Data Paper:

Data For: Development and Use of a Clinical Decision Support System for the Diagnosis of Social Anxiety Disorder

Rachael Beal

Department of Library and Information Science, University of Denver

LIS 4220: Data Curation

Professor Lindsay Gypin

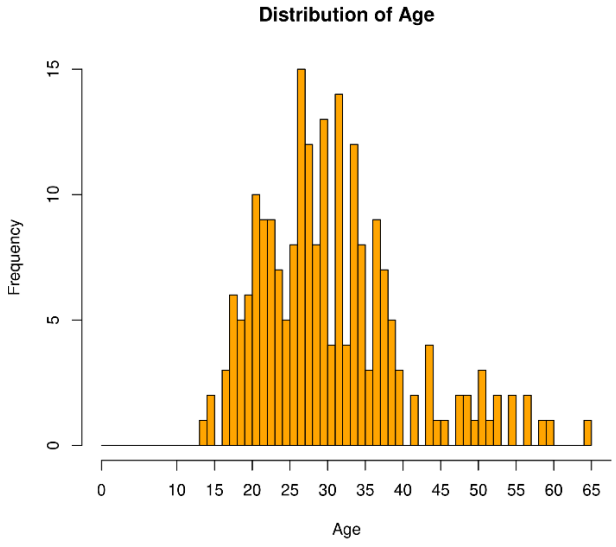
March 7, 2024

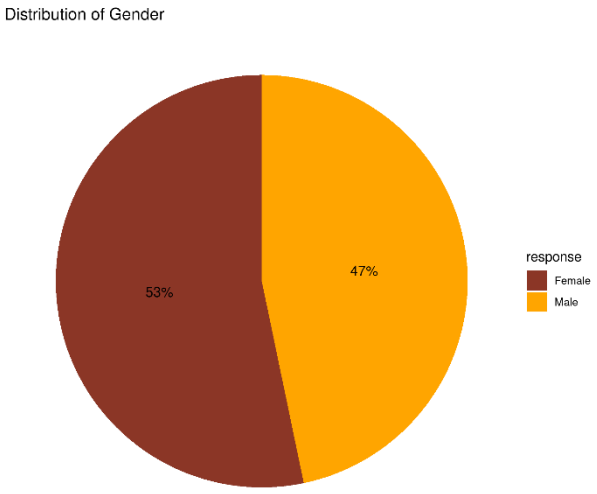
DATA PAPER: CLINICAL DECISION SUPPORT SYSTEM FOR DIAGNOSIS OF SAD

Concept Introduction

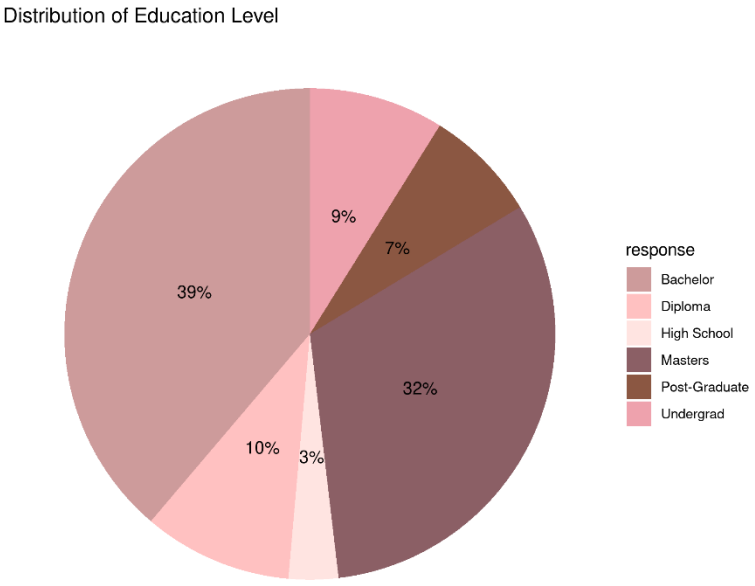
"Data for: Development and use of a clinical decision support system for the diagnosis of social anxiety disorder" contains thirty columns and two hundred and fourteen rows. The first five columns include demographic information such as age, education level, gender, occupation, and whether participants in the study have a family history of social anxiety disorder (SAD) or depression. The next eight columns rate levels of fear of various scenarios based on a ten-point Likert scale. Thirteen columns contain binary responses to personal experiences of physical aspects of SAD. The final three columns contain participants' prior diagnoses and two alternative SAD test results.

