention to the event "Teen Retreat".

Q17:

Q17_1: Whether the respondent’s connection to the bleeding disorder community is **Parent/Guardian** of someone with a bleeding disorder or not(Q2) and whether they are connected to affected by bleeding disorder “**Platelet Disorder**” (Q4) have the statistically significantly effect on their attention to Advocacy Program. Comparatively speaking,

informants who have this connection and are affected by bleeding disorder "Platelet Disorder" are less likely to put emphasis on the importance of the Advocacy Program.

Q17_5: How strongly the informants agree with the statement "My top source of education about bleeding disorders" related to HNC statistically significantly affects their attention on the Mentoring Program. The more they agree with this statement, the more they value the Mentoring Program.

Q17_7: How strongly the informants agree with the statement "My top source of support for living with a bleeding disorder" related to HNC statistically significantly impacts their attention on the SOAR Program. The more they agree with this statement, the more they value the SOAR Program.

Q18:

Q18_3: Whether they are connected to affected by bleeding disorder "Platelet Disorder" (Q4), how strongly they agree with the statement "HNC Puts members first", "My top source of education about bleeding disorders" (Q32), whether the respondent is Black/African American or not (Q34) statistically significantly affect their emphasis given to the service "Emergency financial assistance". If the respondent is **not** connected to affected by bleeding disorder Platelet Disorder, agrees with these two statements more, and is Black/African American, he would tend to pay more attention to the service "Emergency financial assistance".

Q18_6: Whether respondents are connected to affected by bleeding disorder "von Willebrand disease" or not (Q4), how strongly they agree with the statement "Puts members first" (Q32) related to HNC and their genders statistically significantly affect their attention on the service "The Concentrate newsletter". If the respondents is **not** connected to affected by bleeding disorder **von Willebrand disease**, agrees with these two statements more, and the gender is female, the respondent would tend to pay more attention to the service "The Concentrate newsletter".

Q19:

Q19_4, Q19_10 and Q19_12: Similar to the underlying relationship we mention in Q17_5, How strongly the informants agree with the statement "My top source of education about bleeding disorders" related to HNC statistically significantly affects the corresponding feedbacks for Q19_4, Q19_10 and Q19_12 about how highly their need for Emotional support, Emotional support, Ability to navigate healthcare facilities and Raising awareness about bleeding disorders. The more they agree with this statement, the higher their current needs for these three supports.

Q20:

For **Q20_1, Q20_2, Q20_3, Q20_4, Q20_5, Q20_7, Q20_9 and Q20_12**, whose corresponding current needs are "Educational programs about bleeding disorders", "Assistance navigating and understanding health insurance", "Information about research into new treatments for bleeding disorders", "Emotional support", "Financial support", "Local events" and "Legislative advocacy for access to healthcare and treatments", respondents' feeling about "My top source of education about bleeding disorders" has a statistically

significant impact on the these responses. When the respondents agree more with this statement, their current needs for these aspects would be higher.

For **Q20_10 and Q20_11**, whose corresponding current needs are Ability to access healthcare facilities and Access to a Hemophilia Treatment Center, whether their connection to the bleeding disorders community is **Sibling** of someone with a bleeding disorder or not, respondents' feeling about "My top source of education about bleeding disorders" related to HNC and whether the person is American Indian/Alaska Native or not have the statistically significant impact on the these responses. A respondent whose connection to the bleeding disorders community is Sibling of someone with a bleeding disorder, who disagrees "My top source of education about bleeding disorders" related to HNC and who is American Indian/Alaska Native are **less** likely to have the needs Ability to access healthcare facilities and Access to a Hemophilia Treatment Center.

(III)Predict Q24, Q25, Q32 and Q33 by Disorder type (Q4 :What bleeding disorder are you or the person you are connected to affected by?) separately

We treat each part as an indicator in Q4 by using Lasso Regression in part(II). The details about the figure of Lasso Regression is shown in part(I).

Q24: For Q24 (How connected do you feel with HNC?), survey participants are more likely to feel **extremely connected** with HNC when they are **not** Hemophilia B (factor 9 deficiency) bleeding disorder; they tend to feel **very connected** when they are Platelet bleeding disorder; while people with Inhibitor(current or tolerized) bleeding disorder are more likely to feel **fairly connected**.

Q25: For Q25 (How likely would you be to recommend HNC to friends/family/other people affected by a bleeding disorder?), survey participants tend to be **extremely likely** and **very likely** to recommend HNC when they have Rare factor deficiency (factor 1, 2, 5, 7, 10, 12, or 13 deficiency) bleeding disorder.

Q32_4: For Q32_4 (Let us know how you feel about the statement **My top source of education about bleeding disorders** related to HNC?), survey participants are more likely to **strongly agree** this statement when they are **not** Inhibitor bleeding disorder; they tend to **neither agree or agree** this when they are inhibitors bleeding disorder and when they are Glanzmann's bleeding disorder.

Q32_5: For Q32_5 (Let us know how you feel about the **My top source of support for living with a bleeding disorder** related to HNC?), survey participants who do **not** have Hemophilia B bleeding disorder tend to **strongly agree** with this statement.

Q33: For Q33 (Overall, how does HNC meet your needs?), there is no big difference between different bleeding disorders the respondents are or the person they are connected to affected by.

