A Critical Analysis of a Global Health Organization:

Programmatically Enhancing Data Collection Methods

By: Erin O'Neill

Summer 2023 - Spring 2024

The *Kenya Relief Survey* was first conducted in 2021 and gathered information from southwestern Kenya, Africa residents. As part of the Kenya Relief Organization, the survey aimed to gather insights into the factors influencing the health of Kenyans and the impact of Kenya Relief. Two years later, the resulting dataset was analyzed. This document contains the analysis, discussion, visualizations, and improvements related to the survey.

The survey is a collection of 123 responses to 61 questions. The survey combines various data types written or verbally gathered from participants. Data was retrospectively organized into the following categories, where "—" indicates a relational question:

Baseline: gender, age, community, parental status, number of children, education level, literacy, occupation, community involvement, role.

Housing: Number of rooms in the house? What are the walls of your house made of? What is the roof of your house made of? What is the floor of your house made of? Do you have mosquito nets – do you use them? Has the house been sprayed with insecticide? When was the house last sprayed with insecticide? What is the name of the organization that sprayed the house? Does your house have electricity? If you could do something at home, what would you change?

Water: Do you have access to water in your house? Where does your water come from? Do you walk to get water? How far do you walk for water? Do you have a Safe Water system in your community? How far is the access point to gather safe water? Do you have a Safe Water container? Do you use it only for Safe Water collection, or have you used it for other water collection areas that are not safe? Do you boil or do any cleaning process on your water before drinking it? If yes, what do you do?

General Health: Do you have a health center in your community? How far is the health center? Are there any health center professionals in your community – who are they? How many times a year do you get sick? How many days have you missed work this year due to illness? Do you visit a health professional when you get sick – if not, why not? What diseases are more frequent at home? Has your child been vaccinated with Mosquirix (given at 6mo, 7mo, 9mo, and 24mo in the upper arm) to prevent malaria? Have there been any deaths in your family – do you know the cause? Do you know of any deaths in your community – do you know the cause? What are the biggest health problems you and your family are facing? What are the biggest health problems in your community? How would you describe the health of your community?

Specific Health: Have you had a test for schistosomiasis (snail fever or bilharzia)? If so, when was that test? When was the last time you had schistosomiasis? How was it treated? How long were you sick? Have you had African Trypanosomiasis (sleeping sickness from the tse tse fly)? If so, how long was it treated? How long were you sick?

Intervention: If you could access health talks and education, what would you like to learn? Is there any non-governmental organization working in your community— what do they do? Is there any ongoing project from these organizations on the community right now? If you could change something in your community, what would you change? What would you ask if you could make a request for the community?

Preprocessing Data

After importing the survey data into a coding environment, a basic analysis was conducted to assess the data's representativeness. 64 females and 58 males responded to the survey. The average age of all respondents was 44.84 years, and the average number of children was five. The dataset included responses from 75 different communities. The naming conventions for these communities were inconsistent, making it difficult to determine how many represented the same area (e.g., "Karamu Community" vs. "Karamu Water Project"). The community of Nyamome had the highest representation, with seven individuals reporting residency, followed by Luo.

The "How far is the health center?" responses were reported in minutes, hours, miles, meters, and kilometers. The distance was standardized to kilometers, and the data was converted accordingly. Pattern recognition was used to identify the numerical distance followed by unit, accounting for variations in unit naming. If the response followed the pattern, the data was included. After standardizing the distance, a new column called "Standardized Distance" was created with the remaining data. Distance should be recorded in one metric for future survey versions.

The “How many days have you missed work this year due to illness?” responses also required standardization for analysis. The responses that were initially given in weeks were converted to days. A pattern recognition technique was employed, with the understanding that most responses followed the same format. The results were stored in a new column called “Days Unable to Work.” Participants should be encouraged to respond using the metric requested in the question for future survey versions.

