UNIVERSITY OF CALIFORNIA

Los Angeles

Probably Approximately Correct Learnable Fuzzy System

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Yan Wang

2019

ABSTRACT OF THE DISSERTATION

Probably Approximately Correct Learnable Fuzzy System

by

Yan Wang
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2019
Professor Honghu Liu, Chair

This dissertation develops the probably approximately correct (PAC) learnable fuzzy system to predict clinical outcomes from a small number of survey questions (short form). There are five layers in the system: input, fuzzification, inference, defuzzification, and production. The major product in this dissertation is to derive the PAC learnable knowledge-driven machine learning algorithm by growing sample using Bootstrap samples with Gaussian distributed noise. The input layer is the procedure for preparing data input. In the fuzzification layer, sample size is significantly increased using bootstrap re-sampling with replacement. The fuzzy set with proposed membership function is generated by introducing Gaussian distributed noise to survey responses of the bootstrap samples to reflect uncertainty. This is a natural language extension from the point option in survey questions to region input with probabilities from survey design space. The inference layer includes both classification and prediction. Here we use machine learning techniques to derive the algorithms in this layer, e.g. Naive Bayesian method and eXtreme Gradient Boosting (XGBoost). The final predicted values require a defuzzification process in the next layer to remove noise in prediction. There are four types of input after fuzzification, original input, fuzzy input, input required interpolation and input required extrapolation. The defuzzification process is based on weighted means of related information. The last step of the system is the output layer with algorithms, final prediction and validation internally and externally. Lastly, we apply this fuzzy system to derive PAC learnable algorithms to predict oral health clinical outcomes. The input predictors include short forms and demographic information. The

short forms, developed from Graded Response Models in Item Response Theory, have two versions (children and their parents). The clinical outcomes are referral for treatment needs (categorical) and childrens oral health status index score (continuous). The prediction is evaluated internally and externally by sensitivity and specificity of a binary variable, correlation (between original value and predicted value) and root mean square error (RMSE) of a continuous variable. Both internal and external validation show the improvement of prediction when new information is added and generalizability as well as the stability of the algorithm. The best prediction (high sensitivity and relatively high specificity for categorical variables, low RMSE and high correlation) is reached when using child's self-reported short form, plus parent's proxy-reported short form, and demographic characteristics.

The dissertation of Yan Wang is approved.

Thomas R. Belin

Catherine Crespi-Chun

Ronald D. Hays

Honghu Liu, Committee Chair

University of California, Los Angeles

2019

The entire work dedicated to my family, who supported me through the doctoral journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to take the opportunity to thank my advisor Dr. Liu, for his tremendous support and suggestions. I appreciate all my committee members, Dr. Belin and Dr. Crespi for their statistical inputs to my draft. I want to give special thanks to Dr. Hays for not only serving as one of my committee members and providing comments and edits to my manuscripts.

I would also like to take the opportunity to thank everyone in the "PROMIS" oral health group. Dr. Liu is the PI of the project funded by NIH. Dr. Marcus, Dr. Maida, Dr. Coulter, Dr. Crall, Dr. Spolsky, Dr. Lee, Dr. Ramos-Gomez, and Dr. Shen are coauthors for all the publications from this group and showed me how to conduct dental research in practice.

I would like to thank Qing Yang, Wendy Shih, Mengwei Ko, and Di Xiong, who are not only friends but research colleagues, for their support and love during the entire period of dissertation writing.

And many thanks to those who convinced me that I can finish this work gradually. Deeply appreciate the love and support of my family.

VITA

2005	B.A. in Mathematics and B.S. in Computer Science, Beijing Normal University, China
2010	M.S. in Biostatistics, University of California, Los Angeles
2012	Founder and President, Elite Biostats Corp, Los Angeles
2016	Food Studies Graduate Certificate, University of California, Los Angeles
2017	Mary G. and Joseph Natrella Scholarship
2019	GATHER Post Doctoral Fellowship

SELECTED PUBLICATIONS

Wang, Y., Hays, R., Marcus, M., Maida, C., Shen, J., Xiong, D., Lee, S., Spolsky, V., Coulter, I., Crall, J. and Liu, H., 2018. Developing Childrens Oral Health Assessment Toolkits using Machine Learning Algorithm. *JDR Clinical & Translational Research*. Under Revision.

Wang, Y., Yu, W., Liu, S. and Young, S.D., 2019. The Relationship Between Social Media Data and Crime Rates in the United States. Social Media+ Society, 5(1), p.2056305119834585.

Wang, Y., Hays, R., Marcus, M., Maida, C., Shen, J., Xiong, D., Lee, S., Spolsky, V., Coulter, I., Crall, J. and Liu, H., 2018. Development of a parents short form survey of their children's oral health. Int J Paediatr Dent. 2019; 29: 332– 344.

Liu, H., Hays, R., **Wang, Y.**, Marcus, M., Maida, C., Shen, J., Xiong, D., Lee, S., Spolsky,

V., Coulter, I. and Crall, J., 2018. Short form development for oral health patient-reported outcome evaluation in children and adolescents. Quality of Life Research, 27(6), pp.1599–1611.

Marcus, M., Maida, C.A., **Wang, Y.**, Xiong, D., Hays, R.D., Coulter, I.D., Lee, S.Y., Spolsky, V.W., Shen, J., Crall, J.J. and Liu, H., 2018. Child and Parent Demographic Characteristics and Oral Health Perceptions Associated with Clinically Measured Oral Health. JDR Clinical & Translational Research, 3(3), pp.302-313.

Liu, H., Hays, R.D., Marcus, M., Coulter, I., Maida, C., Ramos-Gomez, F., Shen, J., **Wang, Y.**, Spolsky, V., Lee, S. and Cai, L., 2016. Patient-Reported oral health outcome measurement for children and adolescents. BMC oral health, 16(1), p.95.

Wang, Y., Crane, H.M., and Liu, H. Conditional Maximum Likelihood Rasch Model in Data Harmonization. *2014 JSM Proceedings*, Nonparametric Statistics Section. pp. 464–473.

Wang, Y., Rosen, M.I., Shen, J., Moore, B.A., Daar, E.S., and Liu, H. Intensity Estimation for Poisson Processes used to Model a Real-time Medication Event Monitor, Contributed paper. *2014 JSM Proceedings*, Joint Research Conference. pp 4558–4567.

Liu, H., Marcus, M., Maida, C., **Wang, Y.**, Shen, J., and Spolsky, V. (2013) Predictive power of the severity measure of attachment loss for periodontal care need. *Journal of Periodontology*, vol. 84, no. 10, pp. 1409–1415.

Gironda, M.W., Maida, C., Marcus, M., **Wang, Y.**, and Liu, H. (2013). Social Support and Dental Visits. *Journal of American Dental Association.* JADA 144, 188–194.

Wang, Y., Ong, M., and Liu, H. Compare Predicted Counts between Groups of Zero Truncated Poisson Regression Model based on Recycled Predictions Method. In: *2011 JSM Proceedings*, Statistics in Epidemiology Section. pp. 2478–2487

CHAPTER 1

Background

1.1 Introduction

1.1.1 Background

In this dissertation, we develop a fuzzy system that is based on the input value of survey responses to predict clinical outcomes. The motivation of the study is the vagueness of linguistics variables. During the survey design, response options are designed with a variety of cut-off points to measure a potentially continuous variable. For example, frequency might be categorized into six options: always, almost always, often, sometimes, almost never and never. When the participants answer the survey questions, they provide a response on the categorical response scale that they judge to best represent their underlying position. This categorical process could be misleading because of different cut-off options and the vagueness between options, e.g. "always" and "almost always". Items with more quantitative response options are not necessarily better than more qualitative response options [1, 2]. In this dissertation, we propose fuzzy membership functions of the original responses with probabilities (membership function) to fuzzify the input values in surveys options. The fuzzy system does not only expand original observations plausibly but also increase uncertainty by adding noise to original survey options. In this way, we take into account the uncertainty coming from both randomness and vagueness of languages. We will explain this uncertainty next.

In linguistics, the meanings of all terms have a lesser or greater degree of vagueness [3, 4]. The boundary of any term, e.g. any selection in a survey question, is never a point but a region, where the term can move with probability from 0 to 1. Fuzzy set theory is a formal

way of dealing with such vagueness in natural language. In survey selections, the negative markers (e.g. never), adjectives (e.g. good), and adverbs (e.g. very often) are assessed empirically in fuzzy sets. The survey is a process in fuzzy set theory of transmission of quantitative information between the respondents and survey designer. Recently, there is great interest in combining linguistics and computer science to study the role of vagueness in natural language and quantification of meaning, which is very hard to measure precisely. The first attempt of quantifying the meaning of a word was in 1941 [5], defining the meaning of a word as a formula of three components. The constant component is overall meaning over people and over context. The random components represents the variation in the meaning. One part is the variation due to context and the other part is the variation due to the individual. The three parts together represents the meaning of the word. The assumption is that there is a unidimensional meaning of the word that is continuous. The individual variation and context variation are independent of this unidimensional meaning [5]. In the present work, adding random noise to individual cases is intended to reflect that there will be cases in the population that differ somewhat from observed individuals, similar to the way that a given word can have shades of meaning in different settings.

The input observations are usually not comprehensive enough to cover all possible combinations in the survey design space. It is impossible to collect samples that can cover all possible combinations of the item options. A common rule of thumb when using Item Response Theory (IRT) is that we need at least three to five subjects to endorsed each response option in order to yield a stable threshold parameter estimate [6, 7] or using rule of thumb at least using 10 subjects per parameter [8]. The traditional way in IRT theory is to combine the categories with lower endorsement. The naive goal for the fuzzy system is to grow enough sample size to cover the response space and therefore to derive the machine learning algorithms from the available resources to predict the new input. The main aim of the algorithm is to predict the outcome with some level of uncertainty and some level of confidence from new input of the survey. We show the algorithms derived from fuzzy system on certain hypothesis spaces are Probably Approximately Correct (PAC) learnable. The new input may or may not come from the sample population.

Fuzzy random variables were first introduced in 1986 [9] as a generalization of random variables. The uncertainty comes from two important sources, randomness (stochastic variability of all possible situations) and fuzziness (no clear boundaries of parameters) [10], where randomness is for future and fuzziness is for the past and its implication. The fuzzy random variables are a combination of fuzziness (possibilities) and randomness (probabilities) that are naturally compatible. In the context of oral health, the fuzziness and randomness can be explained in distributional differences and membership differences in the context of survey questions [11, 10].

1.1.2 Motivation

The motivation of the study is to develop an algorithm that supplements available data with plausible representations of other values in the underlying population in a way that can be expected to yield the greatest possible predictive accuracy from fitted statistical models, while representing both sampling variability and vagueness in the contextual meaning of latent measurements. The algorithms are from a pre-determined hypothesis space that ensures the prediction can be generalized with certain level of confidence. This is a process in between supervised and unsupervised learning. Besides the efficiency of learning, the supervised learning focuses on the predictive accuracy of a model, which is the most important quality criteria [12]. The predictive accuracy is usually evaluated by the loss function $l(\cdot)$, defined by the difference between expected value and observed value. The supervised knowledge-driven learning approach focuses on the ability to make accurate predictions of so far unseen inputs, rather than to discover the local patterns or associations. The ability of learning from experiences and adapting to new situations, is the integral part of artificial intelligence (AI) [13, 14]. Knowledge-based learning is the bottle neck of AI.

Besides knowledge-based learning, in practice, the boundary between the classifications is not always strict and clearly defined [12]. For example, there is no clear cut-off clinically between urgent need of dental care and necessary early attention. The overlap across categories are very natural (there is no clear cut-off between often and sometimes) and even sometimes

counter-intuitive. For survey response, the boundaries between different categories are often smooth and unclear. The responses of, for example, "often" and "always" are more similar than the ordinal score 2 and 3 assigned to the options. In fuzzy analysis, the same original survey response could be assigned to adjacent categories with membership function. The fuzziness between response options allows the survey options to cover the entire real line of the underlying trait. It is an extension of the classical set theories, where the elements either belong or not belong to the set. In fuzzy set theory, the membership function is used to extend classical set theory, either or not belong to the set, to robust belonging in terms of probabilities, which model the reality better than classical set theory [15]. In the past, the uncertainties of reality were modeled as randomness by probability theory in statistics. We introduce the uncertainty at the data level by generating fuzzy random variables with both randomness and fuzziness. The uncertainty comes from two level fuzziness and randomness [10]. The inferences are made through fuzzy system.

1.1.3 Scientific approach

We develop this fuzzy system based on available resources of observations to extend the information to a certain degree to its boundary. The prediction is not based on only available limited observations, but based on the observations with membership functions defined in the fuzzy set. The widely accepted definition for Knowledge Discovery in database (KDD) is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable structure in data [16]. The core part of KDD is data mining and exploring with an application of a discovery of patterns and associations and eventually knowledge, and automated learning process as well as AI.

In the context of the study, field test data are collected from two metric spaces, the survey response space and the clinical results space. We aimed to predict the results from clinical space from the survey response space. Manually introduced noise is used to generate the fuzzy grid for each response to oral health related survey items based on probabilities. This fuzzy grid is defined on the survey response space according to possible responses from survey

design. On the grid, the original response is expanded throughout the designed options in a random manner, the shorter distance from original option, the higher the probabilities that original value may shift. In this manner, we increase the sample size by adding noise to bootstrap sample to cover the uncertainty.

The uncertainty of final prediction of the outcomes is from three parts: the probability of final machine learning algorithm being calculated, the fuzzification process of the soft boundary among survey responses, and the bootstrap sampling of original observations (crisp data), from which sample size significantly is increased. The learning process is fuzzy and unstable, due to the manually introduced variation in fuzzy data. The possible rule under this is that if subject A's survey response is similar to subject B's survey response, then the clinical result of A is also similar to the clinical result of B. That is the rule of "*similar objects have similar class labels*" [12]. This is the most fundamental assumption of using survey questions and machine learning algorithms to predict clinical outcomes, that is to link self-perceived health and clinical determined health. A crucial ingredient of the present investigation is the availability of "gold-standard" information from a clinical oral health examination that can be used to evaluate predictions from statistical models.

1.2 Organization

In the following, the methodology part is organized as the flow of the proposed fuzzy system shown in Figure 1.1, from input of observations to output of predictions. The fuzzy system includes the process of fuzzification and defuzzification. Machine learning algorithms are derived during the inference layer. After the introduction of the entire flow of the fuzzy system, theoretical proof is provided to show the fuzzy system is Probably Approximately Correct Learnable in large samples. Figure 1.1 provides more details of the connection between different layers. In the fuzzification layer, including the process of growing original sample observations (crisp set) using bootstrap methods with introduced noise. We added Gaussian noise to bootstrap samples.