In part (II), we recognize that different bleeding disorders the survey participants are or the person they are connected to affected by may cause their views on HNC and their satisfaction in various aspects to be different. This may be because people with specific

bleeding disorders would get different levels of satisfaction from HNC, which may be more advantageous in providing service and help for some bleeding disorders.

(IV) Predict Q24, Q25, Q32 and Q33 separately by Disorder type(indicated by Q4), which is divided by three groups:

- (a) A, B, inhibitor (current or tolerized)
- (b) von Willebrand disease
- (c) all others

We tried to see if the groups are all statistically significant from each other.

Nonparametric multi-group comparative analysis was performed: Kruskal–Wallis one-way analysis of variance. As the p-value is all more than the significance level 0.05, we can conclude that there are no significant differences between the reclassified bleeding disorders groups. We gives an example in the main result, and other details are in the appendix.

Q24:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|------|-----------|-----------|
| Col | Mean | I | |
| Row | Mean | a | b |
| -----+----- | | | |
| b | | -0.455074 | |
| | | 0.3245 | |
| | | | |
| c | | -1.067414 | -0.713349 |
| | | 0.4287 | 0.3567 |

alpha = 0.05
Reject Ho if p <= alpha/2

Figure 4: The result of dunn test in Q24

Nonparametric multi-group comparative analysis was performed by Kruskal–Wallis one-way analysis of variance and compared pairwise with "Dun". The output following the Kruskal-Wallis test provides all possible pairwise comparisons (three in the case of three groups), so the one on the first row compares group b with group a and the first on the second row compares group c with group a, etc.

The upper number for each comparison is Dunn's pairwise z test statistic. In this example, the lower number is the raw p-value associated with the test (i.e. we would compare to $\alpha/2$, although this p-value changes depending on the family-wise error rate or false discovery rate multiple comparisons adjustment option.) For stepwise multiple

comparison adjustments, the adjusted p-values will have an asterisk next to them if we reject the null hypothesis at the specified significance level.

Since the p-values(a&b:0.3245 a&c:0.4287) are all greater than 0.05, we reject the null hypothesis, indicating that there is not a significant difference between the reclassified bleeding disorders groups for their feeling about how connected with HNC.

In descriptive statistics, boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles. The box plot displays the distribution of data based on the five number summary (minimum, first quartile, median, third quartile and maximum). Outliers are plotted as individual points. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers. The first quartile is the median of the lower half of the data, and the third quartile is the median of the upper half of the data. The distance between the first and third quartiles is called the Inter Quartile Range (IQR). IQR measures variability in a spirit similar to the median. In boxplots, the box is drawn from the first and third quartiles. A long box indicates a large IQR, so that the middle half of the data is more spread. The horizontal line in the box denotes the median of the data. The minimum and maximum are calculated based on the 1.5 IQR rule, which are the ends of two whiskers (vertical lines outside the box) in the boxplot. Data that are either greater than the calculated maximum or smaller than the calculated minimum are suspected outliers.

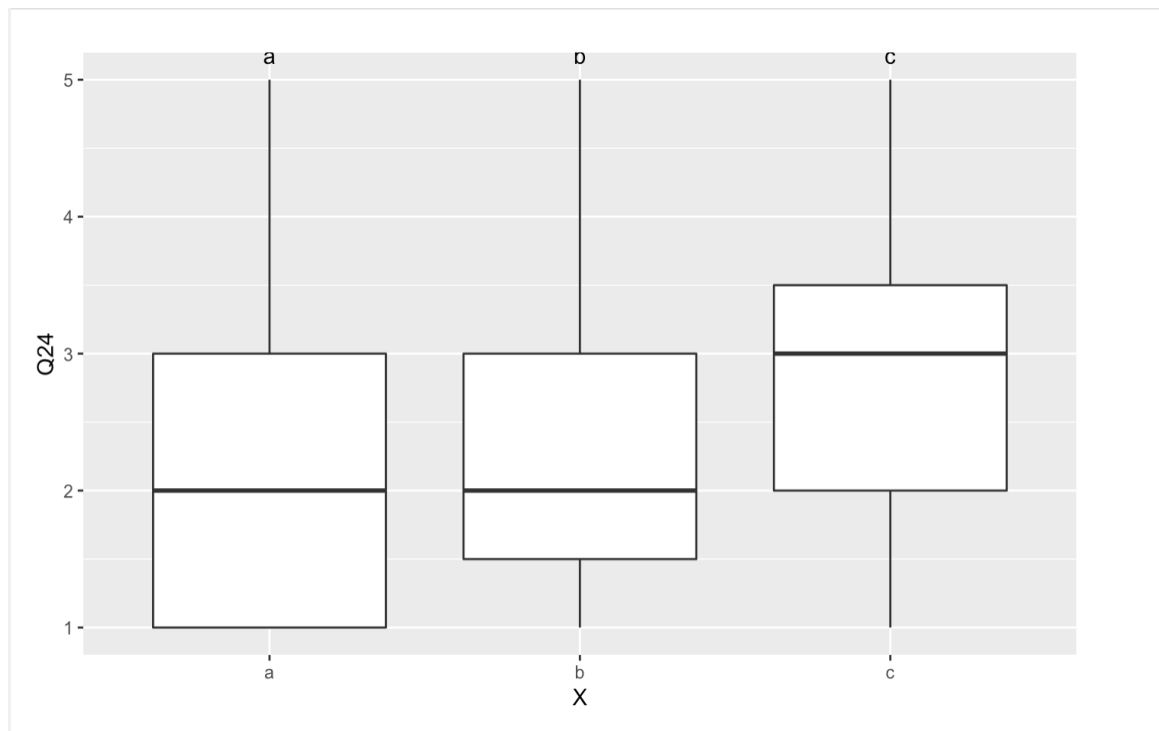


Figure 5: The comparison in the responses for Q24 between group a, b and c

The plot shows the responses for Q24 by different groups (X axis). For y axis, 1 means “extremely connected” and 5 means “Not connected”.