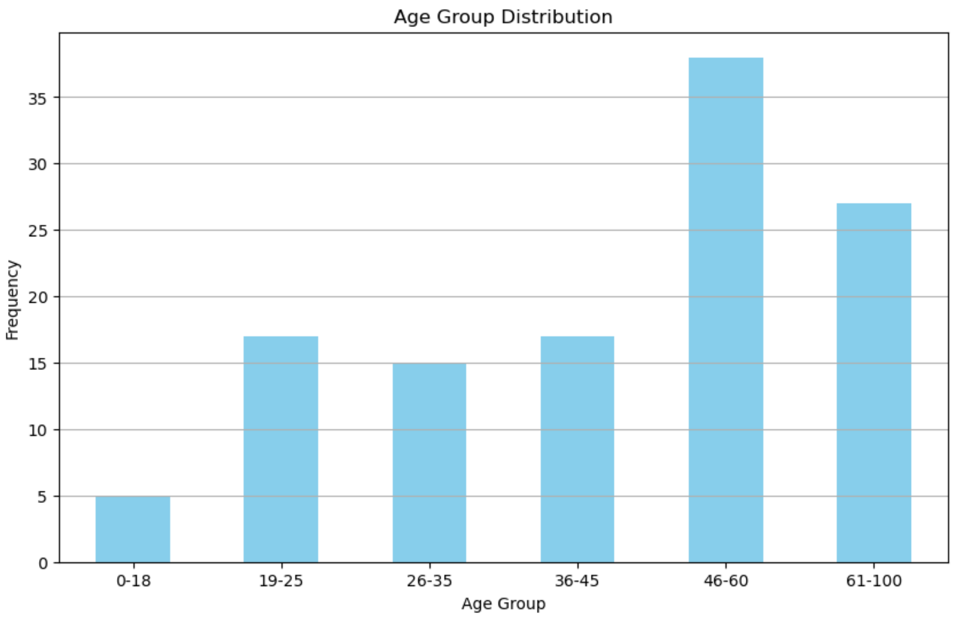

Fig.1. Age Group Distribution

What Were the Top Four Changes Persons in Leadership Would Like to See Occur?

To address this question, the dataset was filtered by leadership. Sub-datasets were created based on the question, "Do you have a role in the community?" If the answer to this question was "yes," the corresponding metadata was assigned to the leadership sub-dataset. There were 50 observations associated with the leadership sub-dataset. A function was developed to identify the five most common words to answer the questions "If you could change something in your community, what would you change?" and "If you could make a request for the community, what would you ask for?" Stop words were removed. Stop words, such as "the," "and" and "is," only carry significant meaning when used in conjunction with other words. The top response to both questions was “water,” with 25 and 22 occurrences, respectively. To gain further insight, categories were created to represent all responses in the dataset.

Health: Health, Medical, Clinic, Hospital, Doctor, Medicine, Care.

Education: Education, School, Teacher, Classroom, Learning, Library, College, University.

Food: Food, Nutrition, Hunger, Meal, Nutrition, Diet, Grocery, Eat.

Housing: Housing, Homeless, Shelter, House, Living, Residence, Accommodation.

Employment: Employment, Job, Work, Career, Occupation, Unemployment, Income.

Environment: Environment, Clean, Pollution, Polluted, Water, Sanitation, Hygiene.

Social Support: Social, Community, Support, Network, Friend, Family, Neighbor, Belonging.

"If you could change something in your community, what would you change?"

Health: 14

Education: 4

Food: 9

Housing: 2

Employment: 4

Environment: 43

Social Support: 5

"If you could make a request for the community, what would you ask for?"

Health: 12

Education: 2

Food: 5

Housing: 1

Employment: 0

Environment: 32

Social Support:4

What Were the Top Four Changes that Persons in Non-Leadership Would Like to See Occur?

The same programmatic process utilized above was employed to answer this question. There were 70 observations associated with the non-leadership sub-dataset. The five common response words were identified, with water having the most counts. In response to “If you could change something in your community, what would you change?” participants responded with water 35 times. In response to “If you could make a request for the community, what would you ask for?” participants responded with water 24 times. Words were then assigned to categories.

“If you could change something in your community, what would you change?”