In practice, the bootstrap step can be extended to sample sizes larger than the original

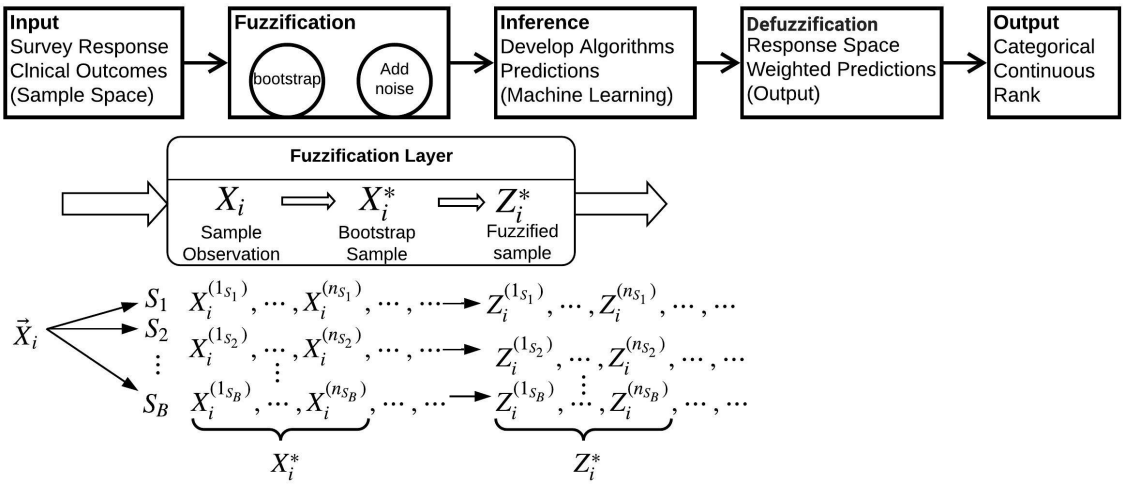

Figure 1.1: Flowchart of the fuzzy system

sample size, and fuzzification can be understood as adding noise to the bootstrap samples. The process thus gives rise to replicate data sets as representations of a broader population that have the potential to yield either better or worse predictive accuracy for statistical modelling purposes depending on the relationship of the additional sample size to the added noise. A goal of the investigation that follows is to gain insight into the corresponding trade offs between sample size and added noise. The larger additional sample size and additional noise may move the original sample with great degree of uncertainty. We will show this by a figure in the following section.

CHAPTER 2

Fuzzy System

2.1 Input Layer

2.1.1 Sample space

Sample space $\mathscr{X}$ consists of the observations collected from the survey response space $\mathbb{R}$. Usually $\mathscr{X} \subseteq \mathbb{R}$, in practice, it is impossible for sample space $\mathscr{X}$ to cover all possible combinations in $\mathbb{R}$. In our scenario, the sample space is the survey responses from children and parents $\mathscr{X}_C$ and $\mathscr{X}_P$, i.e. $\mathscr{X} = [\mathscr{X}_C, \mathscr{X}_P]$. For convenience, we will use the vector $\vec{X}_i, i = (1, \cdots, n)$ to represent the input variables from each family, i.e. the survey response from the subjects (the participating families) $1, \cdots, n$. The corresponding output variable is denoted by $\vec{Y}_i$, i.e. for each input of survey response, there are corresponding clinical outcomes in the training sample. The outcomes are selected as the categorical variable, continuous variable and rank variable to illustrate the methods.

The sample space $\mathscr{X}$ is defined by the observed (or is labeled by function f in machine learning theory) input vectors (survey responses) and output vectors (outcomes) for n subjects as,

$$\begin{bmatrix} \vec{X}_1 & \vec{Y}_1 \\ \vec{X}_2 & \vec{Y}_2 \\ \vdots & \vdots \\ \vec{X}_n & \vec{Y}_n \end{bmatrix} \tag{2.1}$$

Each vector $\vec{X}_i$ represents a family, including the responses from Child's input and Parent's input about child's oral health [17, 18]. We have previously constructed two short forms independently for children [17] and for parents [18]. In PROMIS literature, short forms are

defined as a fixed set of 4-10 items or questions for one domain. More details of the two short forms are discussed in the example (chapter 3 Data Application part). The vector $\vec{X}_i$ includes short form survey items as well as demographic questions [17, 18, 19]. The sample space generated by these vectors is finite (defined on $\mathbb{R}^{p\times d}$) due to the number of survey questions (denoted by p) are fixed and the response options are also fixed (maximum d) as shown in Figure 2.1. The input vectors determine a grid level space with 2 dimension. The survey space R is finite with $p \times d$ dimension. We include an example of input vectors X_i on Figure 2.1 as $x_{ij} = k$, for subject i answered question j with category k, where $i = (1, \cdots, n)$, $j = (1, \cdots, p)$ and $k = (1, \cdots, d)$. The samples are subjects with different lines on the surface R. The clinical outcomes are parameters from another dimension θ. The problems can be stated as using the survey response surface R to predict θ with a smooth function. The samples are lines from the response space R.

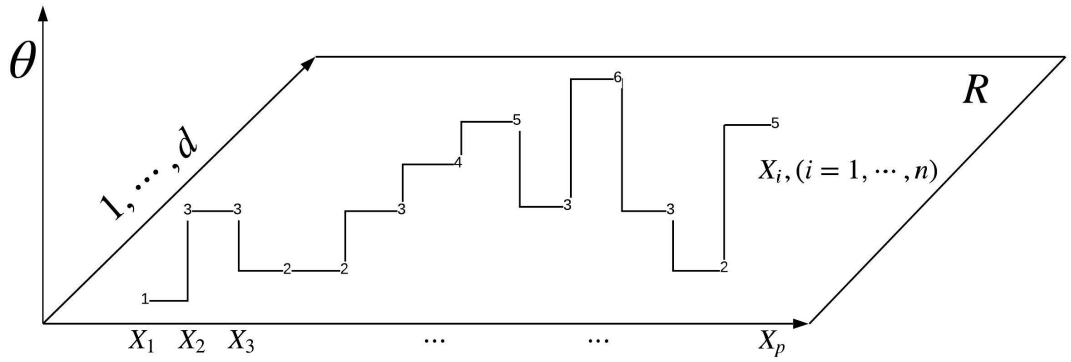

Figure 2.1: Survey response space and θ

Figure 2.1 is a visualization of the relationship between clinical results θ and survey response R, i.e. they are not in the same space. It is possible to use the vectors on R to estimate the corresponding θ for each individual. However, the estimation is never perfectly accurate. In application, it means the survey results can never replace the *"real"* clinical examination. For any algorithms that are developed to link the two space or estimate the relationship between θ and R, the estimation of error should be under consideration.

As d is categorical, it is hard to cover all the points on R. For simplicity, we can assume all questions have the same number of response categories $d_j = d$. The X-axis is the list of

all survey items from $1, 2, \cdots, p$ with options as the score of Y-axis. Assume the options are bounded by d, without loss of generalizability (WLOG). Each input $\vec{X_i}$ corresponds to the output (clinical outcomes) variables $\vec{Y_i}$.

In this entire document, we acknowledge the fact that the survey result could be used to estimate the result of the clinical outcome, but with some error Ξ that is unknown and can never be fully eliminated. We assume,

$$\vec{Y} = f(\vec{X}) + \vec{\Xi} \tag{2.2}$$

The best scenario in this dissertation is to find the function or algorithm that can be close to the result of f. The complete description of a real link function f and system requires far more detailed and elaborate data source, which is usually a challenge in practice.

Definition 2.1. (Realizable) The space $\mathcal{D}$ is **realizable** by hypothesis $\mathcal{H}$ if there exists an optimal $h^* \in \mathcal{H}$ such that the error function is zero, i.e. $error(h^*, \mathscr{D}) = 0$.

Based on above, we know the clinical outcome space is not realizable by any hypothesis $\mathcal{H}$ from survey response space.

2.1.2 Vagueness in Linguistics

In linguistics, the meanings of all terms have a lesser or greater degree of vagueness [3, 4], e.g. the meaning of "*very good*" oral health. This idea was first brought by Labov and Lakoff in 1973 [20, 21]. The boundary of any term, e.g. any selection in a survey question, is never a point but a region, where the term can move with probability from 0 to 1. Linguistics variables were defined by Dr. Zadeh as those variables whose values are not numbers but words or sentences in a natural or artificial languages [4]. The language is vague when describe the different scenario among pre-determined options. The problem of vagueness is a property of natural language .

The first attempt of quantifying the meaning of a word was in 1941 [5]. The meaning of the word M is defined as,

$$M = x + i + c \tag{2.3}$$

In 2.3, the constant component of the word over people, over context is denoted by x. The meaning of the word varies due to individuals is denoted by i. The meaning of the word varies due to context is denoted by c. The data supported the idea that assign each word to a scale value along a unidimensional continuum and the variation in meaning about this scale value was normally distributed [3, 22].

The vagueness of a word implies [3] the variability of using the term by a group of users. Black in 1937 first described vagueness quantitatively using consistency profile. The consistency of using a term T to an element s is defined as,

$$C(T, s) = \lim_{M \to \infty, N \to \infty} \frac{M}{M + N} \tag{2.4}$$

In this definition, M is the number of judgments that T applies to s and N is the number of judgments that not T applies to s. The range of the consistency is from 0 to 1, the probability of T to s. The most doubtful case is 0.5. That is, in this case, the greater vagueness of term T, the more likely that consistency of the profile is close to 0.5.

2.1.3 Linguistics and fuzzy set theory

We will use the simple linguistics concept of temperature (only three levels, cold, warm and hot) to connect with the fuzzy set function. In Figure 2.2, we describe the temperature of the water using a survey item with three categories, cold, warm and hot. The vertical line crosses the functions at about $0.8, 0.4, 0$ may correspond to language that describes temperature as fairly cold, slightly warm and not hot as shown in 2.3 [23]. This is the vagueness of the language. The randomness comes from the percentage of cold, warm, and hot in the population. The fuzziness opens the boundary among cold, warm, and hot.

This is similar to Item Response Theory (IRT) for ordinal response, e.g. Graded Response Models (GRM). The item characteristic curves (ICC) are based on the normality assumption. We combine the normality assumption among the different survey responses and the vagueness in survey process. In this fuzzy system, the membership function is a step function that categorizes the Gaussian membership function, which will be described in detail in the fuzzification layer.

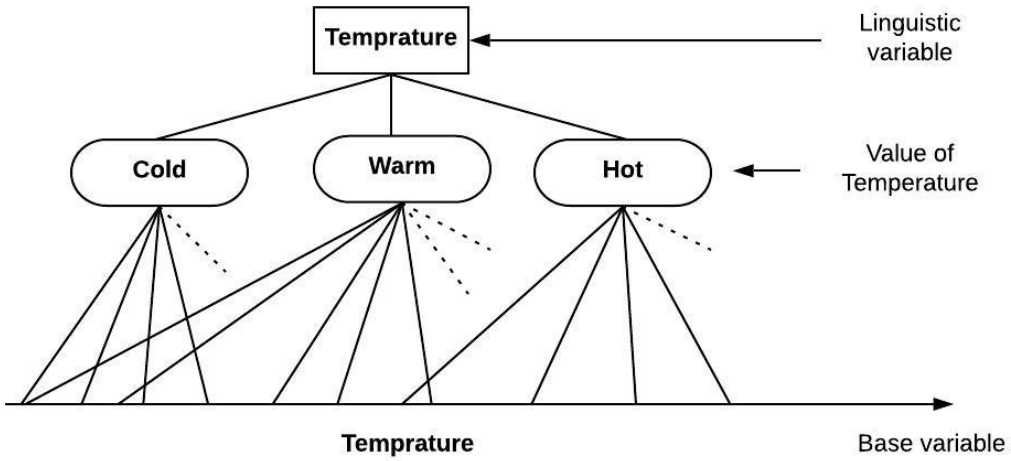

Figure 2.2: The linguistics variable structure

Fuzzy set theory is used when we need to model the uncertainty more than just probabilities and when we face the problems of gradual boundaries, i.e. the boundary of information is not clearly cut. For example, words such as good (oral health), happy (with teeth look), are fuzzy. The concept of good oral health has no clear boundary. Fuzzy set theory is an extension of the classical set theory (CST), where the elements of the set are associated with membership functions. In reality, there is fuzzy knowledge, which involves human thinking and human reasoning, for example, the knowledge and information we learned from the survey response space. Fuzzy set theory deals with unreliable, incomplete and often uncertain information.

The survey questions are designed to incorporate the idea of fuzzy input. The crisp set in CST for a survey question only has yes and no response, which is insufficient to describe human reasoning. It has clear and sharp boundaries. With the fuzzy idea, the survey questions might have these response options: *Very true*, *somewhat true*, *neutral*, *somewhat false*, *false*, and even *do not know*.

An example of using fuzzy logic in reasoning, from Aristotelian logic to inductive logic [24], in the context of oral health, the statements could be formulated in the following as an example, the information from responses a survey item is summarized in A_2. The oral

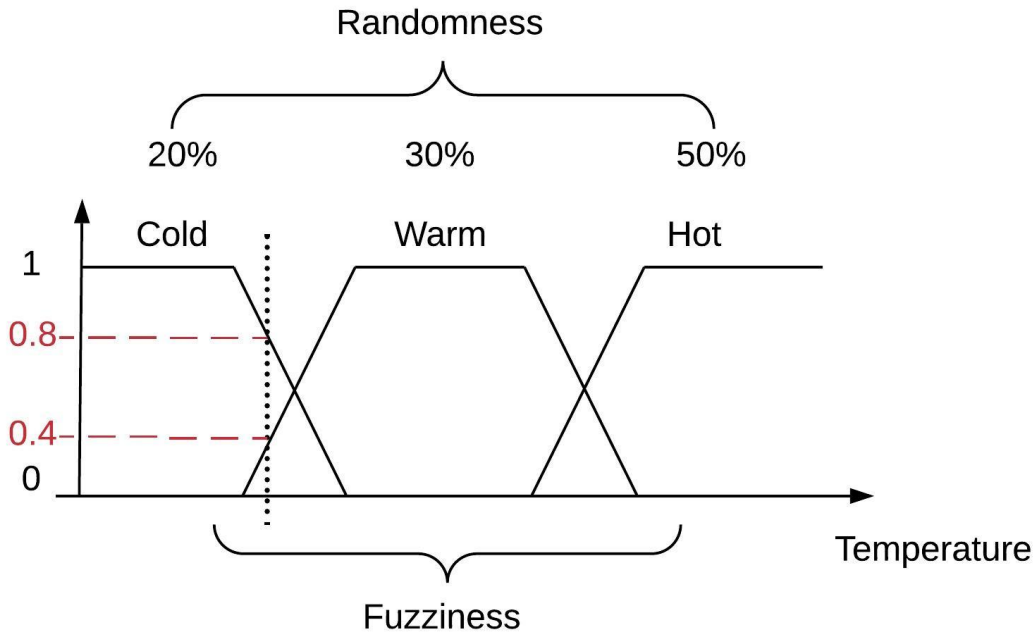

Figure 2.3: The fuzzy function of temperature

health knowledge beyond the survey item is stated is A_1. The fuzzy logic leads to possible conclusion in both A_3 and A_3'. There is no sharp boundary between A_3 and A_3'.

- A_1: Brushing teeth every day leads to good oral health.
- A_2: David brushes his teeth every day.
- A_3: It is likely that David has good oral health.
- A_3': It is very likely that David has good oral health.

Both A_3 and A_3' could be the approximate conclusions of A_1 and A_2.