We can see that the median of responses for Q24 are similar for group a(A, B, inInhibitor (current or tolerized)) and group b (von Willebrand disease), while group a has the lower responses for Q24, indicating group a tends to feel more connected with HNC. However, since there is a higher variance in group a, as shown by the bigger box, which means the data is more spread in this group. Besides, people in group c (all others) are more likely to feel not so connected with HNC compared with those who are in group a and group b.

To conclude, from boxplots, we can see that group a, b and c have a great overlap in response to different issues. Corresponding to the p-values in the dunn-test are all less than 0.5. It further confirms for reclassified different groups, there is no significant difference in their feelings about connection with HNC, their willingness to recommend to others, the degree of affirmation and overall satisfaction with HNC in all aspects.

Discussion/Future Directions

In this report, we used Lasso Regression to accomplish the Variables Selection; Categorical Regression to construct the linear regression equation; Ordinal Regression to interpret an ordinal response, from which we identified several important factors that significantly impact satisfaction level and the probability of retention. We assume that the data we used for this model is representative of a typical year and it would be interesting to see if it applies to years following the pandemic. In the future, we could construct a model aiming at predicting a specific membership's satisfaction. This analysis was based on a set of carefully selected variables. We believe there is the potential for many more interesting variables , which could be studied using a variable selection method such as Tree Model.

Appendix:Methods

Categorical Regression: Categorical Regression quantifies categorical data by assigning numerical values to the categories, resulting in an optimal linear regression equation for the transformed variables.

Lasso Regression: The Lasso (Least Absolute Shrinkage and Selection Operator) is a method of automatic variable selection which can be used to select predictors X_1 of a target variable Y from a larger set of potential or candidate predictors X .

Ordinal Regression: In statistics, ordinal regression is a type of regression analysis used for interpreting an ordinal variable, i.e. a variable whose value exists on an arbitrary scale where only the relative ordering between different values are significant. It can be considered an intermediate problem between regression and classification.

Kruskal-Wallis test: The Kruskal–Wallis test by ranks, is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. It extends the Mann–Whitney U test, which is used for comparing only two groups.

Dunn test: Once the initial ANOVA has found a significant difference in three or more means, Dunn's Test can be used to pinpoint which specific means are significant from the others. Dunn's Multiple Comparison Test is a post hoc (i.e. it's run after an ANOVA) non parametric test (a “distribution free” test that doesn't assume your data comes from a particular distribution).

The details of the Part(III):

Q25:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|--------|----------|----------|
| Col | Mean-I | | |
| Row | Mean | I | |
| | | a | b |
| -----+----- | | | |
| b | I | 0.267157 | |
| | I | 0.3947 | |
| | I | | |
| c | I | 0.793112 | 0.575452 |
| | I | 0.6416 | 0.4237 |
| alpha = 0.05 | | | |
| Reject Ho if p <= alpha/2 | | | |

Figure 6: The result of dunn test in Q25

Since the p-value(a&b:0.3947 a&c:0.6416) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the bleeding disorders groups for their willingness to recommend HNC to others who also affected by bleeding disorder.