Health: 32

Education: 3

Food: 8

Housing: 3

Employment: 1

Environment: 43

Social Support: 13

“If you could make a request for the community, what would you ask for?”

Health: 25

Education: 7

Food: 6

Housing: 2

Employment: 2

Environment: 33

Social Support: 16

What are the most frequent untreated diseases? And determine if vaccination treatment is available.

To answer this question, significant values were counted, excluding nefarious punctuation. Each column was investigated to understand response structure and data cleanliness. The responses to "What diseases are more frequent at home" were determined to contain the cleanest, structured data. The information from this column was transformed into a boxplot graph to account for data loss using the counting method.

"What diseases are more frequent at home?"

The top response to this question was "Malaria," with 77 counts, followed by "Typhoid" with 26.

"What are the biggest health problems you and your family are facing?"

The top response to this question was "Malaria," with 49 counts, followed by "Typhoid," with 15.

"What are the biggest health problems in your community?"

The top response to this question was "Malaria," with 59 counts, followed by "HIV," with 29 counts, and "Typhoid," with 23.

The diseases and health problems documented in the "What diseases are more frequent at home" column were categorized for a more representative understanding into the following:

Respiratory Keywords: Cough, Chest/Respiratory, Pneumonia, Asthma.

Gastrointestinal Keywords: Stomach Virus, Stomachache, Typhoid, Ulcers, Abdominal Pain, Enuresis, Diarrhea, Water-Borne Diseases, Cholera, Amoebiasis.

Vision Keywords: Eye Problems, Cardiovascular Keywords, Chest Problems, Chest Pain, High Blood Pressure, Blood Pressure.

Musculoskeletal Keywords: Arthritis, Back Ache, Knee Problems, Pain in the Knees, Skin Disease.

Neurological Keywords: Headaches, Headache, Epilepsy, Fevers, Lipoma, Neck Problems, Depression.

Dental Keywords: Tooth Decay, Dental.

Metabolic Keywords: Diabetes, Sickle Cell, Goiter, Growth, Thyroid Problems, Numbness

Infectious Keywords: Malaria, Flu, AIDS, HIV, Brucellosis, Wounds, Cholera, Typhoid, Amoebiasis, Water-Borne Diseases, Diarrhea, Flu, Tuberculosis.

Other Keywords: Allergies, Cancer, Urinal Problems, Itching Skin, Nauseated, Sleepless Nights, Kids Suffering, Milk/Chicken Diseases.

The column was filtered for these keywords and assigned a category. The two boxplots produced considered all rows of data.

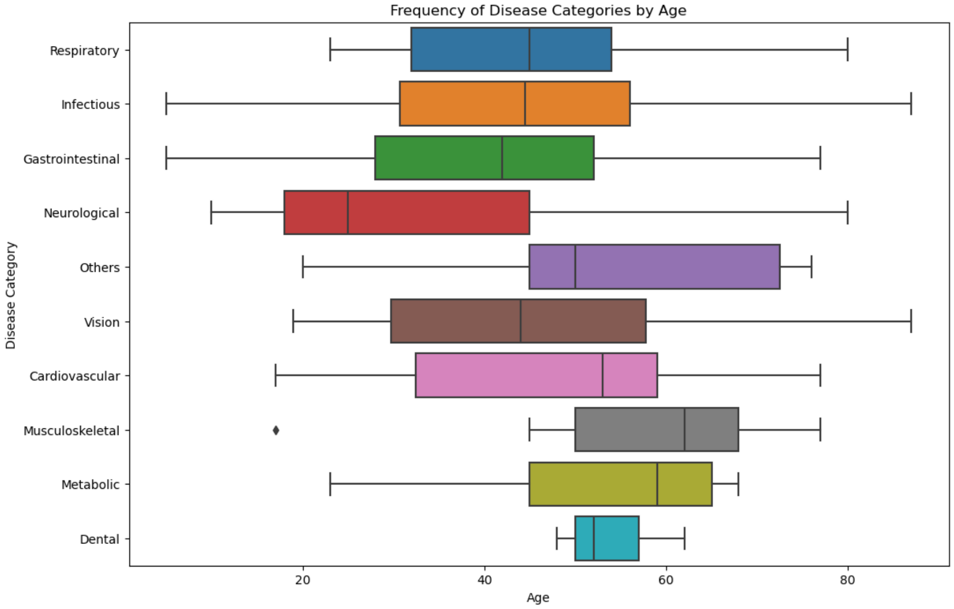

Fig. 2. Frequency of Disease Categories by Age.