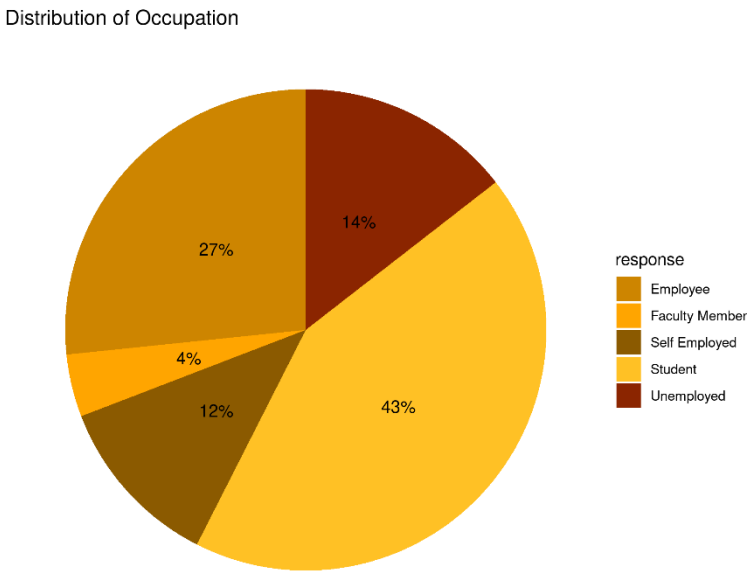
Gender has only been evaluated based on male and female. This could be because participants did not identify as anything otherwise, or because the researchers involved in the study did not consider alternative options and identities. 53% of participants identify as female while the other 47% identify as male. The youngest participant's age is 14. The oldest is 65. The mean age is 31 (30.94), and the mode age is 27.





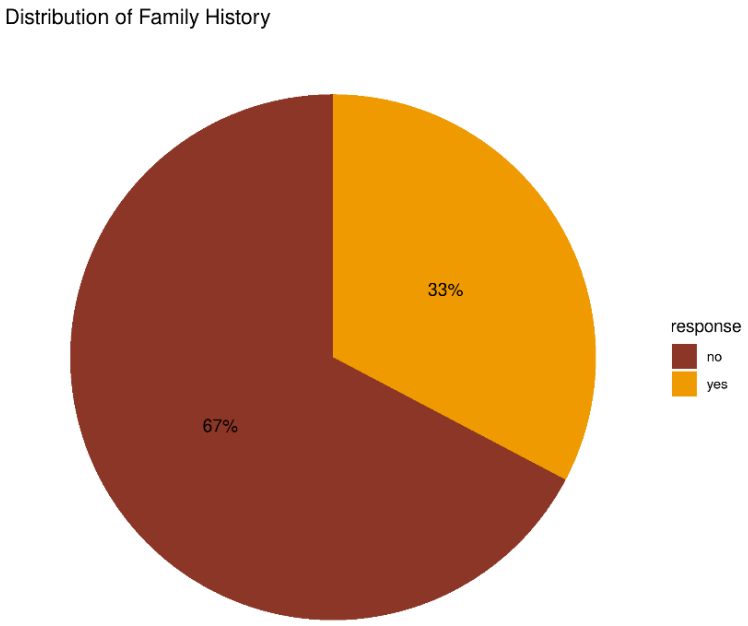
39% of participants reported either currently or already having completed, at most, a bachelor's degree. 32% are currently or have already completed, at most, a master's degree. 10% are currently or have already completed, at most, a diploma. 9% are currently or have already completed, at most, an undergraduate degree. 7% are currently or have already completed, at most, a post-graduate degree. Finally, 3% of participants are currently or have already completed, at most, high school.