2.2 Fuzzification Layer

2.2.1 Bootstrap re-sampling stage

The Bootstrap was first proposed by Efron in 1979 [25, 26, 27] to represent the sampling distribution of complex statistics by using simple random samples with replacement from

the original sample. It is commonly used to estimate the sampling distribution of a test statistic [28]. It is also used in survey re-sampling for small-area estimation [29] that is to produce estimates for smaller geographic areas and subpopulations, for which large samples are not available. However, developing valid bootstrap procedures is a challenge. For small-area estimation, a frequentist predictor, such as an empirical best linear unbiased predictor (EBLUP), or a hierarchical Bayes estimator is often used. Developing valid bootstrap procedure survey samples [29] is challenging because of the complex correlation structure induced by survey design, sampling weights, stratification, imputation schema, and small-area estimation.

The bootstrap is a resampling procedure. Efrons bootstrap sample is defined as a sample of n independent identically distributed random variables $X_1, \cdots, X_n$. The parameter $\hat{\theta}$ is estimated by the empirical distribution F_n, which is assign probability $1/n$ to each observation X_i. Usually, there are two numbers to determine before the procedure, the total number of bootstrap samples B and the number of observations in each bootstrap sample n^*. Usually, the sample sizes of all bootstrap samples are the same and most commonly $n^* = n$, when the aim of using bootstrap is to estimate the sample mean and variance.

For finite population of $N << \infty$, labeled as $\{1, 2, \cdots, N\}$, let $\vec{X}_i$ be the vector of survey responses for $i = (1, 2, \cdots, N)$. We are interested in estimating a nonlinear function of $\vec{X}_i$. We can estimate

- A smooth function of the finite population mean, $\theta = g(\bar{X})$, where $\bar{X} = \sum_{i=1}^{N} \vec{X}_i$
- A smooth function of a vector of function of means, $\theta = g(\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_p)$, common statistics, such as variance, ratio, and correlation can all be written in this format. Note the above situation is a special case of this scenario;
- Nonsmooth function (e.g. quantile).

In traditional randomization survey sampling theory, the response space $X \in R$ is fixed. The unbiased estimator is calculated based on the probability that a response is selected

from the finite population. This is called design unbiasedness, with respect to the sampling design probabilities [29].

In survey sampling, it is important to evaluate the bias of an estimator. For a linear function, the bias of an estimator can be estimated directly. The complex function is expressed in the linear format using Taylor Series. For other cases, when the function is not explicitly expressed in math formula, or for non-linear functions, the bootstrap is used. The common bootstrap steps are [29],

1. Generate resamples from original sample using suitable probability sampling scheme.
2. Calculate $\hat{\theta}_b^*$ from the resamples, here $\hat{\theta}_b^*$ is a nonlinear statistic.
3. Calculate $\hat{\theta}$ from a large number (B) of independent resamples as

$$\hat{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* \tag{2.5}$$

The bias for the bootstrap estimator $\hat{\theta}^*$ is,

$$\text{Bias} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*) \tag{2.6}$$

The variance is estimated by,

$$\text{Variance} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}^*)^2 \tag{2.7}$$

To obtain a confidence interval (CI), one can use the percentile method [30]. It is not appropriate to simply calculate CI using the parameter estimate and variance. Usually, the percentile method works fine to construct a CI for a parameter based on bootstrap estimates. Rank all bootstrap estimates $\hat{\theta}_b^*$. The 90% CI excludes the highest 5% and the lowest 5%. The percentile method is referred to as Efron's percentile method CI. The CI in this method contains the true value of θ with probability $1-\alpha$ but it has some restrictions on the estimator, The Typical Value Theorem only needs to assume the population distributions is symmetric.

For this problem, the true population parameter is both continuous and categorical. The continuous distribution is the childrens oral health status Index (COHSI) score and its percentile. Both COHSI score and percentile have possible ranges from 0 to 100. The percentile is similar to rank value and has lowest and highest percentile. However, in practice, it is impossible to observe extreme values of COHSI. The lowest COHSI happens if a child lost all primary teeth or an adolescent lost all permanent teeth. This rarely happened in practice. The estimation for the population COHSI range is not available in the literature, though the observed range of the COHSI in the field test was 59.18 to 100 [31, 32, 33, 34]. The limitation is that when field tests were conducted at dental clinics among those children who already have a dental home [35] (mostly preventive care), COHSI is highly skewed to the left (negative skewed), with a higher sample mean than the general population. A higher COHSI score was associated with better oral health [36, 31]. Most of the children have better oral health status than the general population. This limits the generalizability of machine learning algorithms to predict oral health of the general population. If the algorithms are derived from a biased sample (Figure 2.4), no matter how good the algorithm is on the training set without overfitting problem, the generalization is still limited.

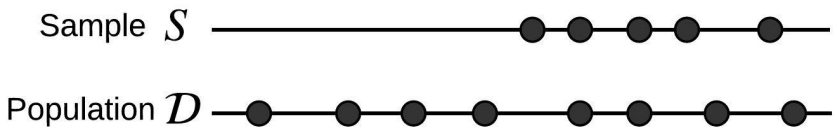

Figure 2.4: Biased sample in population

In our data, the missing or low endorsement of lower categories (the survey options of lower assigned score associated with poor oral health) makes most traditional estimation methods biased. For example, the graded response model needs at least three to five cases per response options in order to estimate threshold parameters. We can combine adjacent categories to increase the sample size. The dental clinic sample makes the lower level of categories (usually predict poor oral health status) hardly being endorsed by those who already have a dental home. Therefore, the distribution of the sample is skewed with a large portion of the sample having better oral health status than population. The combination

of lower adjacent categories reduces the number of threshold parameters and can lead to problems in generalization to other samples.

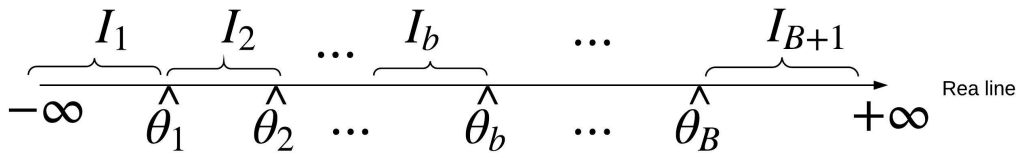

Figure 2.5: Partition real line into $B + 1$ section

2.2.2 Adding noise stage

In this dissertation, to address situations, such as in Figure 2.4, where a re-sampling method would not be expected to represent the full range of values in an underlying population distribution. we developed a technique by adding noise to the original sample so that the sample can better reflect the range of the population. This approach requires knowledge about the population to be estimated. In the scenario of the motivating example, we know the sample is collected from dental clinics with better oral health than the general population. Suppose further that the goal is to generalize to a school-based population, as shown in Figure 2.4. A natural way to expand the original set observations to a broader set of plausible observations is adding noise to fuzzify the sample so that the boundary of the fuzzy set could reach to lower or upper limit of population. The more different the sample is from the population, the more ambitious the added noise may need to be to extend the boundary of original sample, and the more uncertainty of the prediction.

A natural way in statistics to add noise to a sample is with a standard normal distribution, as shown in an illustrative example in Figure (2.6). The figure indicates the probability of transferring X_i symmetrically into six neighbor options $X_i \pm 3$ and itself X_i until the values hit the boundaries of the response options. Figure 2.6 corresponds to the membership function in Figure 2.7.

In this section, we show that the manually introduced noise extended the original data set or crisp set into a fuzzy set with probability defined by the fuzzy membership function,

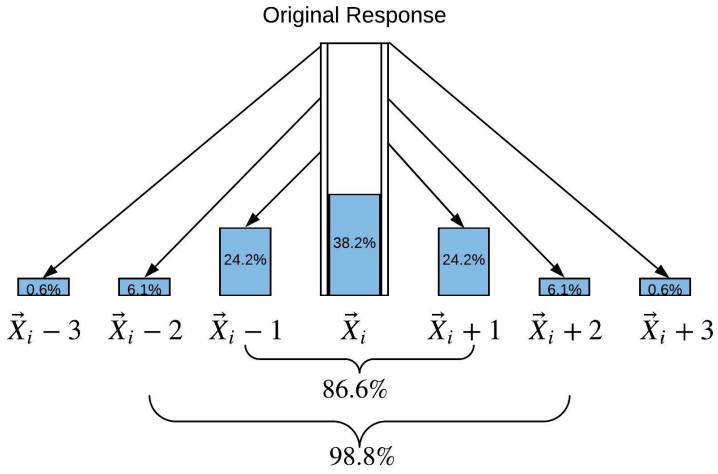

Figure 2.6: The membership function of original response with probabilities

as shown in Figure 2.7. Each unique response forms a crisp set in fuzzy set theory, which belongs to the fuzzy set that is generated by the unique response with probability 1. The noisy set generated by this response x_i is the fuzzy set A_{x_i} defined by,

$$A_{x_i} = \{a_i | a_i \in R, a_i = [x_i + \epsilon], \epsilon \sim N(0,1)\} \tag{2.8}$$

Here $[\cdot]$ means rounded to the nearest integer. This integer is bounded by the response space R. The membership function is restricted on the boundary of the sample space. For example in Figure 2.8, the noise dilutes the extreme response to the neighbor categories. For example, for $x = 1$ in a six response category, after adding the noise, 69.1% of the original response stays at category $X = 1$, the original response. As shown in Figure 2.8, 30.9% of the original responses of $x = 1$ are now distributed as $x = 2$ (24.2%), $x = 3$ (6.1%), and $x = 4$ (0.6%). When the response is at the boundary, the responses can only be distributed to one direction. This is the only restriction when we introduced noise.

As shown in the graph, the extreme response is forced by the manually introduced noise to reallocate to its neighbors by roughly 30%. The nearest neighbor category is increased by 24.2% from the original response category of $X = 1$.

The membership function of a_i is defined by the standard normal distribution of ϵ, i.e.

$$P(a_i \in A_{x_i}) = Z_\epsilon \tag{2.9}$$

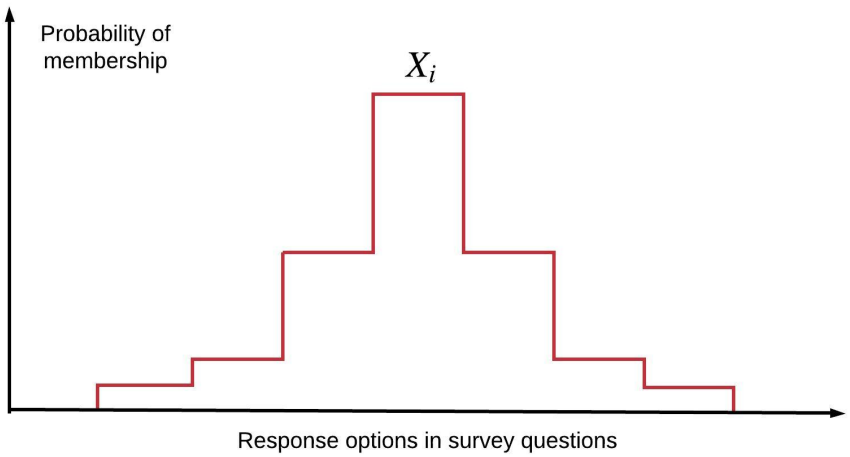

Figure 2.7: The membership function for X_i

This is a categorized version of the Trapezoidal membership function.

For the simplest case, for only one survey item $X \in R^1$ the noise ϵ transfer the response space from X to X', where $X' = X + \epsilon$. Assume X is Multinomial distribution with six level of categories, i.e.

$$X \sim \text{Multinomial}(n, p_1, p_2, \cdots, p_6), \text{ where } \sum_1^6 p_i = 1$$

The distribution for X' is given by,

$$X' \sim \text{Multinomial}(n, p'_1, p'_2, \cdots, p'_6), \text{ where } \sum_1^6 p'_i = 1$$

The probability p'_i can be calculated from the matrix,

$$p'_1 = 0.691p_1 + 0.309p_2 + 0.067p_3 + 0.006p_4$$

In the above formula, the coefficients are directly from the transition matrix as shown in Table 2.1 in the first column.

In vector format, the relationship between p' and p (where transition matrix is defined as Table 2.1) is:

$$p' = T^t p$$

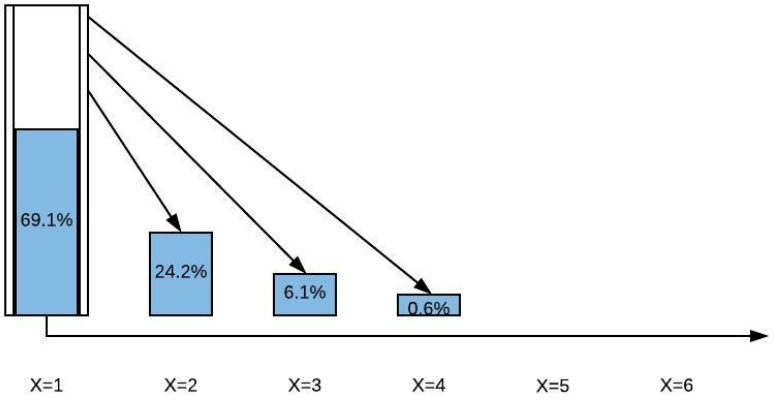

Figure 2.8: The effect of noise on an extreme response in an item with six categories

Table 2.1: Probability of noise for 6-category response transition matrix

Response category	1	2	3	4	5	6
1	0.691	0.242	0.061	0.006	0	0
2	0.309	0.382	0.242	0.061	0.006	0
3	0.067	0.242	0.382	0.242	0.061	0.006
4	0.006	0.061	0.242	0.382	0.242	0.067
5	0	0.006	0.061	0.242	0.382	0.309
6	0	0	0.006	0.061	0.242	0.691

$$\begin{bmatrix} p'_1 \\ p'_2 \\ p'_3 \\ p'_4 \\ p'_5 \\ p'_6 \end{bmatrix} = \begin{bmatrix} 0.691 & 0.242 & 0.061 & 0.006 & 0 & 0 \\ 0.309 & 0.382 & 0.242 & 0.061 & 0.006 & 0 \\ 0.067 & 0.242 & 0.382 & 0.242 & 0.061 & 0.006 \\ 0.006 & 0.061 & 0.242 & 0.382 & 0.242 & 0.067 \\ 0 & 0.006 & 0.061 & 0.242 & 0.382 & 0.309 \\ 0 & 0 & 0.006 & 0.061 & 0.242 & 0.691 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{bmatrix}$$

The matrix is not symmetric because the categories are truncated at the boundaries. For example, the probability for category 1 to 3 (= 0.061) is different from category 3 to 1 (=0.067), because category 1 has small chance to change to 4 (0.006) but it is not possible

for category 3 change to -1 (the fuzzy observations with noise are restricted on the design space).