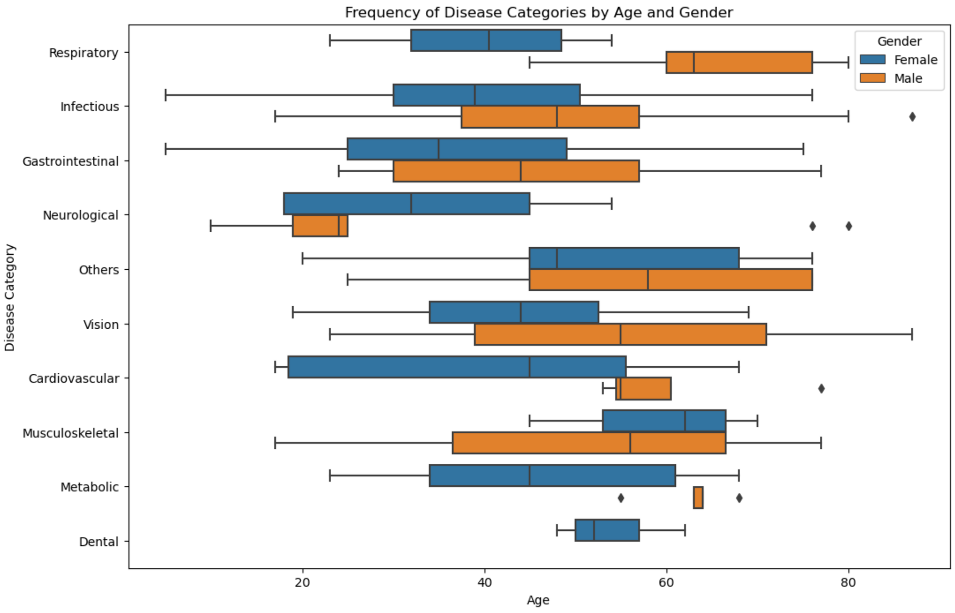

Fig. 3. Frequency of Disease Categories by Age and Gender.

Relationship between Days Unable to Work Due to Illness and Distance to a Clinic

The data was normalized using min-max normalization, and the Pearson correlation coefficient was calculated to investigate the potential relationship between days unable to work due to illness and distance to a clinic. The computed correlation was 0.26, indicating no strong relationship between the two variables in the dataset. A stronger relationship may exist, but data loss may have impacted the results. Many participants provided uninterpretable responses to the question "How many times a year do you get sick," such as "Many times" and "Rarely," which were omitted from the analysis. This trend was observed throughout the data, highlighting the need for improvements such as gathering standardized data to understand better the correlation between days unable to work due to illness and the distance to a clinic. The two scatterplots below display the same plot but have different x and y-axis scales.