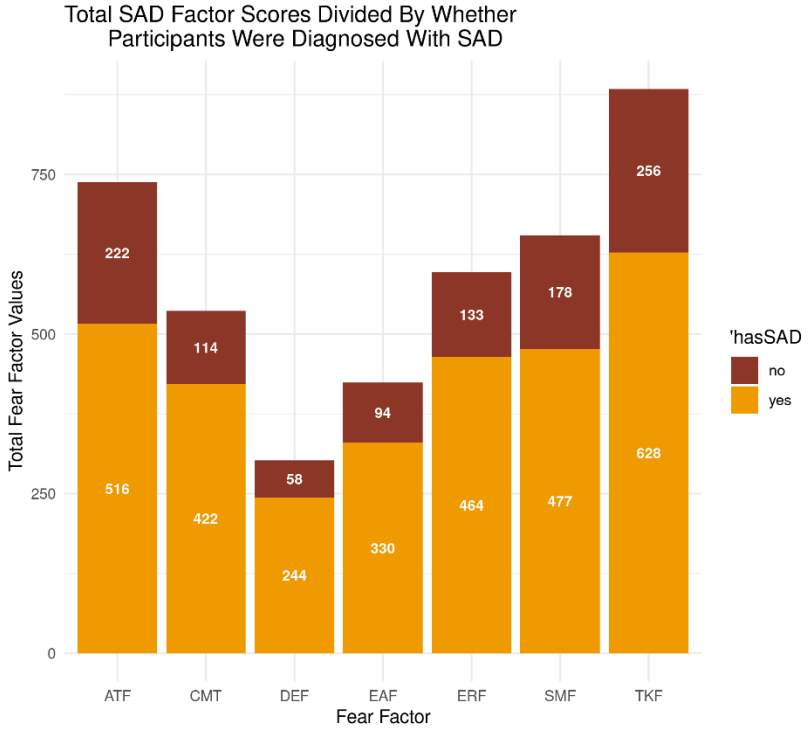
As for occupation, 43% of participants are current students. 27% are current employees. 14% are unemployed. 12% are self-employed. Finally, 4% are current faculty members.

Finally, 67% of participants stated that they do not have a family history of SAD or depression.



Sina Fathi, Maryam Ahmadi, Behrouz Birashk, and Afsaneh Dehnad are the primary researchers who developed this dataset through their study of SAD. They describe SAD as the following: “SAD, also known as social phobia, is one of the most prevalent anxiety disorders described as permanent and severe fear or feeling of embarrassment in social situations such as lecturing or eating in public, being faced with others’ judgement, being the center of attention, etc.” (p. 2, 2020). The researchers’ concern is that patients who suffer from SAD are oftentimes overlooked, and remain undiagnosed or and untreated (Fathi, et. al., 2020). The raw data was originally collected through a website consisting of eleven attributes of SAD based on guidelines developed by DSM-5 and International Classification of Disease, 10th revision, Diagnostic Criteria for Research (Fathi, et. al., pp. 2-3, 2020). These original eleven attributes were

modified and added to by the researchers for the sake of precision and intelligibility on the part of machine learning. It was created for medicinal and mental and physical health approach development. Specifically, to propose and endorse machine learning-based decision support systems for the diagnosis of SAD.



Inside or outside of the data's original intent of use, it could be important to consider the number of participants that have or have not been diagnosed with SAD. Only 52% of participants were previously diagnosed with SAD. This understanding of the data could impact its potential use, especially in comparison to participants' self-indicated scores on individual aspects of SAD.