For a particular response P_{ij} adding noise ε_{ij},

$$P_{ij}^{*} = P_{ij} + \varepsilon_{ij} \tag{2.10}$$

The noise ε_{ij} is with respect to the ith subject and jth question.

$$P_{ij}^{*} = \begin{cases} P_{ij} - 3 & \text{if } \varepsilon \in [-3.5, -2.5), \text{ with probability } 0.006 \\ P_{ij} - 2 & \text{if } \varepsilon \in [-2.5, -1.5), \text{ with probability } 0.061 \\ P_{ij} - 1 & \text{if } \varepsilon \in [-1.5, -0.5), \text{ with probability } 0.242 \\ P_{ij} & \text{if } \varepsilon \in [-0.5, 0.5], \text{ with probability } 0.382 \\ P_{ij} + 1 & \text{if } \varepsilon \in (0.5, 1.5], \text{ with probability } 0.242 \\ P_{ij} + 2 & \text{if } \varepsilon \in (1.5, 2.5], \text{ with probability } 0.061 \\ P_{ij} + 3 & \text{if } \varepsilon \in (2.5, 3.5], \text{ with probability } 0.006 \end{cases} \tag{2.11}$$

The expectation of changing categories from the original response after adding noise is to not change.

For an original response in category 1, the expected category change is,

$$0 \times 0.691 + 1 \times 0.242 + 2 \times 0.061 + 3 \times 0.006 = 0.382$$

For original response in category 2, the expected category change is,

$$-1 \times 0.309 + 0 \times 0.382 + 1 \times 0.242 + 2 \times 0.061 + 3 \times 0.006 = 0.073$$

For an original response in category 3, the expected category change is,

$$-2 \times 0.067 - 1 \times 0.242 + 0 \times 0.382 + 1 \times 0.242 + 2 \times 0.061 + 3 \times 0.006 = 0.006$$

For an original response in category 4, the expected category change is,

$$-3 \times 0.006 - 2 \times 0.061 - 1 \times 0.242 + 0 \times 0.382 + 1 \times 0.242 + 2 \times 0.067 = -0.006$$

For an original response in category 5, the expected category change is,

$$-3 \times 0.006 - 2 \times 0.061 - 1 \times 0.242 + 0 \times 0.382 + 1 \times 0.309 = -0.073$$

For an original response in category 6, the expected category change is,

$$-3 \times 0.006 - 2 \times 0.061 - 1 \times 0.242 + 0 \times 0.691 = -0.382$$

The overall expected change of categories is zero. □

Assume the original sample includes n subjects, in each bootstrap sample with noise, additional sample n with noise will be added to the original sample. In this setting, the sample size n is usually small.

2.2.3 Fuzzy Set Theory in Statistics

The fuzzy set theory (FST) is very popular recently due to the knowledge-based machine learning and artificial computational intelligence [12, 15]. FST was developed in the direction of fuzzy mathematics [15], though it was originally an extension of logic and classical set theory. It models reality better than traditional models in pattern classification and information processing. Data mining and statistical modeling has shifted from analyzing homogeneous data sets of closed populations to the analysis of more complex and diverse data sources from dynamic populations [12]. The uncertainty of the estimation is not only coming from the variance of the population but also from unexpected heterogeneous information. The idea of fuzzy is that the imprecision of the meaning of English word is more from the vagueness of its meaning rather than lack of knowledge of the parameters.

Fuzzy rules are useful for modeling human thinking, perceptions, and judgment. The idea of fuzzy set was first introduced by Lotfi A Zadeh in 1965 [37] to accommodate non-statistical uncertainties, which are sometimes referred to as Vague in Linguistics [38]. It provides a tool that can capture the uncertainties associated with human cognitive functions, such as thinking and reasoning. In FST, the exact reasoning or association of causes with effects derived from the sample is a limited approximation of reasoning, that is truth-values

(meaning or concept of a word) are fuzzy subsets of the unit interval [24].

It is worthwhile to develop the data-driven adaption of fuzzy systems, which is an integration of statistics, machine learning, data management, and computer science. The fuzzy set is a simple definition from the idea of extending one dimension source of information to two dimensions. For example, the easy fuzzifier process is to change the line to a triangle by introducing an additional point not on the line.

FST has been used in data selection and preparation [12], such as condensing several crisp observations into a single fuzzy observation, creating fuzzy summary of data, and modeling vague data, all of which are part of the fuzzy data analysis. There are two different ways of analyzing fuzzy data [12]: (1) extending traditional statistical methods directly to fuzzy data sets and (2) embedding data into the fuzzy metric spaces (a more sophisticated and complex approach).

2.2.4 Preliminary about fuzzy set theory

The Fuzzy Set is defined as below,

Definition 2.2. The set $\bar{A} = \{(x, f_{\bar{A}}(x)), x \in X\}$ is **fuzzy set**, where the function $f_{\bar{A}}$ is the membership function mapping $x \to [0, 1]$

The membership function fully described the fuzzy set and described the degree of similarity between the elements in the fuzzy set and the crisp set. The membership function can be selected based on experiences of the data set or based on a machine learning method, e.g. fuzzy artificial neural networks [38]. The typical shape of membership functions are triangle, trapezoid, or Gaussian. In this dissertation, we created the membership function as a step function shown in Figure 2.7.

Definition 2.3. The **crossover point** in A is defined as the element x with $f_{\bar{A}}(x) = 0.5$.

Sometimes, the crossover point is not considered a member of A.

Definition 2.4. Two fuzzy sets, A and B are equal, if and only if $f_{\bar{A}}(x) = f_{\bar{B}}(x)$ for all $x \in X$.

With similar definition, the relationship between the two fuzzy set $A \subset B$ if and only if $f_{\bar{A}}(x) \leq f_{\bar{B}}(x)$ for all $x \in X$. As part of the axiomatic system of fuzzy set theory, Zadeh (1968) generalized the membership function to the probability of an event A as [39, 3],

$$P(A) = E(f_{\bar{A}}) \tag{2.12}$$

Using the generalization from membership function to probability, the distribution of the event A can be related to the grade of membership of the element in fuzzy set A.

Definition 2.5. The fuzzy set A is a **convex normalized fuzzy set** of the real line $\mathbb{R}$ if there exists exactly one $x_0 \in \mathbb{R}$ such that $f_{\bar{A}}(x_0) = 1$ and the membership function $f_{\bar{A}}(x)$ is piece-wise continuous. In this case, the value x_0 is called the mean value of A.

Definition 2.6. The fuzzy set A is a **convex** if

$$f_{\bar{A}}(\lambda x_1 + (1-\lambda)x_2) \geq \min\{f_{\bar{A}}(x_1), f_{\bar{A}}(x_2)\} \tag{2.13}$$

The classical operations for set, such as union, intersection, and complement of set operations are easy to define in FST [40]. The union is defined as the maximum of the member functions. The intersection is defined as the minimum of the member functions. The complement is one minus the membership functions.

The union of fuzzy sets A and B is defined as,

$$\bar{A} \cup \bar{B} = \{x | f_{\bar{A}\cup\bar{B}}(x) = \max\{f_{\bar{A}}(x), f_{\bar{B}}(x)\}\} \tag{2.14}$$

The intersection of fuzzy sets A and B is defined as,

$$\bar{A} \cap \bar{B} = \{x | f_{\bar{A}\cap\bar{B}}(x) = \min\{f_{\bar{A}}(x), f_{\bar{B}}(x)\}\} \tag{2.15}$$

The complement of fuzzy sets A is defined as,

$$\bar{A}^c = \{x | f_{\bar{A}^c}(x) = 1 - f_{\bar{A}}(x)\} \tag{2.16}$$

The support function of fuzzy set is defined as,

$$Support(\bar{A}) = \{x | f_{\bar{A}}(x) > 0\} \tag{2.17}$$

Fuzzy set with support as a single point x is called fuzzy singleton,

$$f_{\bar{A}}(x) = 1$$

The core of fuzzy set is defined as,

$$Core(\bar{A}) = \{x | f_{\bar{A}}(x) = 1\} \tag{2.18}$$

The power set (or powerset) is defined as the set that included all subsets, including empty set and itself. If A is a finite set with $|A| = n$ elements, then the number of subsets of A is $|P(A)| = 2^n$.

The algebraic product is directly calculated at the function level as,

$$\bar{A} \cdot \bar{B} = \{x | f_{\bar{A} \cdot \bar{B}}(x) = f_{\bar{A}}(x) \cdot f_{\bar{B}}(x)\} \tag{2.19}$$

The algebraic sum of fuzzy set is calculated as,

$$\bar{A} + \bar{B} = \{x | f_{\bar{A}+\bar{B}}(x) = f_{\bar{A}}(x) + f_{\bar{B}}(x) - f_{\bar{A}}(x) \cdot f_{\bar{B}}(x)\} \tag{2.20}$$

For algebraic summation, is can be written as,

$$\bar{A} + \bar{B} = \{x | f_{\bar{A}+\bar{B}}(x) = 1 - (1 - f_{\bar{A}}(x) \cdot (1 - f_{\bar{B}}(x)\} \tag{2.21}$$

There are some bounded operations [40] specifically designed for set operations, because the membership function is defined on the closed interval $[0, 1]$. To ensure the function is still a membership function, the bounded-sum is defined as,

$$\bar{A} \oplus \bar{B} = \{x | f_{\bar{A} \oplus \bar{B}}(x) = \min\{1, f_{\bar{A}}(x) + f_{\bar{B}}(x)\} \tag{2.22}$$

The bounded-difference is defined as,

$$\bar{A} \ominus \bar{B} = \{x | f_{\bar{A} \ominus \bar{B}}(x) = \max\{0, f_{\bar{A}}(x) - f_{\bar{B}}(x)\} \tag{2.23}$$

The bounded-product is defined as,

$$\bar{A} \otimes \bar{B} = \{x | f_{\bar{A}\otimes\bar{B}}(x) = \max\{0, f_{\bar{A}}(x) + f_{\bar{B}}(x) - 1\} \tag{2.24}$$

The classical laws in set theory can all be easily carried over to FST for (2.14 to 2.20), such as Idempotent laws, commutative laws, associative laws, absorption laws, distributive laws, involution laws, De Morgan's laws, and identity laws [40]. In traditional fuzzy set theory, fuzzy sets does not form a Boolean Algebra because the complement laws are not satisfied due to,

$$\bar{A} \cup \bar{A}^c \neq \Omega \text{ and } \bar{A} \cap \bar{A}^c \neq \emptyset \tag{2.25}$$

2.2.5 Bootstrap re-sampling with noise

First we need to prove the fundamental theorem, Typical Value Theorem (TVT) still holds when noise is added to the original samples. The TVT is stated as below.

Theorem 2.1. *(The TVT of Bootstrap samples). For independent identically distributed (i.i.d.) observations, $X_1, X_2, \cdots, X_n$, the observations are indexed by integers $i = (1, 2, \cdots, n)$. Assume S_b, $b = (1, 2, \cdots, B)$ are the bootstrap subsamples randomly selected without replacement from $X_1, X_2, \cdots, X_n$, indexed by $2^n - 1$ non-empty subset of integers $i = (1, 2, \cdots, n)$. Let $\hat{\theta}_b$ be the estimates based on set S_b. These estimates divide the real line into $B+1$ partitions, denoted by $I_1, I_2, \cdots, I_{B+1}$, where $I_1 = (\infty, \hat{\theta}_1), \cdots, I_b = [\hat{\theta}_{b-1}, \hat{\theta}_b), \cdots, I_B = [\hat{\theta}_B, \infty)$. Then the probability for true $\theta \in I_b$ is $1/(B+1)$.*

The TVT defined a procedure that can be used to generate random subsampling (independent). In this way, the estimated values $\hat{\theta}_1, \cdots, \hat{\theta}_B$ are independent. This is because all samples are selected without replacement. We can use the percentile method to generate confidence intervals, which are exactly $j/(B+1)\%$. In this case, the $100(j/(B+1))$ confidence interval can be constructed by selecting the corresponding range in the middle [30]. But for small samples and asymmetric distributions, this method does not work well. Efron [41] has proposed four methods to correct for bias in confidence intervals.

The original sample is $X = (X_1, X_2, \cdots, X_n)$. Each X_i represent a person's response to survey space, i.e. a vector of $p \times 1$. After adding noise, Each X_i is a probability band. Each X_i is not fixed as traditional bootstrap sampling, it is a probability sample on the band instead.

(**Boundary**) Boundary is defined as the extreme responses, where most properties do not hold mathematically. In the noise added to extreme responses, there is 69.1% chance that the response will stay the same after noise. In this way, roughly 30% of the responses in the more extreme categories can be used to represent plausible population values for nearby categories.

The response with noise generates a probability response band centered on current response with probabilities and bounded by the response space, as shown on Figure 2.9. Further away from $\vec{X}_i$ has less probability of being endorsed in the sample with noise.

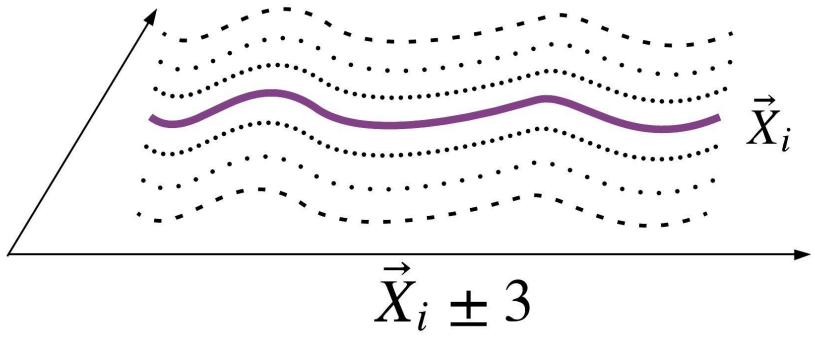

Figure 2.9: The probability response band centered at $\vec{X}_i$

The probability distribution for the probability band is explained on Figure 2.6. The probability band is generated by adding Gaussian noise. Instead of using only $\vec{X}_i$ to predict θ, we use sampled $\vec{X}^*_{i\epsilon}$ from the probability band centered at $\vec{X}_i$ bounded by R to predict θ.

Theorem 2.2. *(Pointwise Bounded Theorem) For any continuous and differentiable function $g : X \to R$,*

$$\|g(\vec{X}^*_{i\epsilon}) - g(\vec{X}_i)\| \leq g'(\epsilon)\|\vec{X}^*_{i\epsilon} - \vec{X}_i\| = \epsilon g'(\epsilon) \tag{2.26}$$

and

$$E(g(\vec{X}^*_{i\epsilon}) - g(\vec{X}_i)) = 0 \tag{2.27}$$

This theorem is easy to prove by *Middle Value Theorem* or *Rolle's theorem.* □

Now we can generate the very similar noisy version of TVT. We defined the original observations as, $X_1, X_2, \cdots, X_n$ and bootstrap sample S_b, $b = (1, 2, \cdots, B)$ the same as Theorem 2.1. The noise is introduced to bootstrap samples by $X^*_i = [X_i + \epsilon] \in R$, bounded within R, with $\epsilon \sim N(0, 1)$. The bootstrap sample with noise is denoted as S^*_b. Let $\hat{\theta}^*_b$ be the estimates based on set $X^*_{i_b} \in S^*_b$. Then we have the following theorem.