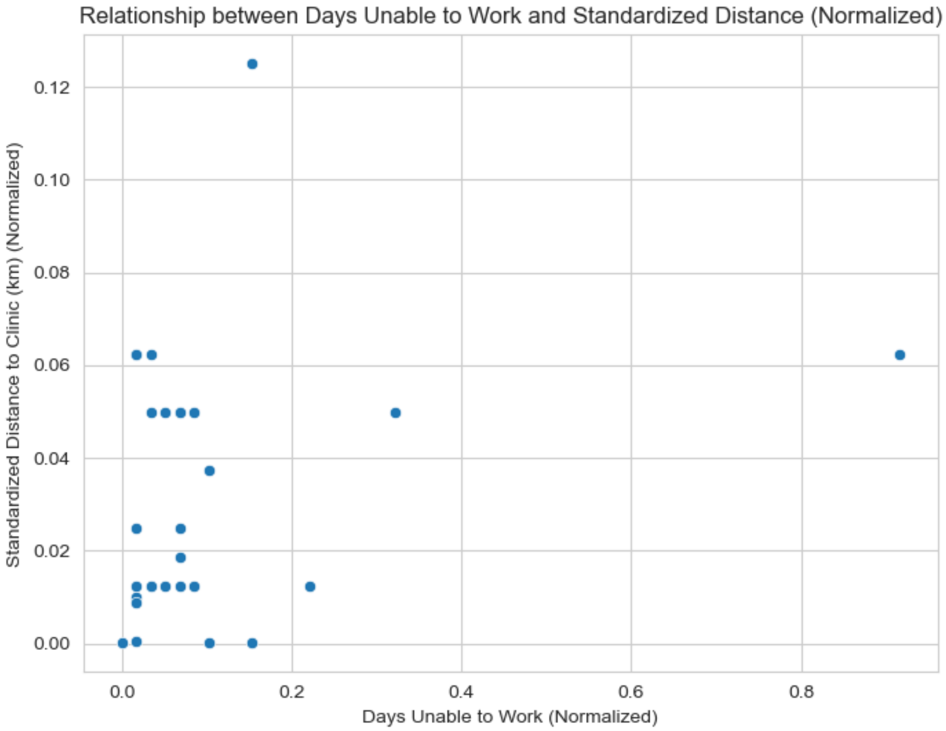

Fig. 4. Relationship between Days Unable to Work and Standardized Distance (Normalized)

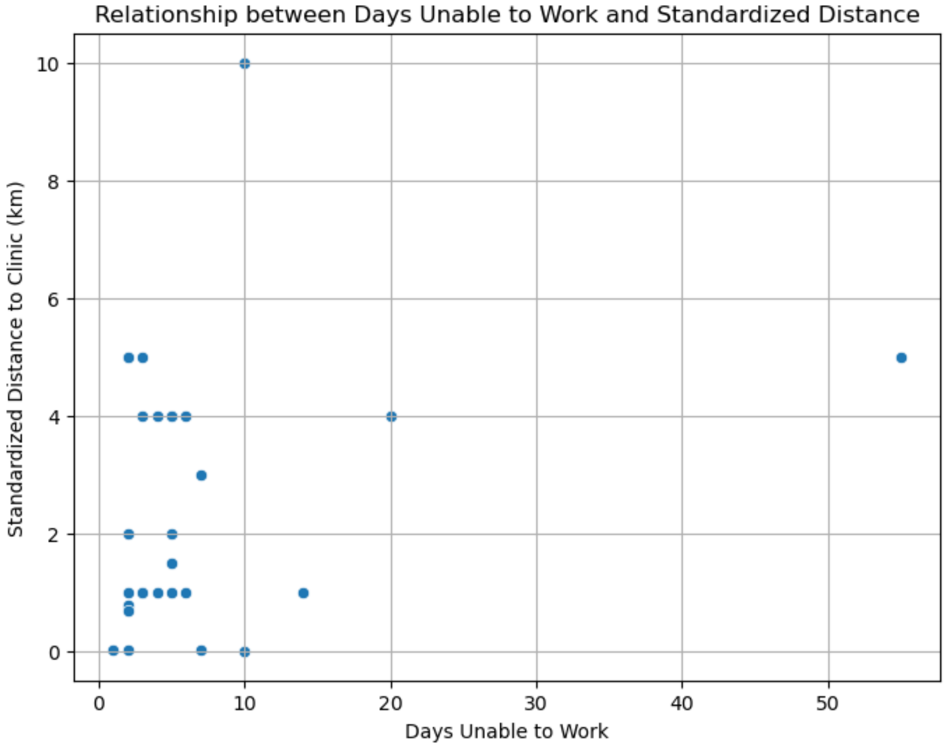

Fig. 5. Relationship between Days Unable to Work and Standardized Distance

What would you like to learn if you could access health talks and education?