On the other hand, the dataset was constructed for ease of comprehension and use by everyday people. The accompanying paper is filled with medical jargon and emphatic research methods and rationales (Fathi, et. al., 2020), so the dataset has been cleaned so that laypeople are able to identify the purpose of each of the columns and their values without much assistance

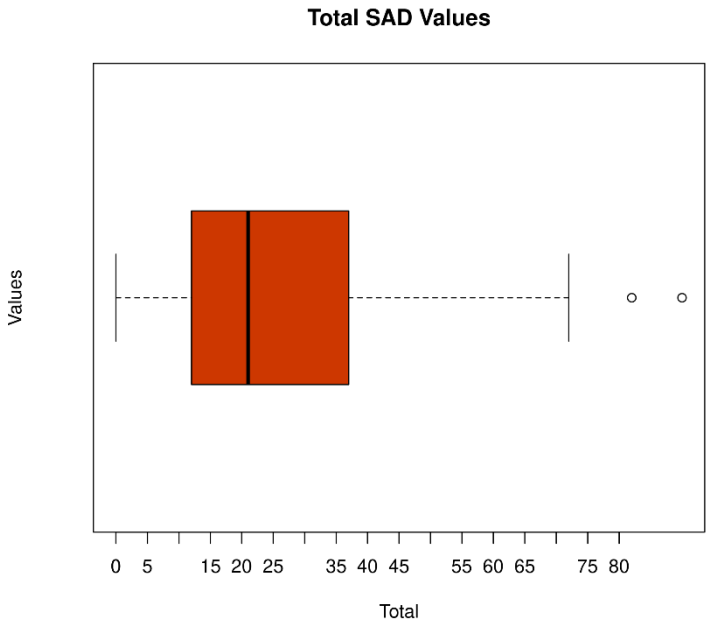
from the supplementary literature. Once cleaned, the data can be used to comprehend SAD on an individual and personal level outside of the medical field.

Data Cleaning

OpenRefine and Excel have been utilized to clean the data for wider use and understanding. There are two versions of the dataset stored. Both contain identical changes in terms of the data included, organization, and the recognized format of data values. Columns containing binary responses represented as 0 or 1 have been converted to ‘no’ and ‘yes’, respectively. Columns with these such responses are ‘HR’ through ‘hasSAD’. A similar change was made for the ‘Gender’ column, which contained 0 and 1 for ‘female’ and ‘male’, respectively. The same procedure was used to change the cell values from ‘Education’. Values 1 through 6 have been changed to ‘High School’, ‘Diploma’, ‘Undergrad’, ‘Bachelor’, ‘Master’, and ‘Post-Graduate’, respectively. And cell values 1 through 5 from the ‘Occupation’ column have been changed to ‘Student’, ‘Faculty Member’, ‘Employee’, ‘Self-Employed’, and ‘Unemployed’, respectively. These changes are based on the original key developed by the researchers (Fathi & Ahmadi, 2020).

All columns containing numeric values have been altered so they are consistently recognized as numeric values rather than textual strings, especially when using coding programs to evaluate the data. All rows have been reordered based on the ‘Age’ column, from youngest to oldest. This is because age may be a common factor in many of the other column values. The dataset can still be freely reordered based on any other factor.

Finally, a new column was added to the dataset based on the study participants' responses to columns 'ATF' through 'TG'. Values from 'ATF' through 'DAF' are reported on a scale 0 through 10, while 'HR' through 'TG' are 'yes' or 'no' responses evaluated based on 0 or 1 values. The sum of these scores for each participant are included in a separate column called 'Total' (0-93), which is meant to show researchers an all-encompassing view of which participants show greater or harsher symptoms of SAD. The 'Total' column can be used comparatively to individual scores for the various aspects of SAD or can help observe opposition to the previous diagnosis of individuals, their Social Phobia Inventory (SPIN) test scores, or their Liebowitz Social Anxiety Scale (LSAS) test scores. The highest total SAD value is 90 while the lowest is 0. The mean of the total values is 26 (25.57). The mode is 19 and the median is 21.



The difference between these two versions is the column names. The first, named SAD_Data_v1.csv, possesses all original column names which are abbreviated based on professional, medicinal, and computational use and understanding. The second version, named SAD_Data_v2.csv, contains columns that have been renamed for ease of comprehension by those inside or outside of the medical field. Users may find it easier to use the data or alter

column names to ones that are most comprehensible to them if they do not need to reference a code key each time they use the data.