Theorem 2.3. *(**Noisy Typical Value Theorem**) The estimates $\hat{\theta}^*_b$ divide the real line into $B+1$ partitions, denoted by $I^*_1, I^*_2, \cdots, I^*_{B+1}$, where $I^*_1 = (\infty, \hat{\theta}^*_1), \cdots, I^*_b = [\hat{\theta}^*_{b-1}, \hat{\theta}^*_b), \cdots, I^*_B = [\hat{\theta}^*_B, \infty)$. Then the probability for true $\theta^* \in I^*_b$ is $1/(B+1)$.* □

Theorem 2.4. *For fixed continuous and differentiable function g, we can show $E(\hat{\theta}^*_b - \hat{\theta}_b) = 0$. If further g' is uniformly bounded, i.e. $g' \leq M$, we can show,*

$$\hat{\theta}^*_b \approx \hat{\theta}_b + \epsilon M \tag{2.28}$$

Here ϵ is normally distributed with mean 0 and variance 1. M is a constant.

By Taylor Series, $\hat{\theta}^*_b$ is estimated based on the bootstrap sample S^*_b. We use $X^*_{i_b}$ to denote the observations contained in the subsample S^*_b.The N^{th} order of Taylor Series for function g at $X^*_{i_b} \in S_b$ is,

$$\hat{\theta}^*_b = g(X^*_{i_b}) \tag{2.29}$$

and

$$g(X^*_{i_b}) = g(X_{i_b}) + g'(X_{i_b})(X^*_{i_b} - X_{i_b}) + \frac{g''(X_{i_b})}{2!}(X^*_{i_b} - X_{i_b})^2 + \cdots \tag{2.30}$$

Therefore, we can show

$$E(\hat{\theta}^*_b - \hat{\theta}_b) = E(g(X^*_{i_b}) - g(X_{i_b})) \approx E(g'(X_{i_b})) \times E(\epsilon) = 0 \tag{2.31}$$

This theorem indicates that the expectation of noisy bootstrap sample estimation is similar to the original sample for the same smooth function.

The standard error for bootstrap sample is, here $\bar{\hat{\theta}}$ is the average of all the bootstrap sample estimates.

$$Var(\hat{\theta}_b) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}})^2 \tag{2.32}$$

For noisy version, we directly calculated the variance from the bootstrap sample,

$$Var(\hat{\theta}_b^*) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2 \tag{2.33}$$

We can use Theorem 2.4 and the fact $Var(x) = Ex^2 - (Ex)^2$ to show the difference between the two variances is,

$$\begin{aligned}
Var(\hat{\theta}_b^*) - Var(\hat{\theta}_b) &= E((\hat{\theta}_b^*)^2) - E(\hat{\theta}_b^*)^2 - E(\hat{\theta}_b^2) + E(\hat{\theta}_b)^2 \\
&= E((\hat{\theta}_b^*)^2) - E(\hat{\theta}_b^2) + (E\hat{\theta}_b - E\hat{\theta}_b^*)(E\hat{\theta}_b + E\hat{\theta}_b^*) \\
&= E((\hat{\theta}_b^*)^2) - E(\hat{\theta}_b^2) \\
&\approx E((\hat{\theta}_b + \epsilon M)^2) - E(\hat{\theta}_b^2) \\
&= E(\hat{\theta}_b^2 + 2\epsilon M\hat{\theta}_b + \epsilon^2 M^2) - E(\hat{\theta}_b^2) \\
&= 2ME(\epsilon)E(\hat{\theta}_b) + M^2E(\epsilon^2) \\
&= M^2
\end{aligned} \tag{2.34}$$

As expected, after introducing standard normally distributed noise, the variance of the bootstrap sample increases but the level of increment is bounded as a constant.

2.2.6 Fuzzification process

Fuzzification is the process of converting each data point X_i^* from the bootstrapped samples (sampled from X_i) to fuzzy data point Z_i^*. The process could be either a lookup table or a transform function with probabilities. The membership function of a fuzzy set is defined in this step. After this process, each re-sampled response X_i^* is fuzzified with a probability that the fuzzy response Z_i^* is close to the corresponding data point X_i in original sample space, which is referred to crisp set C_r. The fuzzified observations are denoted as $Z_i^* \in A_{X_i}$. The fuzzy set A_{X_i} is generated by the original sample X_i after bootstrapping and introducing noise.

The fuzzy set A_{X_i} is generated by $X_i \in C_r$. The fuzzification process is only applied to the short form items of survey response vector X_i. We do not add noise to demographic questions under the assumption that the population should have similar characteristics as the sample. The fuzzy observations include fuzzy short form responses, similar demographics, and associated oral health outcomes. Therefore, for each input pair (X_i, Y_i) in the bootstrap samples, the fuzzification layer generates observations with the probability $(\vec{X}_i, \vec{P}_\epsilon)$. The probability will be later used as a weight in estimating the fitted outcomes. For each observation, an identically independently distributed (i.i.d.) noise from the standard normal distribution will be added as,

$$\vec{Z}_{X_i} = \vec{X}_i + \vec{\varepsilon} \tag{2.35}$$

The probability of Z_{X_i} is the probability of the maximum noise element of $\vec{\epsilon}$,

$$P_{Z_{X_i}} = \|\vec{\varepsilon}\|_\infty \tag{2.36}$$

The fuzzy set A_{X_i} is defined as all observations in the bootstrap sample that is generated by X_i with its probability as membership function.

Theorem 2.5. *The fuzzy set A_{X_i} is a convex normalized fuzzy set with mean value X_i and variance greater than $Var(X_i)$.*

Proof:

The convex normalized fuzzy set is defined in 2.5. For fuzzy set A_{X_i}, we let the probability of $X_i \in A_{X_i} = 1$. The membership function defined as above is a piece-wise continuous as shown in Figure 2.7.

All the noise ε_{bj}, for $j = 1, \cdots, n_b$ and $b = 1, \cdots, B$ are i.i.d. from standard normal distribution. Given all X_i^* are equal, we have,

$$\begin{aligned} E(Z_i) &= E(X_i + \varepsilon) \\ &= E(X_i) + E(\varepsilon) \\ &= X_i \end{aligned} \tag{2.37}$$

For categorized noise version, we have

$$\begin{aligned} E(Z_i) &= \sum_{Z_i=X_i-3}^{Z_i=X_i+3} Z_i P_{Z_i} \\ &= (X_i-3)P_{Z_i=X_i-3} + (X_i-2)P_{Z_i=X_i-2} + \cdots + (X_i+3)P_{Z_i=X_i+3} \\ &= X_i \times \sum_{Z_i=X_i-3}^{Z_i=X_i+3} P_{Z_i} \\ &= X_i \end{aligned} \tag{2.38}$$

and

$$\begin{aligned} Var(Z_i) &= Var(X_i+\varepsilon) \\ &= Var(X_i) + Var(\varepsilon) \\ &= Var(X_i) + 1 \end{aligned} \tag{2.39}$$

For categorical noise version, we have

$$\begin{aligned} Var(Z_i) &= E(Z_i - E(Z_i))^2 \\ &= \sum_{Z_i=X_i-3}^{Z_i=X_i+3} P_{Z_i}(Z_i - E(X_i))^2 \\ &= \sum_{Z_i=X_i-3}^{Z_i=X_i+3} P_{Z_i}(X_i - E(X_i))^2(9P_{X_i-3} + 4P_{X_i-2} + P_{X_i-1}) \times 2 \\ &= Var(X_i) + 1.08 \\ &\approx Var(X_i) + 1 \end{aligned} \tag{2.40}$$

In the variance part, there is rounding error when the probabilities are squared. In this way, we show that adding noise at response options does not change the means and variances when rounding the noisy responses at design options. □

The aim of the fuzzification layer is to combine the randomness and fuzziness and therefore generate fuzzy random variables for the next layer to make inferences. The simple definition is from [42, 43] as the random variables with values as fuzzy numbers. It is a natural combination of randomness and fuzziness as shown in Figure 2.10 using previous simple temperature example. The fuzzification is one way to re-construct the uncertainty

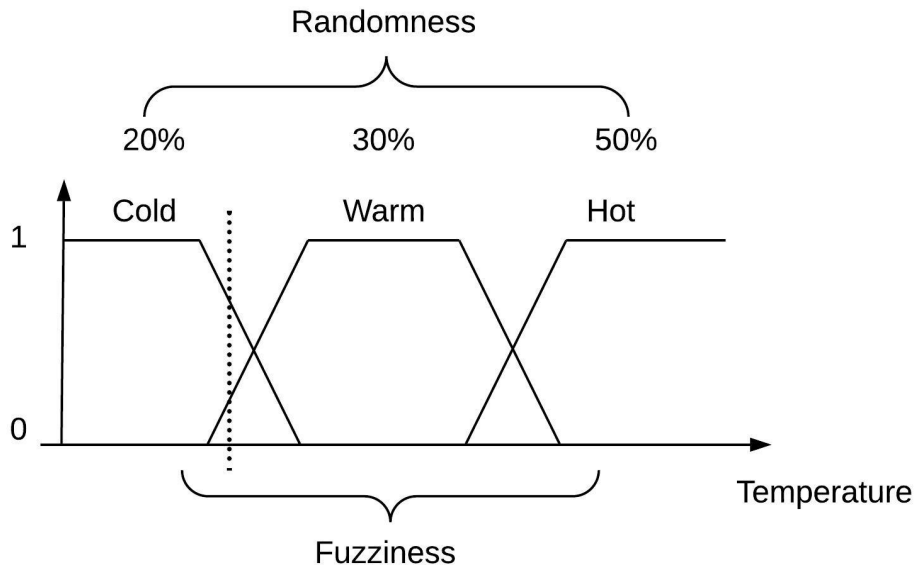

Figure 2.10: Combine fuzziness and randomness

in statistical literature. In this dissertation, the uncertainty of fuzzy system is from two parts. One is from the randomness when using bootstrap method to resample from original observations. Another part is using fuzzy set to blur the boundary among survey options by adding a simple noise structure to observations. In this way, the sample is expanded to plausible observations with roughly Gaussian distribution with mean at itself except at boundaries.

2.3 Inference Layer

The inference layer develops the algorithms that are used for prediction. The algorithm is developed based on the sample size (N) in the previous section. In the following, we discuss several algorithms, that are different based on the outcomes.

2.3.1 Categorical variables - Classification

For categorical outcome, the simplest form of classification is using the Naive Bayesian (NB) classifier [44]. NB has been used intensively for classification problems because of its

simplicity but surprisingly high accuracy [45, 46]. In classification, it has been recently used for diagnosis, disease prediction, and classification [47, 48, 49]. NB [50, 51] is based on Bayes Theorem and has a smaller error rate than other tree-based algorithms for classification [49]. It directly converts the conditional probabilities between clinical exam results and survey responses. We use $\mathbb{R}$ to denote the response space of survey questions. The outcome from clinical exam is denoted by E, with E_{cat} denotes the categorical outcomes referral for treatment needs (RFTN), with k denoting categorical level. I use E_{con} to represent the continuous outcome, children's oral health status index (COHSI).

Using Bayes' theorem, the conditional probability can be expressed in general,

$$P(E|\mathbb{R}) = \frac{P(E) \times P(\mathbb{R}|E)}{P(\mathbb{R})} \tag{2.41}$$

In this notation, $\mathbb{R} = (P_1, P_2, \cdots, P_J, C_1, C_2, \cdots, C_H)$. In plain language, the equation can be explained,

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Evidence}} \tag{2.42}$$

The classification is determined by the posterior probability that is calculated by the prior information, the sample statistics (Likelihood in 2.42), and the sample response space (Evidence in 2.42). In the context of the study, the posterior information is the prediction of the clinical outcomes, given the survey responses. The prior information is estimated by the probability of treatment needs in the field test sample. The sample statistics are estimated by the conditional probability of the survey responses, given RFTN. The translated version is,

$$\text{Prediction from survey} = \frac{\text{Incidence} \times \text{Field Test}}{\text{Survey Response}} \tag{2.43}$$

The predicted RFTN falls onto the category with the highest posterior probability. This is called maximum a posteriori (MAP). NB is widely used in the prediction of categorical outcomes. The only assumption in NB is conditional independence among predictors (short form and demographics), which is the same assumption used in item response theory (IRT). The conditional independence is used in Graded Response models among the short forms items.

The estimation procedures are separated for if only children's information is available, if only the response from parents' are available, if both information are available, and if additional information are available, e.g. demographics. Below are the derivation for categorical outcome E_k, i.e.$E_{\text{cat}} = k$. The posterior probability ($P(E|\mathbb{R})$) can be estimated by maximum probable hypothesis (MAP). The chain rule is used to calculate the joint probability of the responses from children ($\mathbb{R}_C$) and outcomes E,

$$P(\mathbb{R}_C, E_k) = P(C_1|C_2, \cdots, C_H, E_k)P(C_2|C_3, \cdots, C_H, E_k)\cdots P(C_H|E_k)P(E_k)$$

If we assume "naive independence", the joint probability is,

$$P(\mathbb{R}_C, E_k) = P(C_1|E_k)P(C_2|E_k)\cdots P(C_H|E_k)P(E_k)$$

The assumption is very strong in practice, but even if the assumption does not hold, the results are not affected much. Therefore, the posterior probability of Naive Bayesian is,

$$\begin{aligned} P(E_k|\mathbb{R}_C) &= \frac{P(\mathbb{R}_C, E_k)}{P(\mathbb{R}_C)} \\ &= \frac{P(E_k) \times \prod_{h=1}^{H} P(C_h|E_k)}{P(\mathbb{R}_C)} \\ &= \frac{P(E_k) \times \prod_{h=1}^{H} P(C_h|E_k)}{\sum_k P(E_k) \times \prod_{h=1}^{H} P(C_h|E_k)} \end{aligned} \tag{2.44}$$

The classifier for the outcome of subject i, given the self-reported survey responses is y_i is $(i = 1, \cdots, N)$,

$$y_i = k, \text{argmax}_k \left(P(E_k|\mathbb{R}_C)\right) \tag{2.45}$$

Using the fact the response space $P(\mathbb{R}_C)$ does not change from population to population, only numerator counts in the final prediction. Further expanded the above formula, with consideration of being very conservative about the prediction (always assign smaller category),

$$y_i = \min_k \left\{ \text{argmax}_k \left(P(E_k) \times \prod_{h=1}^{H} P(C_h|E_k) \right) \right\} \tag{2.46}$$

In this model, the probability is estimated by two parts $P(E_k)$ and the products of all conditional probabilities $P(C_h|E_k)$. The probability $P(E_K)$ can be modified if the information

is available in new population. The other part of the formula is the probability estimated based on the field test data. Given the oral health status, the probability of response falls into different categories across all questions answered by children.

In certain situations, only the answers from parents are available and appropriate to use, for example, when the children are too young for the survey questions. The parent can provide proxy responses for the surveys, $\mathbb{R}_P$. Except for the level of accuracy for some questions related to the child's personal experiences (e.g. pain, quality of life, satisfaction), the response from parents has a broad range. Though the accuracy of the estimation still depends on the dental literacy of parents, the responses are treated an additional information for children's response. Similar as children's estimation, the posterior probability for NB of parents is,

$$P(E_k|\mathbb{R}_P) = \frac{P(E_k) \times \prod_{j=1}^{J} P(P_j|E_k)}{\sum_k P(E_k) \times P(\mathbb{R}_P|E_k)} \tag{2.47}$$

Then the classifier for the outcome of subject i, given only the survey response of his or her parent,

$$y_i = k, \text{argmax}_k \left(P(E_k|\mathbb{R}_P)\right) \tag{2.48}$$

Therefore, the final estimation for parents response only can be estimated as,

$$y_i = \min_k \left\{ \text{argmax}_k \left(P(E_k) \times \prod_{j=1}^{J} P(P_j|E_k) \right) \right\} \tag{2.49}$$

The explanation for the parents' model is the same as the children's model. The prediction is based on two parts of the probabilities, the incidence rate of the current population and the field test sample estimation.