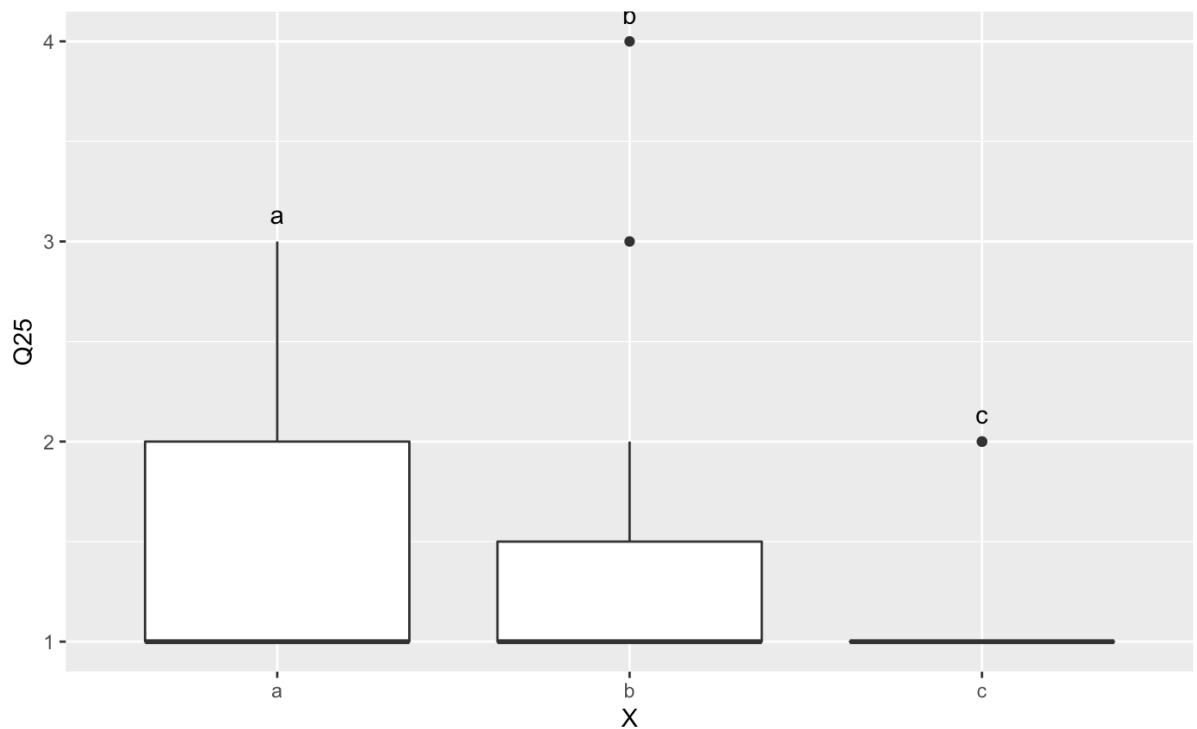


Figure 7: The comparison in the responses for Q25 between group a, b and c

The plot shows the responses for Q25 by different groups (X axis). For y axis, 1 means “extremely likely” and 4 means “probably not likely”.

We can see that the same median of responses for Q25 in group a (A, B, inInhibitor (current or tolerized)), group b (von Willebrand disease), and group c (all others). However, the responses of group a has a higher variance, indicating the responses for Q25 is more spread in this group. While in group c, all of the survey participants in this group are almost extremely likely to recommend HNC to others who are affected by bleeding disorders. Except one outlier in group b indicating that one respondent is probably not likely to recommend HNC, most of the respondents tend to extremely likely and very likely to recommend HNC to others, which means HNC have done so great in service that members are very willing to recommend it.

Q32_1:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|------|----------|-----------|
| Col | Mean | I | |
| Row | Mean | a | b |
| -----+----- | | | |
| b | | 0.533628 | |
| | | 0.8904 | |
| | | | |
| c | | 0.233023 | -0.115492 |
| | | 0.6118 | 0.4540 |

alpha = 0.05
Reject Ho if p <= alpha/2

Figure 8: The result of dunn test in Q32_1

Since the p-value(a&b:0.533628 a&c:0.6118) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the reclassified bleeding disorders groups for their agreement with the statement "HNC put members first".

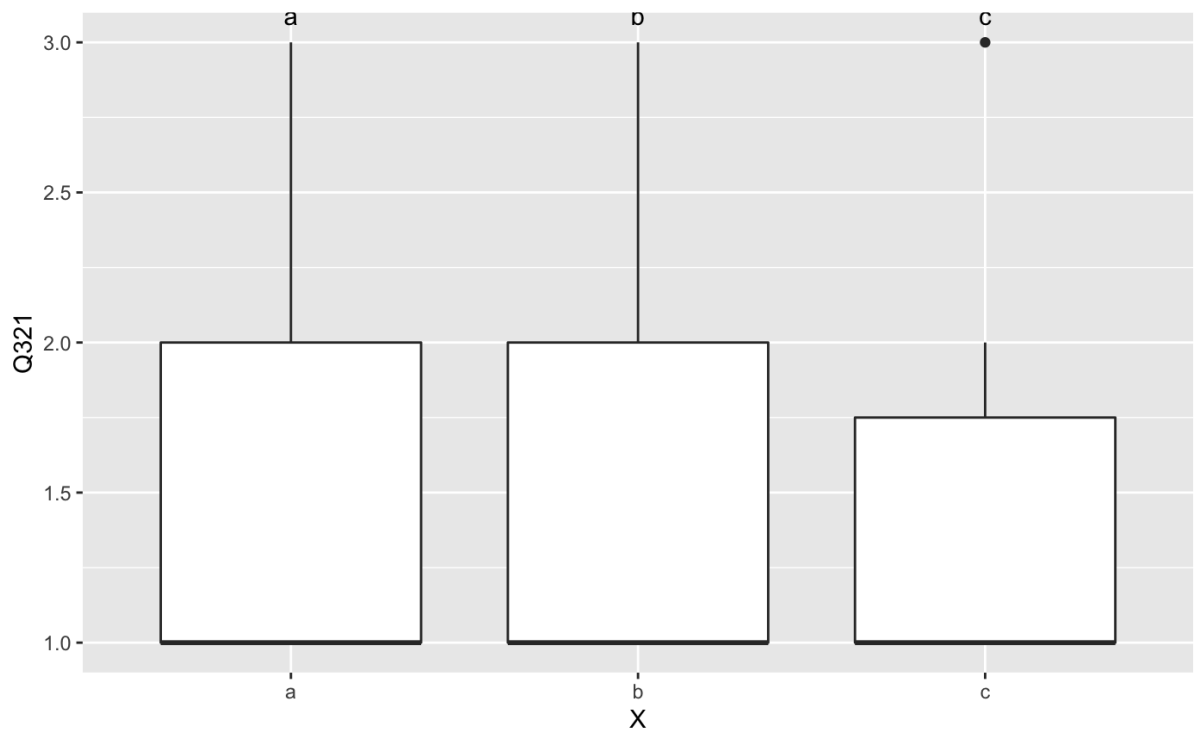


Figure 9: The comparison in the responses for Q32_1 between group a, b and c

The plot shows the responses for Q32_1 (How do you feel about the statement HNC put members first?) by different groups (X axis). For y axis, 1 means "strongly agree" and 3 means "Neither agree or disagree".