Table 1

Relabeled Column Names

Original Column Names	Relabeled Column Names	Original Source Definition
EducationLevel	Education	Education level (1=High School; 2=Diploma; 3=Undergraduate; 4=Bachelor degree; 5=Master degree; 6=Post-graduate
HasFamilyHistory	Family_History	Family history of anxiety or depression (1 = yes; 0 = no)
ATF	Fear_of_Attention	The fear of being at the center of attention (Range=0-10)
EAF	Fear_of_Eating_Around_People	EAF: The fear of eating in front of another person (Range=0-10)
TKF	Fear_of_Public_Speaking	TKF: The fear of speaking in public (Range=0-10)
CMT	Fear_to_Attend_Parties	CMT: The fear of attending parties (Range=0-10)
DEF	Fear_of_Public_Eating	CMT: The fear of attending parties (Range=0-10)
SMF	Fear_of_Contact_with_Strangers	SMF: The fear of meeting or contact with strangers (Range=0-10)
ERF	Fear_to_Enter_Occupied_Rooms	The fear of getting in a room where others are sitting (Range=0-10)
DAF	Fear_of_Disagreement	DAF: The fear of disagreement with strangers (Range=0-10)
HR	Heart_Palpitations	Has heart palpitations (1=yes; 0=no)
SW	Sweating	Has sweating (1=yes; 0=no)
TR	Tremor	Has a tremor (1=yes; 0=no)
DR	Dry_Mouth	Has dry mouth (1=yes; 0=no)
BR	Hard_Breathing	Has hard breathing (1=yes; 0=no)

CK	Feel_Suffocation	Has a feeling of suffocation (1=yes; 0=no)
CP	Chest_Pain	Has chest pain (1=yes; 0=no)
NS	Nausea	Has gastrointestinal discomfort and nausea (1=yes; 0=no)
DZ	Dizzy_Weak_and_Sick	Has a feeling of dizzy, weak and sick (1=yes; 0=no)
UR	Feeling_of_Being_Unreal	Has a feeling of being unreal (1=yes; 0=no)
UB	Fear_of_Losing_Balance	Has a fear of losing balance (1=yes; 0=no)
MD	Fear_of_Being_Crazy	Has a fear of being crazy (1=yes; 0=no)
TG	Numbness_and_Moaning	Has numbness or moaning (1=yes; 0=no)
LSAS	LSAS_Score	The result of the Liebowitz Social Anxiety Scale questionnaire (Range=0-144)
SPIN	SPIN_Score	The result of the Social Phobia Inventory questionnaire (Range=0-68)

Challenges and Strengths

One of the most difficult things about this dataset is that it was originally created for a very particular use—for machine learning and the development of psychiatric care and decision-making. The original data and its provenance are reported in a way that is comprehensible most only to those people that the original researchers assumed would find interest in its contents and for its original intended use. Steps taken to clean the dataset have been done with this in mind in hopes that the data will be more coherent to those outside of the data's original intended use. This data could be extremely insightful and supportive of various fields and individual needs. That is why it is important that the dataset be generalizable to people from various disciplines, organizations, occupations, expertise, and walks of life.

Access, Use, & Re-Use

There are countless potential applications of this data outside of those initial intended uses. Researchers and other users involved or interested in psychiatry, psychology, social work, humanities, anthology, mental and physical health work, counseling, education, or other similar fields may find this data incredibly serviceable. Even those involved in understanding artificial intelligence, machine learning, and other computer science-related fields may find the data's structure and methods of development useful for determining future projects and possibilities. Individuals dealing with mental or physical unwellness for themselves or from people around them may also find this data very informative. The data provides a new set of questions that can help to inform the degree to which individuals are struggling with SAD and can lead people to better understand what resources and assistance struggling individuals need. This is especially true considering the number of undiagnosed individuals who have no knowledge of the disorder that they may be suffering from (Fathi, et. al., 2020). The data can also lead to a better understanding of what symptoms or experiences individuals diagnosed with SAD have. Or better: the *variance* in experiences that these individuals have.