2.3.2 Continuous variables - Prediction

For the continuous outcomes, we elected to use XGBoost [52, 53], eXtreme Gradient Boosting, among many available algorithms. XGBoost is a scalable tree boosting algorithm widely used recently on many machine learning challenges and can achieve better prediction results while using less resources for complicated computations [54, 55]. It is the most popular machine learning algorithm since it was introduced in 2014 and is widely used in the re-

cent decade on many machine challenges. It is highly effective for sparse data due to its scalability. In this section, we briefly review the techniques used in XGBoost. It is also a supervised learning algorithm built on a gradient boosting method. We predicted COHSI scores obtained from the dental exam by several additive functions [31, 34] developed from the short form responses and demographic information. The parameters of the model are derived from the loss function (better prediction) and regulation function (less overfitting). XGBoost is very efficient on sparse data and large data sets.

Assume we have a data set $\{(X_1, Y_1), (X_2, X_2), \cdots, (X_n, Y_n)\}$ with size n, where X_i is a vector of $\mathbb{R}^p$. Here we assume Y_i is a continuous variable in $\mathbb{R}$. Then it is estimated by K additive functions in $\mathcal{F}$ [52],

$$\hat{Y}_i = \sum_{k=1}^{K} f_k(\vec{X}_i), \forall f_k \in \mathcal{F} \tag{2.50}$$

Here f_k denotes an independent tree structure with leaf weights $\vec{\omega}$ and $\mathcal{F}$ is the space of regression trees. The weight is,

$$\vec{\omega} = (\omega_1, \omega_2, \cdots, \omega_T)$$

The number of leaves is T_k for each tree for $k = 1, 2, \cdots, K$. The total number of leaves is T. The risk function is defined as the loss function and the regulation function [53],

$$L = \sum_{i=1}^{n} l(\hat{Y}_i, Y_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.51}$$

In the risk function 2.51, the loss function is choosing among differentiable and convex functions. The regulation function is defined as,

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2.52}$$

The regulation function aims to penalize the complexity of the model to avoid over-fitting. The regulation parameters γ is for total number of leaves T and λ is for the weights of leaves when using L_2 norm.

In gradient tree boosting, use the fact that Y_i can be estimated by the additive function in $\mathcal{F}$, the t^{th} iteration of $\hat{Y}_i$ is,

$$\hat{Y}_i^{(t)} = (\hat{Y}_i^{(t-1)} + f_t(X_i)$$

The risk at $(t)^{th}$ iteration can be written as below [53],

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l(\hat{Y}_i^{(t-1)} + f_t(X_i), Y_i) + \Omega(f_t) \\ &= \sum_{i=1}^{n} l(\hat{Y}_i^{(t-1)}, Y_i) + g_i f_t(X_i) + \frac{h_i}{2} f_t^2(X_i) + \Omega(f_t) \end{aligned} \tag{2.53}$$

The above formula is proved by Taylor expansion to the second order expanding at $Y_i^{(t-1)}$. The first order partial derivatives in 2.53 is,

$$g_i = \frac{\partial l(\hat{Y}_i^{(t-1)}, Y_i)}{\partial Y_i^{(t-1)}}$$

The second order partial derivatives in 2.53 is,

$$h_i = \frac{\partial^2 l(\hat{Y}_i^{(t-1)}, Y_i)}{(\partial Y_i^{(t-1)})^2}$$

Now the goal is to find the best set of $\vec{\omega}_k$ that minimize 2.53. Revisiting the leaves on the tree, assume $q(x)$ is a function that map each observation $X_i \in \mathbb{R}^p$ to the leaves of tree (T). Denote the set of all leaves j as I_j,

$$I_j = \{i | q(X_i) = j\} \tag{2.54}$$

Rewrite the risk function in 2.53 as (remove constant items) below, by rearranging the summation items by leaves [53],

$$\begin{aligned} L^{(t)} &\propto \sum_{i=1}^{n} g_i f_t(X_i) + \frac{h_i}{2} f_t^2(X_i) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} (\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2 + \gamma T \end{aligned} \tag{2.55}$$

For any fixed structure (fixed T), we can find ω^* that minimizes the risk function 2.55 by setting the first derivative with respect to each ω_j to zero. The weight for each leaf j is,

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{2.56}$$

The risk function is equal to (the quality of the tree structure),

$$L^{(t)} \propto -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{2.57}$$

Once the structure of the tree with weights on each leaf is determined, the outcome can be directly predicted from the regression trees. The R function is fully available based on greedy algorithm [56], which starts from one leaf and adding branches in each step.

2.3.3 Training and Testing

When developing an algorithm using Machine Learning theory, we need to split the original data into a training set and a test set. The training set and test set are used in machine learning to derive the algorithms. The test data if separated from the original sample is called an internal test set. If a new source of information is collected, independent of the training set and test set, then the test set is called an external test set. The aim of the test set is to estimate the true error of the learning algorithm. The training set and test set are used in machine learning to derive the algorithms.

When the learning fails, the main approaches to fix the failure is noted in the literature [57]. The most common situation is to get enough sample size for training the algorithm. Alternatively, one can update the hypothesis class about the available algorithms or models, in order that a model of good fit can be learned. Another remedy could be updating the features that represent the data. If the selected X_i's cannot be linked to Y_i's, the least used remedy is to change the criteria of optimization. For example changing the rule to accept error rate.

Both NB and XGBoost are supervised learning and require a large sample size for the training set to achieve accurate prediction, stable parameters, and generalizable algorithms. The supervised learning algorithm means that the algorithm is taught by the training data set with the existing mapping between the outcome and predictors. The test set is used to correct the learning from training. The training process stops if the prediction satisfies the criteria. Then the training algorithm is used on the test data set.

Given a fixed sample size, there are several recommendations for splitting data into training set and test set, for example 60% for training and 40% for testing, 70% and 30%, or 80% and 20% . No matter how the sample is split, the sample size gets smaller overall. If

one spends too much of a share of available resources on training (e.g. 80% in training), we may not be able to fully evaluate the performance of the model. Hence, overfitting problems may exist. Meanwhile, if one spends a lot of resources on evaluation (e.g. 40% in testing), the parameters in the model may not be stable and well-trained. We use 70% and 30% as our training set and internal test set. The way used to split the data is completely random.

Therefore, the original data were divided into 70% for training the algorithm, and 30% for testing the generalizability and stability of the algorithm as suggested above and in the literature [58, 59]. In this case, we can reduce the overfitting problem during training the algorithms or models and develop a more stable and generalizable algorithm. The commonly used statistics for validating the prediction are sensitivity and specificity for categorical outcome. For continuous outcomes, we commonly use Pearson correlation and root-mean-square error (RMSE) to evaluate the predictive of the algorithms. The 30% test data is used in each step to estimate whether the algorithm satisfied the pre-determined criteria (e.g. 90% sensitivity and higher specificity). The error estimated from this part of data is internal test. To fully estimate the error of the learning, we also recruit data from one more dental clinic using the same method to collect additional data to evaluate the performance of the algorithm. We expected the algorithm from the fuzzy set system to perform better than the original observations, due to it being more "knowledgeable" on the response space $\mathbb{R}$. The new data were used only for external test and never used for training the algorithm.

2.4 Defuzzification Layer

In the literature, there are many different methods of defuzzification [60]. In this fuzzy system, we use two steps, at fuzzy set level and at prediction level. At the fuzzy set level, the observations are converted to the response on a grid (i.e. the categorical response on the survey space). This step ensures all fuzzy observations still belong to response space $\mathbb{R}$. It is a trivial step of defuzzification. The other step is using the predicted algorithm at the estimation stage after making inferences. At this step, the defuzzification process uses the weighted average of the estimates from the derived algorithm to accommodate the fuzzy

observations and membership functions so that any new input is predicted by all available learning resources. In previous sections, we list several methods that can be used in the inference layer to either make predictions or develop rules for classification. We will assume in this section the algorithm $f \in \mathscr{F}$ is learned from the above inference layer.

2.4.1 Type of new input

We know in almost of all the surveys, $\mathscr{X} \subseteq \mathbb{R}$, i.e. the survey response can not cover the entire survey space completely. For example, a simple survey of 10 questions with only yes and no options in response may have $2^{10} = 1024$ possible combinations in response space. We must collect at least 1024 samples with unique responses to cover the complete response space in design. This is very difficult in practice.

This problems exists in most samples and survey designs. There is not enough sample size to cover most of the scenario in the survey response space $\mathscr{X}$. That is, the sample space is incomplete. There is always a new input that is not covered in the sample space. The set $C_r = \{X_1, X_2, \cdots, X_n\}$ includes all observations that can be reached by the original sample. The new input is X_{n+1}. The fuzzy set generated by X_i is denoted as A_{X_i}.

In the following, we define four type of new inputs as show in Figure 2.11,

- Observations from Crisp set, i.e. $X_{n+1} \in C_r$. This is sometimes referred as estimation;
- Observations from fuzzy set but not in Crisp set C_r that need interpolation, i.e. $X_{n+1} \in A_{X_q}$ for some $X_q \in C_r$ but $X_{n+1} \notin C_r$. There exist observations in Crisp set $X_l \in C_r$ and $X_m \in C_r$, such that $X_l < X_{n+1} < X_m$, i.e. X_{n+1} is bounded by two observations in C_r;
- Observations from fuzzy set but not in Crisp set C_r that need extrapolation, i.e. $X_{n+1} \in A_{X_q}$ but $X_{n+1} \notin C_r$, i.e. X_{n+1} cannot be bounded by two observations in C_r;
- New observations neither in fuzzy set or Crisp set, i.e. new data input, $\forall 1 \leq q \leq n, X_{n+1} \notin A_{X_q}$ and $X_{n+1} \notin C_r$;

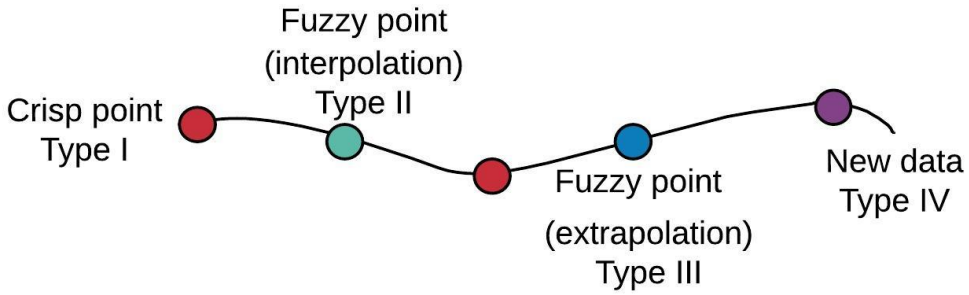

Figure 2.11: Four types of new input values

The fuzzy system is designed to predict new observations, with the "smallest" generalization error, defined as,

Definition 2.7. (Generalization error). The generalization error is in general defined as the difference between the prediction function $f \in \mathscr{F}$ and real function f_T,

$$P_{\forall x}(f(x) \neq f_T(x)) = E_x(1_{f(x)\neq f_T(x)}) \tag{2.58}$$

2.4.2 Defuzzification for Crisp set (Type I)

In Crisp set C_r, all observations are from sample data. For the new observation $X_{n+1} \in C_r$, the estimation is from all potentially related observations. All fuzzy sets potentially contained X_{n+1} contribute the estimation. Assume the new observation $X_{n+1} = X_q \in C_r$ It is the weighted average based on probabilities of crisp set and fuzzy set as,

$$Y_{n+1} = \frac{\sum_{i=1,i\neq q}^{n} P(X_{n+1} \in A_{X_i}) f(X_i) + \sum_{b=1}^{B} \sum_{j=1}^{n_b} P(Z_q^{(b)_j} \in A_{X_q}) f(Z_q^{(b)_j})}{\sum_{i=1,i\neq q}^{n} P(X_{n+1} \in A_{X_i}) + \sum_{b=1}^{B} \sum_{j=1}^{n_b} P(Z_q^{(b)_j} \in A_{X_q})} \tag{2.59}$$

By using all the available information to predict the data point in crisp set, the prediction is a weighted average of the prediction functions of fuzzy set data that are close (by probabilities) to the observation X_{n+1}.

2.4.3 Defuzzification for fuzzy set (Type II)

In fuzzy set, the predicted outcome (type II) is the weighted average based on probabilities of fuzzy set observations. Since the fact that interpolation input is bounded from below and up by fuzzy observations, the estimation is directly made from the weighted averages with all covered fuzzy observations,

$$Y_{n+1} = \frac{\sum_{i=1}^{n} P(X_{n+1} \in A_{X_i}) f(X_i)}{\sum_{i=1}^{n} P(X_{n+1} \in A_{X_i})} \tag{2.60}$$

2.4.4 Defuzzification for fuzzy set (Type III)

In fuzzy set, the observations (type III) is related to the weighted average based on probabilities of fuzzy set observations but is not fully bounded by the fuzzy observations. Part of the uncertainty is from outside of the fuzzy set. The estimation is from both the weighted average of nearby fuzzy observations and the algorithm directly applied onto the data point.

$$Y_{n+1} = \frac{[\sum_{i=1}^{n} P(X_{n+1} \in A_{X_i}) f(X_i)] + f(X_{n+1})}{\sum_{i=1}^{n} P(X_{n+1} \in A_{X_i}) + 1} \tag{2.61}$$

The prediction for this type both depends on the information from the fuzzy set, and the algorithm itself. The observations is usually outside the original sample. The uncertainty of the estimation is larger than Type I and Type II above.

2.4.5 Defuzzification for new input (Type IV)

If the observation is completely new (neither from the original sample or the fuzzy set), the only information is the function or algorithm developed from the fuzzy system, it is the direct estimation from,

$$Y_{n+1} = f(X_{n+1}) \tag{2.62}$$

In this way, the estimation is beyond the representativeness of the sample. The system can generate the estimation value, but the uncertainty is the largest among all types of predictions.

2.5 Production Layer

2.5.1 Sample Size and Sample Complexity

In Machine Learning Theory, sample complexity needs to be taken into account to determine the required sample size. This is well-defined in Probably Approximately Correct (PAC) Learning, which is the first to define learning in Machine Learning Theory. Before defining what is PAC learning, we first introduce the sample complexity (SC) [61]. The sample complexity can be defined as a function of the two approximation parameters in PAC, the accuracy parameter ϵ and the confidence parameter δ [57]. The accuracy parameter ϵ describes how close the estimate is to the optimization, i.e. approximately correct. The confidence parameter δ indicates how confident the learning algorithm meet the accuracy requirements. The function defined on the hypothesis space $\mathcal{H}$ maps the two parameter to an integer $\mathbb{N}$. It is the requirement of minimum sample size for PAC learnable. It has shown that [57] for finite hypothesis class, it is PAC learnable if the sample complexity,

$$N_{\mathcal{H}} \leq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \tag{2.63}$$

We will introduce ϵ-representative for sample complexity and sample size of finite space $\mathbb{R}$. First we define the representativeness of the sample for [57] as below.