A rigorous programmatic procedure determined what participants would like to learn about regarding health talks and education. First, the common words and phrases used in response to the question were counted, again using stop words and natural language processing (NLP) methods like tokenization and n-grams. Tokenization splits text into smaller units, and n-grams provide word relationships. Target words were identified by counting the top five common phrases. The target words were prevention, diseases, hygiene, malaria, and cancer. The common phrases were extracted using these target words and assigned to their corresponding target word category. The top five target word categories and their phrases are below:

Phrases Containing Prevent: Prevent malaria: 7 occurrences; Malaria prevention: 4 occurrences; Prevent diseases: 3 occurrences; Prevent illnesses: 2 occurrences; Preventative measures: 2 occurrences; Help prevent: 1 occurrence; Prevent AIDS: 1 occurrence; Prevention measures: 1 occurrence; Preventing illnesses: 1 occurrence; Prevent HIV/AIDS: 1 occurrence; Preventing diseases: 1 occurrence; Prevent hypertension: 1 occurrence; Prevent mosquito-borne: 1 occurrence; Prevent cancer: 1 occurrence; Causes prevention: 1 occurrence; Prevent different: 1 occurrence; Prevent infections: 1 occurrence; Prevent disease: 1 occurrence.

Phrases Containing Diseases: Prevent diseases: 3 occurrences; Infectious diseases: 2 occurrences; Blood diseases: 1 occurrence; Contact/contract diseases: 1 occurrence; Preventing diseases: 1 occurrence; Mosquito-borne diseases: 1 occurrence; Different diseases: 1 occurrence; Transmitted diseases: 1 occurrence; Many diseases: 1 occurrence; Lethal diseases: 1 occurrence; Lifestyle diseases: 1 occurrence; Waterborne diseases: 1 occurrence.

Phrases Containing Malaria: Prevent malaria: 7 occurrences; malaria prevention: 4 occurrences; stop malaria: 2 occurrences; HIV malaria: 2 occurrences; malaria safe: 1

occurrence; malaria headaches: 1 occurrence; AIDS malaria: 1 occurrence; eliminate malaria: 1 occurrence.

Phrases Containing Hygiene: Cleaning/personal hygiene: 1 occurrence; Hygiene using: 1 occurrence; Personal hygiene: 1 occurrence; Clean hygiene: 1 occurrence; Hygiene diabetes: 1 occurrence; Cancer hygiene: 1 occurrence; Hygiene AIDS: 1 occurrence; Hygiene education: 1 occurrence; Hygiene talks: 1 occurrence; Hygiene environment: 1 occurrence

Phrases Containing Health: General health: 2 occurrences; health talks: 1 occurrence; health issues: 1 occurrence; health concerns: 1 occurrence; live healthy: 1 occurrence; healthy lives: 1 occurrence; mental health: 1 occurrence; healthy eating: 1 occurrence; healthy body: 1 occurrence; health requirements: 1 occurrence; stay healthy: 1 occurrence.

Relate morbidity and mortality differences against each community's level of NGO intervention. Positive perceived impact or not?

Mortality data was not numerically reported, but a connection could be established between illness-related work absence and the presence of non-governmental organizations (NGOs). Due to the amount, it would not provide significant insights to analyze the data by individual communities, but conducting future analyses based on regional breakdowns could be informative. A boxplot and swarm plot were used with consistent coloring to help illustrate the breakdown of data.