Further exposure of this dataset can lead to the expansion of awareness and research related to this topic. It can encourage similar research that includes new elements such as open-ended or qualitative research. Additional research could also include observing various alternative potential factors in SAD or its physical and mental aspects that can be considered alongside the current study or similar research.

The current dataset is meant to assist in clinical or medical workers' ability to distribute accurate and high-quality care to patients, and to be able to better handle the decision-making

processes involved in diagnosing and tending to patients. It is also conscientious of the "imprecise" natures of symptoms (Fathi, et. al., 2020)—those things that people experience that may or may not be listed in an official, regulatory medical or clinical journal or book. The data is meant to allow machines to consider and accurately approach vagueness of responses. So, the data is easily processed because of the careful consideration taken to create and collect the data.

To stimulate and encourage fresh, unique use of the data, facets were developed in OpenRefine during the cleaning process to allow researchers to access and observe groups or aspects included in the data that might not be immediately obvious or convenient. Multiple facets have been developed to manage the scores between 0 and 10 of characteristics reported by participants for each aspect of SAD. Facets have also been created for when researchers seek scores based on the LSAS or SPIN test scores, or affiliated responses and results to the various other columns. Facets can help researchers in answering questions such as: how responses across scenarios impact or oppose one another on the scale 1 through 10? Are there relationships between participants' reactions to certain scenarios as seen through this scale? Are there scenarios when participants' reactions are seemingly or significantly independent of other reactions? Are there differences in participants' responses to seemingly or significantly similar concepts or situations? Why might these observations occur? What differences exist between those who have previously been diagnosed with SAD and those that have not? A brief look at the data shows that quite a few participants who were previously diagnosed with SAD scored, generally, very low throughout the survey. These are only a few of the numerous possibilities.

Preservation and Storage

The original data is published by Elsevier and stored in its designated repositories (Fathi & Ahmadi, 2020). The newly cleaned versions of the data are available through GitHub (Beal, 2024), which is a free and open access repository. The data could also be kept in a repository such as the Canadian Institute for Health Information (CIHI, 2024), HealthData.gov (U.S. Department of Health and Human Services, 2024), or Open Machine Learning (OpenML, 2024) and attributed back to the original dataset in Elsevier. Health sciences and machine learning are the data's original function, but certain repositories can extend the data to not only these fields, but other relative ones previously discussed.

Transformation

The original dataset, named final_data.csv, as well as the recently cleaned versions, are saved as CSV files for ease of reformatting, transferability, and interoperability of their resources. CSV files are also compatible with various scripting programs which will allow for researchers and other potential users to import the data into systems or programs that allow them to analyze, organize, and visualize the data. The affiliated articles are available in PDF format, so they are consistent and durable in the long-term.

Licensing and Privacy

The original dataset and article follow a CC BY-NC-ND license (Elsevier, 2024). The article is openly accessible and ensures that authors and creators receive proper attribution and credit for their published work. Where the cleaned versions are available in GitHub, they follow a CC0 license, so there are no restrictions on its use.

References

Beal, R. (2024). SAD-Data. *GitHub*. https://github.com/BealRL/SAD-Data?tab=readme-ov-file#readme

CIHI. (2024). About CIHI. https://www.cihi.ca/en/about-cihi

Elsevier. (2024). Copyright: Overview. https://www.elsevier.com/about/policies-and-standards/copyright

Fathi, S. & Ahmadi, M. (2020). Data for: Development and use of a clinical decision support system for the diagnosis of social anxiety disorder. *Mendeley Data, 1*. [Data set]. doi: 10.17632/4jycfwhb4g.1

Fathi, S., Ahmadi, M., Birashk, B. & Dehnad, A. (2020). Development and use of a clinical decision support system for the diagnosis of social anxiety disorder. *Computer Methods and Programs in Biomedicine, 190.* https://doi.org/10.1016/j.cmpb.2020.105354

OpenML. 2024. OpenML: A worldwide machine learning lab. https://www.openml.org/

U.S. Department of Health and Human Services. (2024). Welcome to HealthData.gov. *HealthData.gov.* https://healthdata.gov/