Definition 2.8. (ϵ-representative). A training set T is called ϵ-representative, with respect to domain $\mathcal{Z}$, hypothesis function class $\mathcal{H}$, loss function l, and the distribution $\mathscr{D}$, if $\forall h \in \mathcal{H}$,

$$|L_T(h) - L_{\mathscr{D}}(h)| \leq \epsilon \tag{2.64}$$

Here $L(h)$ is the risk function, defined as the expected loss of an algorithm in $\mathcal{H}$,

$$L_{\mathscr{D}}(h) \triangleq E_{\mathscr{D}} l(h, (X, Y)) \tag{2.65}$$

The empirical risk or training error is defined as the expected loss over the sample T,

$$L_T(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} l(h, (X_i, Y_i)) \tag{2.66}$$

Sometimes, we can simply write the loss function as below when l is defined in the context of X_i,

$$l(h,(X_i,Y_i)) = l(h(X_i),Y_i) \tag{2.67}$$

For most of learning algorithm, it is searched by minimizing the risk function on training set. Find an algorithm h to minimize $L_T(h)$, which is called Empirical Risk Minimization (ERM). To avoid overfitting, the searching process is restricted on a finite class of functions $\mathcal{H}$. The algorithm $h_T \in \mathcal{H}$ is the results of applying the $\text{ERM}_{\mathcal{H}}$ to training set T [57],

$$h_T = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\ L_T(h) \tag{2.68}$$

A simple theorem has been proved [57] that if the sample is $\epsilon/2$-representative, then ERM process will return an algorithm with risk close to the distribution $\mathscr{D}$, even though $\mathscr{D}$ is unknown. Now we extend this theorem to the fuzzy set T_ε, T with noise.

Theorem 2.6. *If the sample set T is ϵ-representative, the fuzzy set with noise on T is also ϵ-representative.*

Proof: For all algorithm $h_\varepsilon \in \mathcal{H}$, we have

$$\begin{aligned} |L_{T_\varepsilon}(h) - L_{\mathscr{D}}(h)| &\le |L_{T_\varepsilon}(h) - L_T(h)| + |L_T(h) - L_{\mathscr{D}}(h)| \\ &\le |L_{T_\varepsilon}(h) - L_T(h)| + \epsilon \end{aligned} \tag{2.69}$$

For the motivating example considered here with a seven-level categorical variable, for each observation $(X_i = x, Y_i = y) \in T$, the corresponding fuzzy set of observations is,

$$(X_i = x, Y_i = y) \Leftrightarrow \begin{cases} (X_i = x-3, Y_i = y), & \text{with probability } 0.006 \\ (X_i = x-2, Y_i = y), & \text{with probability } 0.061 \\ (X_i = x-1, Y_i = y), & \text{with probability } 0.242 \\ (X_i = x, Y_i = y), & \text{with probability } 0.382 \\ (X_i = x+1, Y_i = y), & \text{with probability } 0.242 \\ (X_i = x+2, Y_i = y), & \text{with probability } 0.061 \\ (X_i = x+3, Y_i = y), & \text{with probability } 0.006 \end{cases} \tag{2.70}$$

Therefore, the risk of fuzzy set (T_ε without noise) is,

$$\begin{aligned} L_{T_\varepsilon}(h) &= E_{T_\varepsilon} l(h,(X,Y)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{-3}^{3} w(k) l(h,(x_i = x+k, Y_i))}{\sum_{-3}^{3} w(k)} \end{aligned} \tag{2.71}$$

The crisp set (T without noise) is estimated directly as,

$$\begin{aligned} L_T(h) &= E_T l(h,(X,Y)) \\ &= \frac{1}{n}\sum_{i=1}^{n} l(h,(x_i = x, Y_i)) \end{aligned} \tag{2.72}$$

As long as the loss function in (2.72) can be estimated by the weighted average of loss function in (2.71), the sample space with noise is also ϵ-representative. □

We assume the weighted average loss function is approximately estimate the the loss function of ϵ-representative the expected loss function on sample space with a small bias ξ, we will later show this bias is zero almost everywhere in the fuzzy set when the fuzzy set is large enough. A theorem can be immediately derived as below,

Theorem 2.7. *If the sample set T is ϵ-representative, w.r.t. domain $\mathcal{Z}$, hypothesis function class $\mathcal{H}$, loss function l, and the distribution $\mathscr{D}$, then any algorithm derived from the ERM process, i.e., has the following property,*

$$L_{\mathscr{D}}(h_{T_\varepsilon}) \leq \min_{h\in\mathcal{H}} L_{\mathscr{D}}(h) + \textit{small amount} \tag{2.73}$$

The ERM algorithm is defined as,

$$h_{T_\varepsilon} = \underset{h\in\mathcal{H}}{argmin}\ L_{T_\varepsilon}(h) \tag{2.74}$$

Proof: The risk function of h_{T_ε} from fuzzy set is,

$$\begin{aligned} L_{\mathscr{D}}(h_{T_\varepsilon}) &\leq L_T(h_{T_\varepsilon}) + \epsilon \\ &\approx L_{T_\varepsilon}(h_{T_\varepsilon}) + \epsilon + \xi \\ &\approx L_{T_\varepsilon}(h) + \epsilon + \xi \\ &\approx L_T(h) + \epsilon + 2\xi \\ &\approx L_{\mathscr{D}}(h) + 2\epsilon + 2\xi \end{aligned} \tag{2.75}$$

The small amount $2\epsilon + 2\xi$ in probability very small. □

The "No-Free-Lunch" theorem [57] indicates the estimation bias and sample complexity are trade off to ensure the learning algorithm is PAC learnable. If the sample space is arbitrarily complex, e.g. infinite domain set, and the hypothesis space $\mathcal{H}$ is the set of all functions, then it is not PAC learnable. There is an important class of hypothesis space $\mathcal{H}$ with uniform convergence property [62, 57].

Definition 2.9. (**Uniform convergence property**) The hypothesis function class $\mathcal{H}$ satisfied the uniform convergence property w.r.t. domain $\mathcal{Z}$ and loss function l if $\exists N_{\mathcal{H}} < \infty$ and $\forall \epsilon, \delta \in (0,1)$, for any distribution $\mathscr{D}$ over $\mathcal{Z}$, a sample $X_1, X_2, \cdots, X_n$ with $n = N_{\mathcal{H}}$, we have

$$P\left(\sup_{h\in\mathcal{H}} |\frac{1}{n}\sum_{i=1}^{n} l_h(X_i) - E_{\mathscr{D}(l_h)}| \leq \epsilon\right) \geq 1 - \delta \tag{2.76}$$

This definition guaranteed PAC learnable of hypothesis space $\mathcal{H}$ using empirical risk minimization (ERM). □

When the sample is not fully represent the response space $\mathbb{R}$, we will study the subset $\mathscr{X} \in \mathbb{R}$, which is the available sample due to design, missing response, or skewed distribution. We show by extending the sample space using fuzzy responses, the derived algorithm is improved. Before prove the major theorem, we introduced a few definitions in PAC learning theory.

Another important definition in PAC learning is VapnikChervonenkis (VC) dimension [57]. When restricted the hypothesis space $\mathcal{H}$ to $T = \{t_1, t_2, \cdots, t_m\} \subset \mathscr{X}$ on only a set of functions derived from $\mathcal{H}$ that map T to $\{0,1\}$. The $\mathcal{H}$ restricted on T is denoted by $\mathcal{H}_T$, which is a subset of $\{0,1\}^{|T|}$. We say $\mathcal{H}$ shatters T if $\mathcal{H}$ restricted on T is all the functions from T to $\{0,1\}$, i.e.,

$$|\mathcal{H}_T| = 2^{|T|}$$

Definition 2.10. (VC dimension) The VC dimension of hypothesis $\mathcal{H}$ is defined as the maximal size of T shattered by $\mathcal{H}$.

When we add noise to set $T \in \mathbb{R}$, because $T_\varepsilon \in \mathbb{R}$, then $|T_\varepsilon \cup T| \geq |T|$ and hence,

$$\mathcal{H}_{T_\varepsilon \cup T} \subseteq \mathcal{H}_T \tag{2.77}$$

The possible new observations generated based on the probability provided in 2.70 in fuzzy set from $\{t_i\} \to \{0, 1\}$ is,

$$\{t_i, t_i \pm 1, t_i \pm 2, t_i \pm 3\} \to \{0, 1\}$$

Therefore, $\mathcal{H}$ restricted on $T_\varepsilon \cup T$ is

$$|\mathcal{H}_{T_\varepsilon \cup T}| = 2^{7|T|} \tag{2.78}$$

Then for finite classes $\mathcal{H}$, for any set T, we have,

$$|\mathcal{H}_T| \leq |\mathcal{H}_{T_\varepsilon \cup T}| \leq |\mathcal{H}| \tag{2.79}$$

Thus, T cannot be shattered if,

$$|\mathcal{H}| \leq 2^{|T|} \tag{2.80}$$

and T_ε cannot be shattered if,

$$|\mathcal{H}| \leq 2^{7 \times |T|} \tag{2.81}$$

The fuzzy set has finite but larger VC dimension if the sample has finite VC dimension. This implies the following for VC dimension of the fuzzy set T_ε,

$$\text{Dim}_{vc}(\mathcal{H}_{T_\varepsilon \cup T}) \leq \log_2(|\mathcal{H}|) \leq 7 \times |T| \tag{2.82}$$

This will further implies the PAC learnability of $\mathcal{H}_{T_\varepsilon \cup T}$ from fuzzy set T_ε. □

Note, the VC dimension can be much smaller than the above bound.

2.5.2 Loss function in PAC learning

In this entire documents, we introduce two types of loss functions that are commonly used in machine learning theory. For categorical outcomes $Y = y$, we use 0-1 loss function l_{01} in ERM process. It is defined as, for algorithm $h \in \mathcal{H}$,

$$l_{01}(h, (x, y)) = 0, \text{ if } h(x) = y; \text{ else } l_{01}(h, (x, y)) = 1 \tag{2.83}$$

It is very common to use this type of loss function in classification, either classified right or complete wrong [57].

When the continuous outcomes $Y = y$, we commonly use square loss function l_{sq} as,

$$l_{sq}(h, (x, y)) = (h(x) - y)^2 \tag{2.84}$$

In this case, the expected risk minimization process is the same with least square in regression, and the loss function is the same as the square error.

We defined ε-close by the L_1 norm, i.e. the distance between $Z^*_{X_i}$ and X_i is defined as $\|\varepsilon\|_1$. Suppose X_i is r-dimensional. Recall the fuzzification process,

$$\vec{Z}^*_i = \vec{X}^*_i + \vec{\varepsilon}$$

All values in the above process are r-dimensional. The distance between the fuzzified vector and original sample vector is A ball covered X_i and generated by Z_i is defined as,

$$\|\varepsilon\|_1 = \max_{1 \le k \le r} \varepsilon_{ik} \tag{2.85}$$

It is the smallest ball generated by the additional noise centered at X_i such that all the fuzzified observations lie in the ball B_ε. In this way, we define the prediction is ε-close with probability $F_Z(|\varepsilon|)$, where F is the cumulative distribution function (CDF) of standard normal distribution.

2.5.3 Probably Approximately Correct Learning

"*Not everything that can be defined can be computed*" [63]. But using enough sample with great representativeness, we can find the approximately correct prediction with certain level of confidence that it is almost right.

The major result we show in this section is that the fuzzy system outlined in this dissertation is PAC learnable. We first recall the definition of (Agnostic) PAC learnable and predictive PAC learnable. In this entire chapter, we restrict $\mathcal{R}$ on the finite response space, i.e. grid point only. Here we always refer to (Agnostic) PAC learnable, without assuming

the perfect algorithm that link the response space $\mathbb{R}$ to clinical outcomes Ω. Given the finite training sample T_n and hypothesis space $\mathcal{H}$, the leaning algorithm is derived from,

$$h_{T_n} = \Upsilon(T_n, \mathcal{H}) \tag{2.86}$$

Definition 2.11. (PAC learnable). A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}}$ of $|\mathcal{H}|, \epsilon, \delta$, such that when the sample size $m \geq m_{\mathcal{H}}$, then the derived algorithm h has the following property of,

$$\underset{x \in \mathbb{R}}{P}\left(\sup_{h \in \mathcal{H}} |L_T(h) - L_{\mathbb{R}}(h)| \geq \epsilon\right) \leq \delta \tag{2.87}$$

The idea is, given the finite hypothesis space, the approximate error, the confidence level, we can derive the PAC-learnable algorithm from the larger than calculated samples size such that the prediction error compared to all available algorithms in hypothesis space is small enough with high chance. The difference between PAC learnable and Agnostic PAC learnable is that whether there exists a perfect algorithm h^* that link every observation in $\mathbb{R}$ to Ω.

Now we prove that when the hypothesis space is finite, and when training set is PAC learnable, the learning error from the noisy data can be bounded by enlarge the sample size. We introduced an important inequality, Hoeffding inequality [57], which is used to prove the later theorem.

Definition 2.12. (Hoeffding inequality) Let X_1, X_2, X_n be i.i.d. observations with expectation μ and $\forall i, P(X_i \in [a, b]) = 1$. Then for any $\epsilon > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{2.88}$$

Theorem 2.8. *Assume the hypothesis space is finite defined as $\mathcal{H} = \{h_1, h_2, \cdots, h_m\}$. The training observations is T of size n and fuzzy set with noise generated directly from T is denoted as T_ε. The approximation error is bounded and it is PAC learnable.*

Proof: We define the possible misleading samples as $\mathcal{B}_k, k = 1, \cdots, m$, the loss for these data points are larger than expected with respect to the algorithm h_k,

$$\mathcal{B}_k = \{(X_i, Y_i) \in \mathbb{R}, i = 1, \cdots, n : L_{T_\varepsilon}(h_k) - L_T(h_k) \geq \epsilon\} \tag{2.89}$$

Then probability of existing a $h \in \mathcal{H}$ with large approximation error in the fuzzy set system is,

$$\begin{aligned}
&P(\exists h \in \mathcal{H} : L_{T_\varepsilon}(h) - L_T(h) \geq \epsilon) \\
=&P\left(\cup_{k=1}^{m}(L_{T_\varepsilon}(h_k) - L_T(h_k) \geq \epsilon)\right) \\
\leq&\sum_{k=1}^{m} P((L_{T_\varepsilon}(h_k) - L_T(h_k) \geq \epsilon)) \\
\leq&\sum_{k=1}^{m} \left(P(L_{T_\varepsilon}(h_k) - L_{\mathbb{R}}(h_k) \geq \epsilon) + P(L_{\mathbb{R}}(h_k) - L_T(h_k) \geq \epsilon)\right) \\
\leq&2me^{-2n\epsilon^2}
\end{aligned} \tag{2.90}$$

The last part of the proof applied Hoeffding's Inequality. Let the above result (2.90)=δ, we have the following,

$$\epsilon = \sqrt{\frac{1}{2n} \times \log \frac{2m}{\delta}} \tag{2.91}$$

By control the level of two parameters ϵ and δ, we can estimate the sample size for each iteration when adding noise for the fuzzy system. □

Now we are ready to prove one of the major results. For any response in the sample space $X_i = \{x_i\}_{j=1}^{p} \in \mathscr{X}$. Usually $\mathscr{X}$ is not complete, i.e. not all points on the grid level can be covered. $\mathscr{X}$ is finite, with at most νp grid points. Here ν is the number of options for each $\{x_{ij}\}$. Without loss of generalizability, we assume the number of options are equal for any j. Hence, on the sample space $\mathscr{X}$, the ranging of the points is $[1, 2, \cdots, \nu p]$.