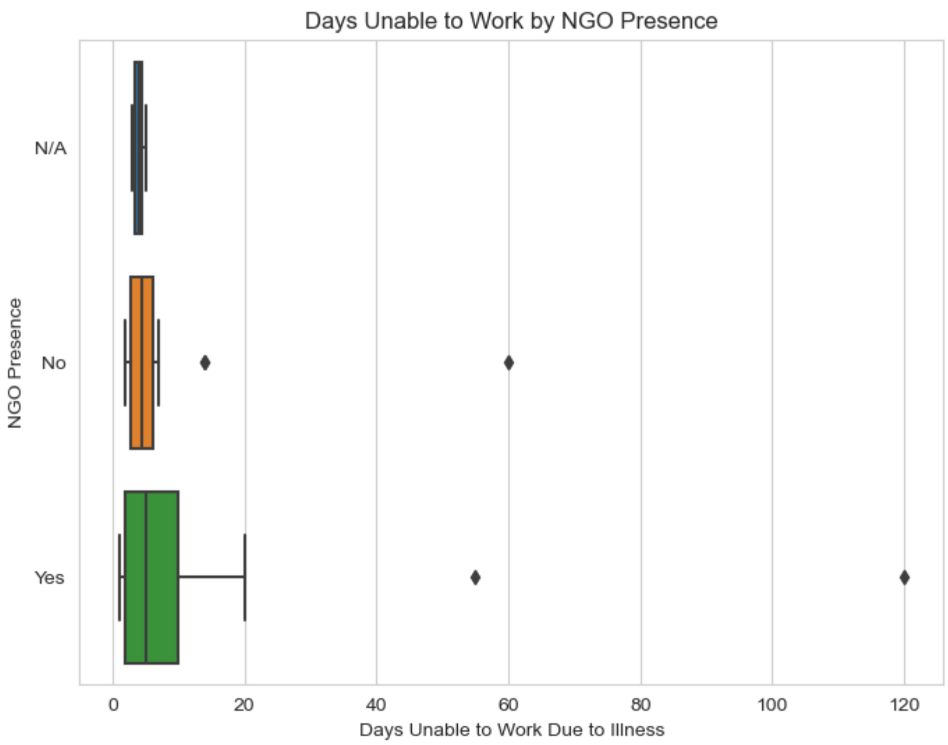

Fig. 6. Days Unable to Work by NGO Presence

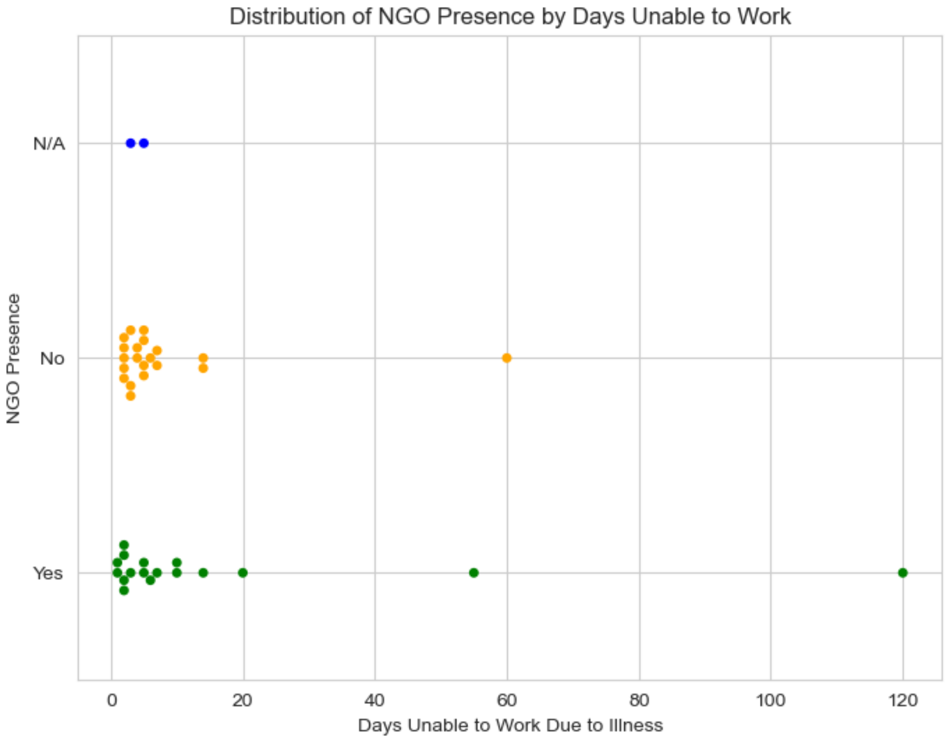

Fig. 7. Distribution of NGO Presence by Days Unable to Work

Based on the swarm plot, it is evident that there are 20 data points for "No," 17 for "Yes," and 2 for "N/A." The effectiveness of these plots relied on whether the participant responded to the NGO presence question and reported the number of days unable to work due to illness. Some participants answered "N/A" to the NGO question but reported their days unable to work due to illness, represented in blue. A t-statistic and p-value were calculated. The t-statistic was 1.067, and the p-value was 0.293. Based on the data, there is no significant difference in days unable to work between communities with and without NGOs. However, these results may only be partially representative due to data missingness.