We can see that the same median of responses for Q32_1 in group a (A, B, inInhibitor (current or tolerized)), group b (von Willebrand disease), and group c(all others), indicating there are at least half of the informants strongly agree HNC puts members first. Besides, as we can see the responses of group a and b have a higher variance compared with group c, indicating the responses for Q321(agree HNC put members at first) is more spread in this group. In general, in terms of “put members first”, HNC does a pretty good job.

Q32_2:

| Comparison of x by group (Benjamini-Hochberg) | | | | |
|--|------|-----------|-----------|--|
| Col | Mean | I | | |
| Row | Mean | I | | |
| | | a | b | |
| -----+----- | | | | |
| b | | 0.134277 | | |
| | | 1.0000 | | |
| | | | | |
| c | | -0.035182 | -0.116906 | |
| | | 0.4860 | 0.6802 | |

$\alpha = 0.05$

Reject H_0 if $p \leq \alpha/2$

Figure 10: The result of dunn test in Q32_2

Since the p-value(a&b:1 a&c:0.486) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the reclassified bleeding disorders groups for their agreement with the statement "HNC acts in the best interest of members".

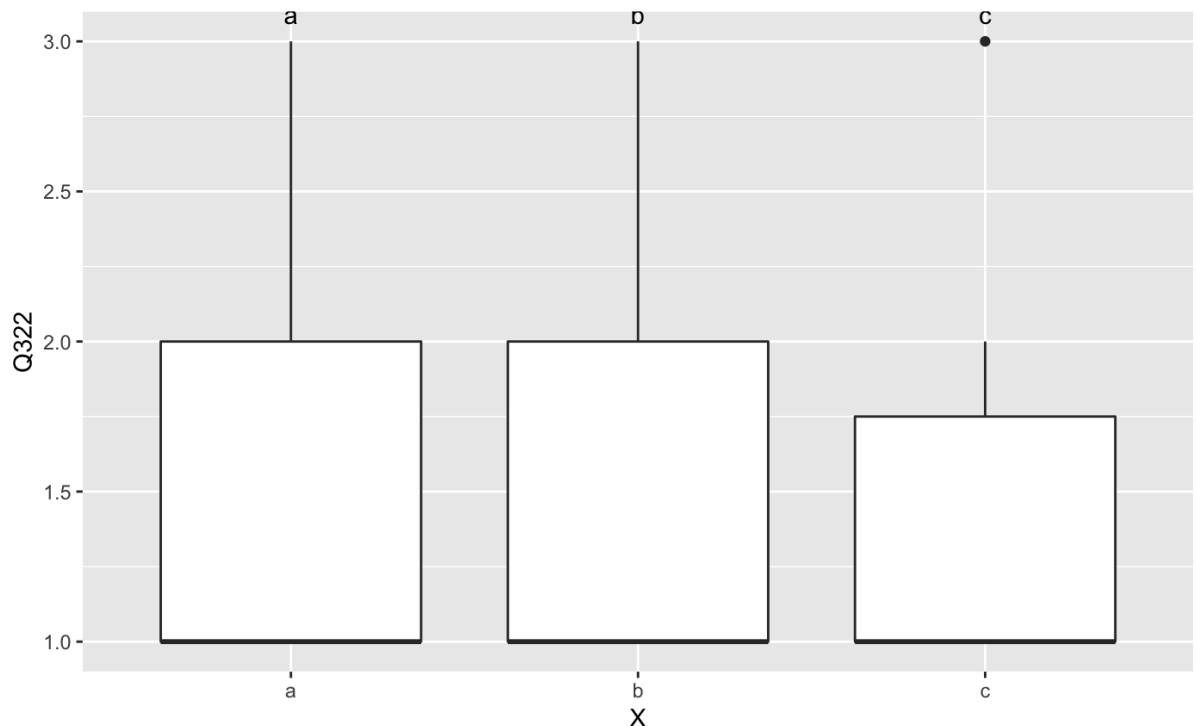


Figure 11: The comparison in the responses for Q32_2 between group a, b and c

The plot shows the responses for Q32_2(How do you feel about the statement HNC acts in the best interest of members?) by different groups (X axis). For y axis, 1 means “strongly agree” and 3 means “Neither agree or disagree”.

We can see that the responses for Q32_2 in group a (A, B, inInhibitor (current or tolerized)), group b (von Willebrand disease) are the same totally, indicating that the members in these two group tend to have the same feeling when it comes to how good HNC acts in the best interest of members. On the other hand, group c has the same median as group a and b, but it has the smaller variance, suggesting that members in group c are more likely to keep the same idea--HNC acts in the best interest of members. In general, more than 75% respondents believe that HNC acts in the best interest of members. Meanwhile, at least half of members strongly agree that HNC was doing a good job in terms of acting in the best interest of members.

Q32_3:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|------|-----------|-----------|
| Col | Mean | | |
| Row | Mean | a | b |
| b | | 0.241843 | |
| | | 0.4045 | |
| | | | |
| c | | -0.875595 | -0.976335 |
| | | 0.2859 | 0.4933 |

$\alpha = 0.05$

Reject H_0 if $p \leq \alpha/2$

Figure 12: The result of dunn test in Q32_3

Since the p-value(a&b:0.4045 a&c:0.2859) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the bleeding disorders groups for their agreement with the statement "HNC is a leader in advocacy".

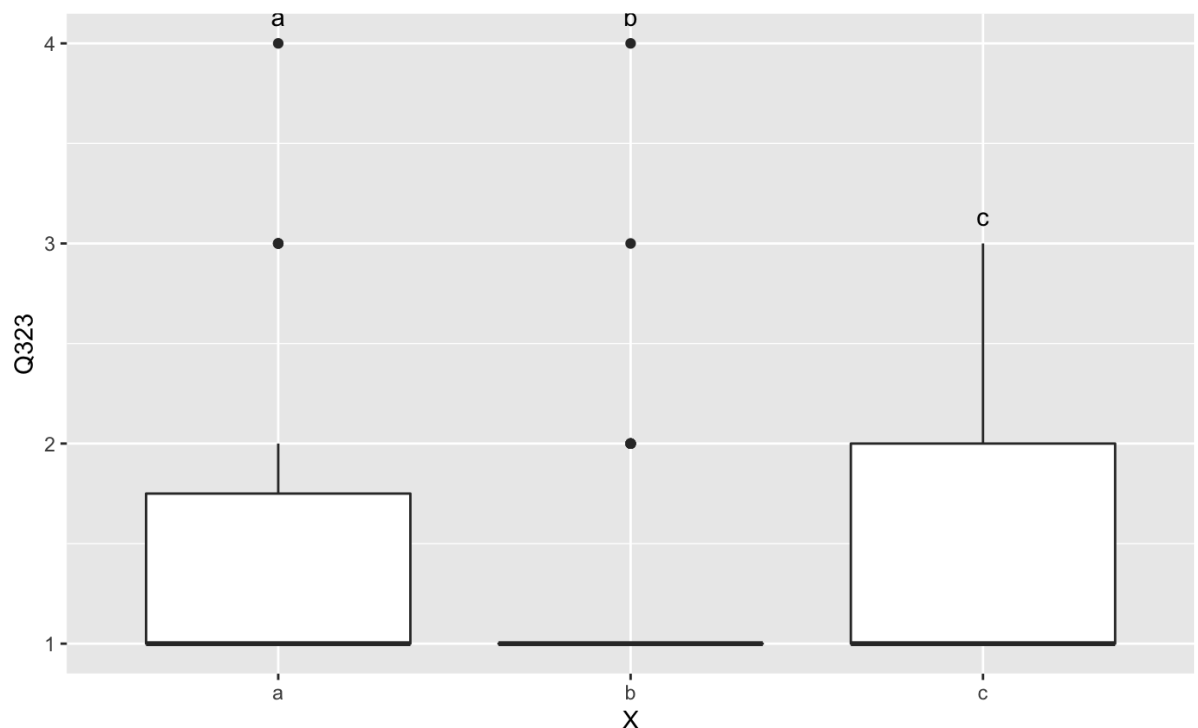


Figure 13: The comparison in the responses for Q32_3 between group a, b and c

The plot shows the responses for Q32_3(How do you feel about the statement HNC is a leader in advocacy?) by different groups (X axis). For y axis, 1 means "strongly agree" and 4 means "Disagree".

As we can see, when it comes to the feeling about “A leader on advocacy”, the data is most concentrated in group b compared with other groups, and the responses of this group are mostly concentrated on “Strongly agree”. So, members who are von Willebrand disease bleeding tend to have a very high appraisal for HNC about its advocated leadership.

Q32_4:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|------|-----------|-----------|
| Col | Mean | | |
| Row | Mean | a | b |
| -----+----- | | | |
| b | | 0.216761 | |
| | | 0.4142 | |
| | | | |
| c | | -0.319099 | -0.435157 |
| | | 0.5622 | 0.9952 |

$\alpha = 0.05$

Reject H_0 if $p \leq \alpha/2$

Figure 14: The result of dunn test in Q32_4

Since the p-value(a&b:0.4142 a&c:0.5622) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the reclassified bleeding disorders groups for their agreement with the statement “HNC is my top source of education about bleeding disorders”.