Discussion

The analysis provided numerous insights. The dataset displays a well-distributed age range (5-87), and the gender and community role data were also evenly collected. As expected, the main drawback of the data is that it was not collected in a standardized manner, which limits the informativeness of the results. Additional data collection would enhance the results.

Based on the survey, the top changes individuals identifying as leaders and non-leaders would like to see are environmental improvements, especially water-related. A greater proportion of non-leadership participants reported wanting to see health-related improvements. When combining the responses for “If you could change something in your community, what would you change?” and “If you could make a request for the community, what would you ask for?” 26/137 (~19%) of the responses fell into the health category for leadership participants versus 57/194 (~29%) for non-leadership.

Participants consistently reported problems with malaria and typhoid, both of which have vaccine treatments available. Notably, a greater proportion of participants responded with “HIV” to the question “What are the biggest health problems in your community?” compared to similar questions like “What are the biggest health problems you and your family are facing?” and “What diseases are more frequent at home?” HIV does not have a vaccine, and health education could be a viable solution to combatting the issue.

As shown in Figure 2, a higher proportion of individuals under the age of 30 reported experiencing neurological health issues on average. Conversely, individuals over the age of 60 reported musculoskeletal problems at a higher rate on average. Figure 3 shows that younger females are more likely to report respiratory problems than their male counterparts. Additionally, females reported experiencing infectious, gastrointestinal, vision, cardiovascular, and metabolic issues earlier than males. On the other hand, males reported neurological and musculoskeletal problems at an earlier age than females.

The data presented in Figures 4 and 5 show no discernible relationship between the number of days unable to work due to illness and the distance to clinics. This is supported by the corresponding Pearson correlation coefficient of 0.27, indicating a weak positive relationship. The plot only contains 21 points because the data needed to be standardized for both variables. More data is required to create a fully representative plot of the relationship between the two variables.

The participants expressed a strong interest in learning about malaria, hygiene, cancer, and general prevention of infectious diseases. Participants indicated a high demand for malaria education, with 32 counts showing a desire for more information on this topic.

Figures 6 and 7 show the number of days unable to work between communities with and without NGOs, which is the same. Out of 123 participants, only 39 data points were plotted, suggesting that the graph may only partially represent the actual relationship. More data is needed for a thorough analysis. Additionally, a major limitation of this question is that participants may inaccurately remember how many workdays were missed due to illness. Overestimating or underestimating could significantly impact the results. Questions like this should be asked in intervals of five or ten to mitigate recall issues.

Future Work

Standardization of the data collection methods would improve the survey's results. Additionally, future analysis could benefit from filtering by community or region once the data has been reconciled. The following specific and general improvements are recommended:

- Community: provide predetermined options for participants and a write-in area
- Education: provide predetermined options (primary, secondary, tertiary, none) and explain what grade levels fall into which
- How many times a year do you get sick/how many days have you missed work due to illness: predetermined intervals (0-5), (10-15), (15-20), etc.
- Have there been any deaths in your family: the question should be changed to ask how many deaths there have been in your family, followed by a question asking the cause. Currently, the latter question data is in Y/N and free response format.
- Any questions asking "how far" should be recorded in a standard distance, following the same format (number, unit), or provide participants with predetermined intervals (e.g., 0-5 kilometers).
- Any questions asking for numerical answers should be recorded in float format (1,2,3), not string (one, two, three)
- Any questions asking for dates (e.g., "When was the house last sprayed with insecticide?"), should all have the same date response format (e.g., "M/D/Y")

- Any questions that may allow for multiple answers (e.g., “What diseases are more frequent at home?”) should have responses separated by commas or consistent delimiters.