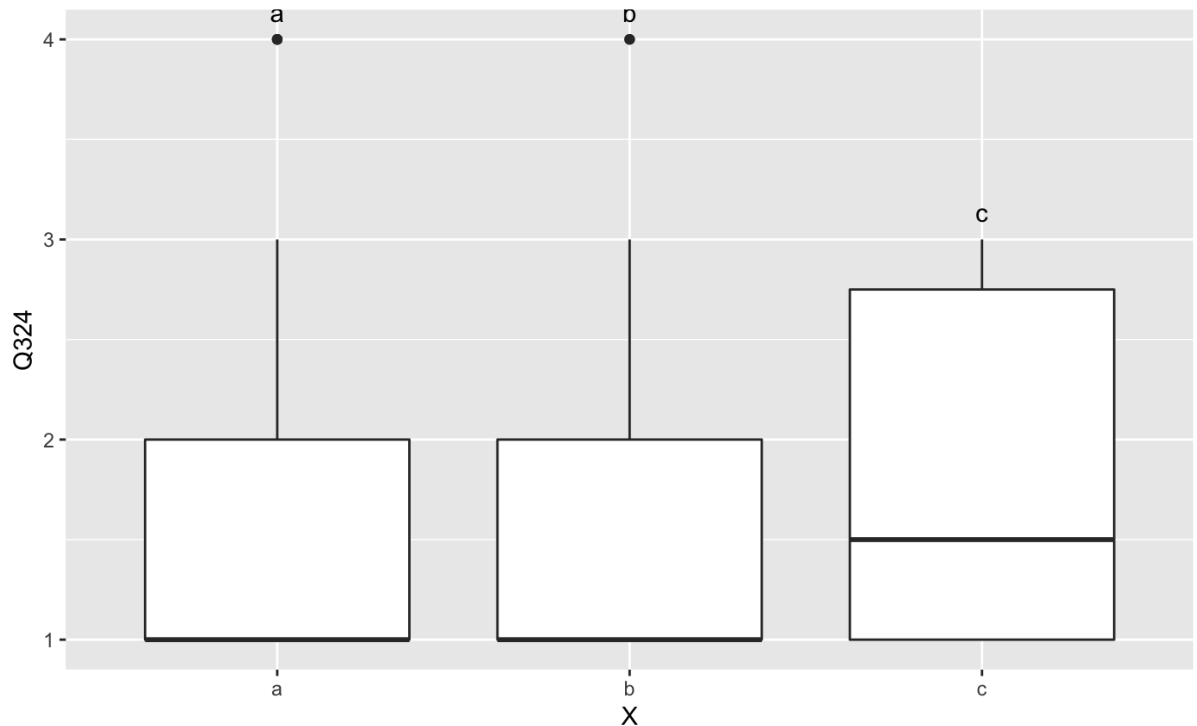


Figure 15: The comparison in the responses for Q32_4 between group a, b and c

The plot shows the responses for Q32_4(How do you feel about the statement HNC is my top source of education about bleeding disorders?) by different groups (X axis). For y axis, 1 means “strongly agree” and 4 means “Disagree”.

As shown in Figure 12, the responses for Q32_4 in group a (A, B, inInhibitor (current or tolerized)), group b (von Willebrand disease) are the same totally, indicating that the members in these two group tend to have the same feeling when it comes to how good HNC work as their top source of education about bleeding disorders.

Furthermore, the members in these two groups are more likely to strongly agree with the statement that HNC is their top source of education about bleeding disorders. In group c, it is obvious that the data is more spread compared with other two groups, indicating the respondents of this group’s agreement with this statement tend to vary from others.

In general, more than 50% respondents agree that HNC is their top source of education about bleeding disorders.

Q32_5:

| Comparison of x by group (Benjamini-Hochberg) | | | | |
|--|------|-----------|-----------|---|
| Col | Mean | - | | |
| Row | Mean | | a | b |
| -----+----- | | | | |
| b | | -0.577218 | | |
| | | 0.2819 | | |
| | | | | |
| c | | -1.188994 | -0.760612 | |
| | | 0.3517 | 0.3352 | |

alpha = 0.05
Reject Ho if $p \leq \alpha/2$

Figure 16: The result of dunn test in Q32_5

Since the p-value(a&b:0.2819 a&c:0.3517) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the reclassified bleeding disorders groups for their agreement with the statement "HNC is my top source of support for living with a bleeding disorder".

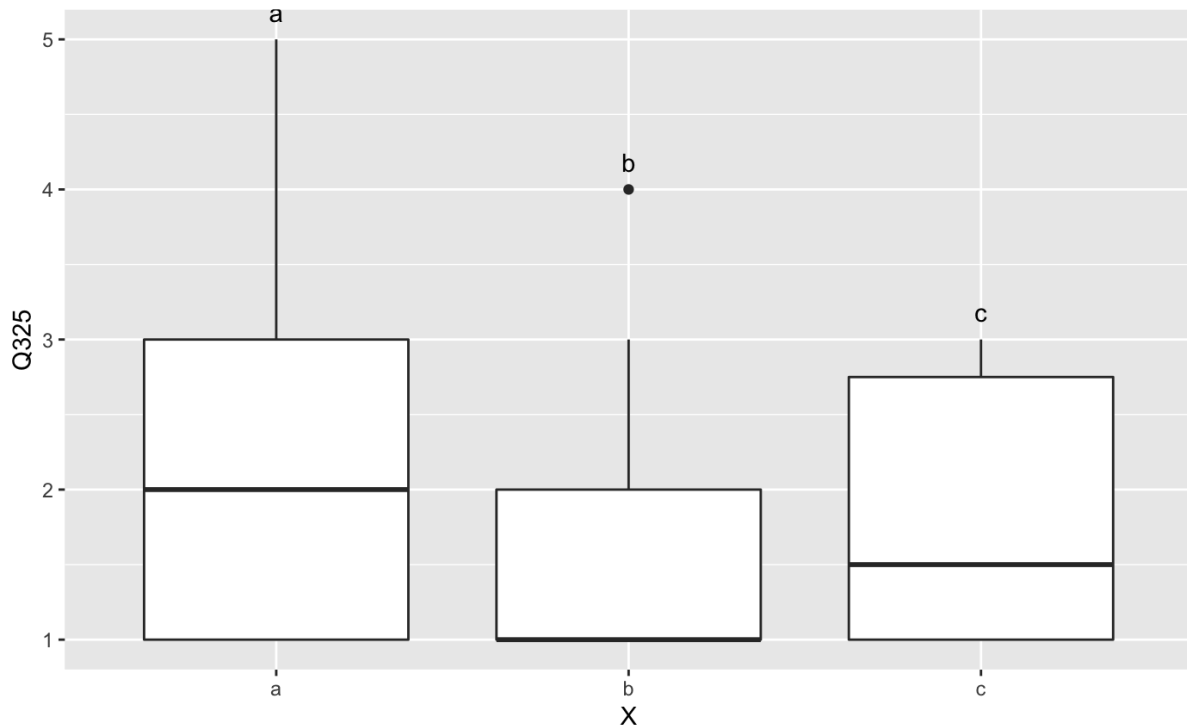


Figure 17: The comparison in the responses for Q32_5 between group a, b and c

The plot shows the responses for Q32_5 (How do you feel about the statement *HNC is my top source of support for living with a bleeding disorder?*) by different groups (X axis). For y axis, 1 means "Strongly Agree" and 5 means "Strongly Disagree".

As shown in Figure 14, the informants in group a are more likely to disagree with the statement *HNC is their top source of support for living with a bleeding disorder*, while the informants in group b tend to have a higher evaluation for this character of HNC.

Q33:

| Comparison of x by group (Benjamini-Hochberg) | | | |
|--|------|----------|-----------|
| Col | Mean | | |
| Row | Mean | a | b |
| -----+----- | | | |
| b | | 1.072182 | |
| | | 0.4255 | |
| | | | |
| c | | 0.114095 | -0.559031 |
| | | 0.4546 | 0.4321 |

alpha = 0.05
Reject Ho if p <= alpha/2

Figure 18: The result of dunn test in Q33

Since the p-values(a&b: 0.4255 a&c:0.4546) are all greater than 0.05, we reject the null hypothesis, indicating that there is not the significant difference between the reclassified bleeding disorders groups for their overall satisfaction of HNC.

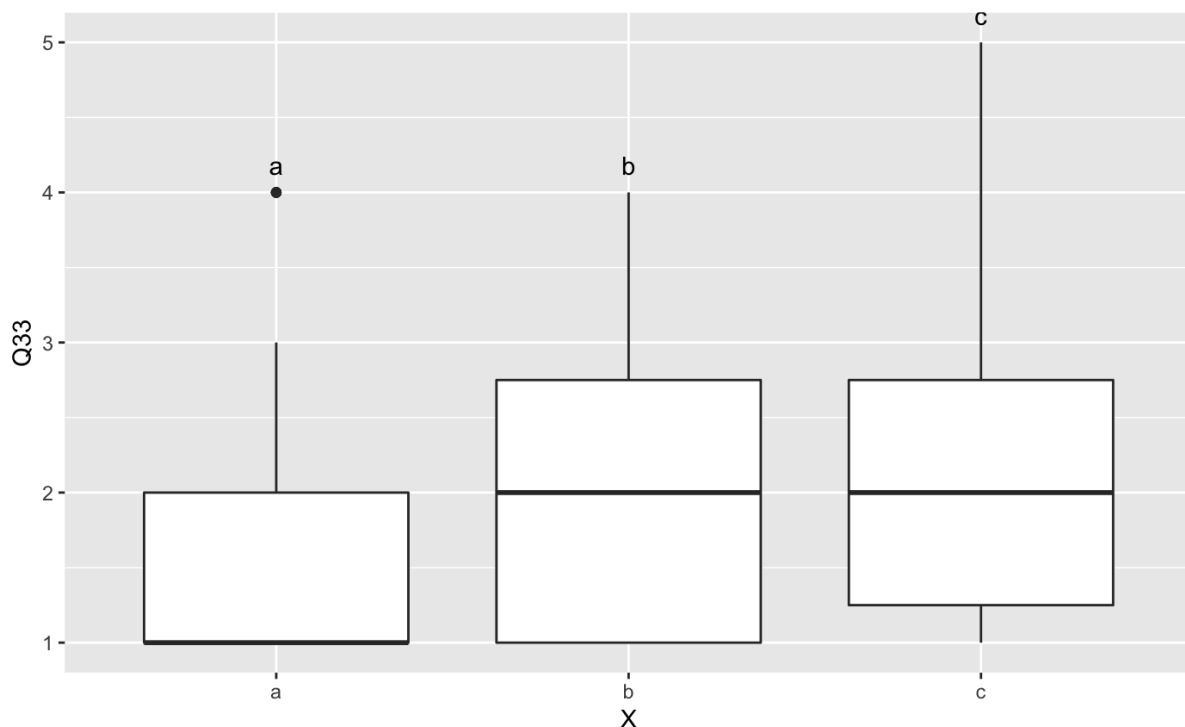


Figure 19: The comparison in the responses for Q33 between group a, b and c

The plot shows the responses for Q33 (Overall, how does HNC meet your needs?) by different groups (X axis). For y axis, 1 means “strongly agree” and 5 means “Strongly Disagree”.

Generally speaking, there is a higher overall evaluation of HNC within the members who are A, B, inInhibitor bleeding disorder(group a). The overall evaluation of HNC in group c (all others) is the most unevenly distributed, so we speculate that members of group c are more likely to have greater differences in overall evaluation of HNC due to their own circumstances. Perhaps, HNC can consider providing personalized services and assistance to members of group c.