Theorem 2.9. *Let $\mathcal{H}$ be the hypothesis space of functions from the the sample space $\mathscr{X} \in \mathbb{R}$ to dichotomous outcomes $\{0, 1\}$. To find the optimal threshold in $\mathcal{H}$ to classify $X_i = \{x_i\}_{j=1}^{p}$ that is PAC learnable, with respect the approximation parameter and confidence parameter $\epsilon \times \delta \in (0, 1) \times (0, 1)$, we show $\mathcal{H}$ is PAC learnable as long as the sample satisfy,*

$$n_{\mathcal{H}} \geq \frac{\log(2\nu p) - \log \delta}{2\epsilon^2} \tag{2.92}$$

Proof: The formula $n_{\mathcal{H}}$ is directly derived from the error bound in (2.91). The size of hypothesis space is $|\mathcal{H}| = \nu p$. During the defuzzification process at observation level, we

already project the fuzzy observations on the grid of $\mathscr{X}$. In this case, $|H|$ did not change, though training observations may shift. □

This theorem ensures the learning from fuzzy system with 0-1 loss function, is PAC learnable. A simple example for the PAC learnable sample requirement is as below. Assume we have a total of 20 questions, with each question 5 options. If we need the approximation of the error is less than 0.01 and with at least 80% confidence, then the required learning sample size would be,

$$N \geq \frac{\log(2 \times 5 \times 20) - \log(1 - 0.8)}{2 \times 0.01^2} \approx 17270 \tag{2.93}$$

When applied the fuzzy system, we can generate enough bootstrap sample observations, together with the original observations in order that sample space is as close as $\mathscr{X}$. Hence the learning is PAC learnable with approximate error less than 0.01 and the sample size is greater than 17270.

In practice, if we can tolerate approximate learning with error 0.1 and but with a confidence of roughly 90% level of confidence. The sample size is reduced significantly to about 380, which is later on our case in practice.

2.5.4 PAC learning for tree-based model

The most popular way to construct machine learning algorithm is to use a hypothesis base class B. We will show the VC-dimension is finite for the fuzzy system with a linear combination of base hypothesis. Assume the hypothesis class of tree-based model has the VC-dimension d.

We state the follow lemma first before the main theorem of tree based algorithms. The following lemma about the growth function $\tau_{\mathcal{H}}$ is called Sauer-Shelah-Perles Lemma [57].

Let $\mathcal{H}$ be a finite hypothesis class with VC-dimension less than or equal to d. Then for all m, we have the growth function,

$$\tau_{\mathcal{H}} \leq \sum_{i=1}^{d} {}_m\mathrm{C}_i \tag{2.94}$$

If $m \geq (d+1)$, we have,

$$\tau_{\mathcal{H}} \leq \left(\frac{em}{d}\right)^d \tag{2.95}$$

The growth function is the number of different functions from a set of size m to $\{0, 1\}$.

Theorem 2.10. *The VC-dimension in the fuzzy system is finite and bounded by the following formula.*

$$T(2d+2)(3\log(T(2d+2)+2) \tag{2.96}$$

Hence the system with tree-based model is PAC-learnable.

Proof: Assume C is a set that can be shattered by fuzzy system, with $\{X_1, X_2, \cdots, X_m\}$. Using Sauer's lemma [57], there are at most

$$\left(\frac{em}{d}\right)^d \tag{2.97}$$

different ways of combinations induced by the hypothesis B over its shattered set C. If we need to choose T hypothesis for the tree leaves of the observed data, there are at most below ways to do it for the fuzzy system,

$$\left(\frac{em}{d}\right)^{2dT} \tag{2.98}$$

As the manually introduced noise may added one more possible leaves for some observations, with probability less than or equal to 0.5. We assume it is 0.5 for simplicity. Using Sauer's lemma again, we have at most following number of choice for the linear predictors,

$$\left(\frac{em}{2T}\right)^{2T} \tag{2.99}$$

Then total number is the product of 2.97 and 2.99,

$$\left(\frac{em}{d}\right)^{2dT} \times \left(\frac{em}{T}\right)^{2T} = \frac{e^{2dT}e^{2T} \times (m)^{(2d+2)T}}{d^{2dT}T^{2T}} \leq m^{(2d+2)T} \tag{2.100}$$

Use the fact d and T are greater than 2. As C is shattered, then

$$2^m \leq m^{(2d+2)T} \tag{2.101}$$

Use the fact [57], if $a > 0$ and $x < a\log(x)$, then we have $x < 2a\log(a)$. Hence we have the bounded function below,

$$m \leq T(2d+2)(3\log(T(2d+2)+2) \tag{2.102}$$

This concludes the fuzzy system with tree-based model is PAC-learnable, as long as we can bound the total number of leaves T and total number of trees d. □

2.5.5 Final prediction

The learning module and prediction module can be simplified as,

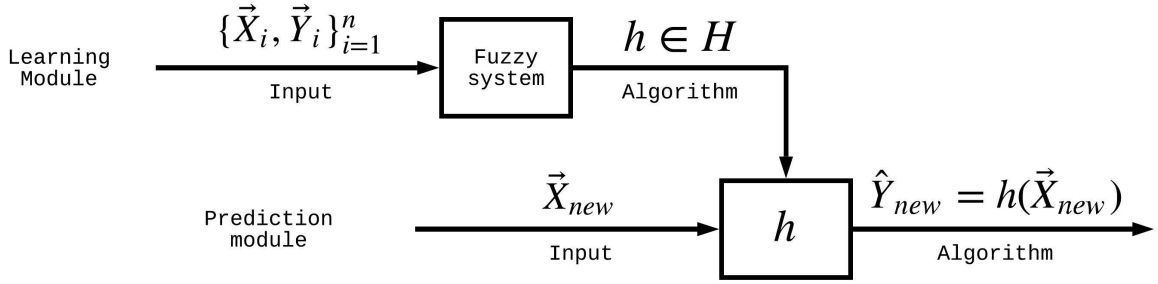

Figure 2.12: The learning and prediction modules for fuzzy system

The generalization version of valiant's Probable Approximately Correct (PAC) learning first answered the fundamental question about what is learning [57].

We assume the real relationship between the survey data and clinical outcome can be described by the unknown function f, i.e. $Y_i = f(X_i) + \Psi$, where Ψ is the unattainable error between the survey response space and clinical outcomes. It indicates that the survey questions can never replace a diagnosis procedure. There is always some distance that can not be measured by surveys.

Assume the survey is from a certain type of distribution $\vec{X} \sim \mathcal{D}$, the all possible algorithm space is $\mathcal{H}$. The true error of the learning from the training sample is $h \in \mathcal{H}$ is,

Definition 2.13. (True error). With respect some loss function $L(h, \mathcal{D}, f)$ is,

$$L(h, \mathcal{D}, f) = P_{\vec{X} \sim \mathcal{D}}(h(\vec{X}) \neq f(\vec{X})) \tag{2.103}$$

As in most of the case, the distribution of the survey responses $\mathcal{D}$ and the true relationship f are not known. The training error is used to estimate the true error, which is defined as,

Definition 2.14. (Training error). With respect some loss function $L(h, \mathcal{D}, f)$ is,

$$L(h, \{\vec{X}_i, \vec{Y}_i\}_{i=1}^{n_{train}}) = \frac{\#|\{i \in [1, n_{train}] : h(X_i) \neq Y_i\}|}{n_{train}} \tag{2.104}$$

The training error is also called Empirical Risk Minimization (ERM) if we consider $\mathcal{P} = P_{X_n}$ and $\mathcal{D} = D_{X_n}$, where the two quantities are defined as,

$$P_{X_n}(X_n = x) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \vec{1}_{\{X_n = x\}} \tag{2.105}$$

The distribution function is,

$$f_{X_n}(x) = Y_i, \text{ if } x = X_i; \text{ otherwise } f_{X_n}(x) = 0 \tag{2.106}$$

In this case, the training error is ERM,

$$L(h, \{\vec{X}_i, \vec{Y}_i\}_{i=1}^{n_{train}}) = L(h, \mathcal{D}, f_{X_n}) \tag{2.107}$$

The algorithms developed based on ERM have perfect performance on the training set, but very poor on the test set, i.e. overfitting [57]. One of the solution in the literature is to searching the algorithms within a pre-determined hypothesis space $\mathcal{H}$. This pre-determined hypothesis space is restricting the learning system to find a good predictor, sometimes called an inductive bias. It is always determined before collecting the data based on some previous knowledge. This idea can be formulated in plain language as,

$$\text{Data} + \text{Prior Knowledge} = \text{Generalization} \tag{2.108}$$

The fundamental questions in machine learning theory is that how to choose the hypothesis space $\mathcal{H}$ in order that the resulting algorithm will not result overfitting.

In PAC learning, we have two level of uncertainty. One level is the accuracy parameter ϵ and the other level is the confidence parameter δ. It has been proved by Haussler in 1990 [64] when the hypothesis space $\mathcal{H}$ is finite, then algorithm h from $\text{ERM}_{\mathcal{H}}$ is generalizable when sample size is sufficient large.

CHAPTER 3

Data Application

In this chapter, our fuzzy system is applied to data collected from the Patient Reported Outcome Measurement Information System (PROMIS) Oral Health project [65]. In previous publications, we described more details about the process of designing survey items [66, 67], calibrating item banks [36], and developing short forms [19, 17]. We developed two versions of short forms, a parents' version and a children's version. The short forms consisted of a limited number of survey questions to measure oral health, but had information (inverse of reliability) comparable to the long forms. In this project, we use two short forms to collect information from children and their parents about the child. Dental examinations were conducted independently from the survey by two dentists. We checked the consistency and agreement between two dentists and recorders using both kappa and prevalence-adjusted and bias-adjusted kappa (PABAK) [68]. The agreement between the dentists was high ($> 85\%$). We use the fuzzy system to predict clinical outcomes from survey questions, using parents' short form only, children's short form only, combination of two short forms, and with additional demographic information. The application part of this dissertation has been submitted to Journal of Dental Research Clinical & Translational Research. The statistical methodology was presented at the Joint Statistical Meetings in 2018. The application was presented at International Association for Dental Research (IADR) at London, United Kingdom in 2018.

3.1 PROMIS Oral Health overview

In this example, we predict three clinical outcomes, referral for treatment needs (RFTN), children's oral health status index (COHSI) and COHSI percentile. These three outcomes cover three common types of predictions in Machine Learning: the categorical outcomes for classification; the continuous variable with a fixed range; and the rank variable or percentile.

3.1.1 Survey development

The process of reviewing existing items and developing new survey items was described in previous publications. Before creating a PROMIS oral health survey item bank, we reviewed all existing instruments that measure children's oral health by both self-reported and parent-reported methods. Our target population was children aged from two to seventeen. In PROMIS, survey questions are only applicable to children of eight years old and above. We organized focus group interview sessions to collect opinions on measuring oral health from both parents and two children groups aged 8 to 12 and 13 to 17. Based on existing survey instruments and focus group results, we drafted the first version of survey items. We organized five cognitive interviews per age group of children and parents to ensure the survey questions were understood by children and by parents. All questions were designed using the PROMIS approach.

In PROMIS, health is measured by self-reported items from four components: global health, physical, mental and social health domains. Each component has been further extended to the next level [65, 17], such as pain, symptoms, etc. The conceptual model of the domains, sub-domains are illustrated in Figure 3.1. The conceptualized model expanded oral health into three components: physical, mental and social health. Each component (orange in Figure 3.1) was further extended to sub-components (green), domains (purple), and sub-domains (blue). The colors of the block in Figure 3.1 indicated different levels of structure. The gray colored domains did not directly measure oral health status and therefore were not included in the item bank.

3.1.2 Field testing

In field testing, we targeted to collect survey responses and dental examinations from at least 500 families [36]. The data were collected using a convenience sample from dental clinics located in Los Angeles County from August 2015 to October 2018 as shown in Figure 3.2. The location of the clinics ranged from Torrance (south) to Valencia (north) and from Agoura Hills (west) to Whittier (east), to represent children and adolescents who had visited dental clinics in Los Angeles County. Only one child was chosen from each family as a stratified systematic convenient sample aimed at reflecting the race and ethnicity distribution of the general California population with a very similar percentage representation. Those who did not speak English and who were in orthodontic treatment were excluded from the study. The survey responses were directly entered into the computer using an audio computer-assisted self-interview software (ACASI). In this case, each survey question must be answered in order to move to the next step. There is no missing input in the survey part. The only possible missing values may happen because the younger kids do not cooperate with the clinical exam. This was a rare situation and happened less than 2%. These observations were excluded from the analysis because there was no clinical outcomes available.

A total of 545 surveys were collected from parents and 363 surveys were collected from children aged eight or older. The target sample sizes were based on rules of thumb for estimating IRT models, i.e. in most scenarios, a sample size of 500 could derive accurate estimation for Graded Response Models (GRMs) parameters [69, 70]. One limitation of the study is that the study participants were sampled from dental clinics selected conveniently from Los Angeles County. The majority of families are considered as having a usual source of preventive dental care. Therefore the overall oral health status of the sample is better than the general population.

3.1.3 Clinical outcomes

There are two summary clinical oral health outcomes from the on-site dental exam, the Children Oral Health Status Index (COHSI) score and referral for treatment needs (RFTN).

COHSI is a weighted regression of facial profile, occlusion status, presence of active caries, abnormal position, missing and filled teeth [31, 34], etc. The COHSI has a best possible score of 100 and decrements from that are estimated by multiplying previously derived regression coefficients by number of missing teeth (-2.27 for primary anterior teeth, -4.55 for primary posterior teeth and permanent teeth), number of decayed teeth (-1.12 for primary anterior teeth, -2.24 for primary posterior teeth and permanent teeth), occlusion status (-4.38), and abnormal positions (-1.73) (see reference 32 for details). The worst possible COHSI score is -27.4 for adolescents with all permanent teeth missing) and 18.16 for younger children with all primary teeth missing. The observed range of the COHSI in the field test was 59.18 to 100.

RFTN is a treatment or dental clinic visit need with "1" denoted for having at least one teeth with active caries, or at least 12 teeth bleeding upon probing. It is derived from the 4-level guidelines used in the National Health and Nutrition Examination Survey (NHANES).

3.1.4 Short forms

The samples collected in field testing were used to develop two versions of short forms for children [17] and for parents [18] separately using GRMs in Item Response Theory (IRT). Through the model, the items are calibrated with slope (discrimination parameter) and threshold (difficulty) parameters to quantify the item characteristics [71]. The short form questions were selected based on the two above parameters (discrimination and difficulty), domain coverage of oral health conceptual model, and the suggestions from dental experts.

The response space is defined from the input of both the child [17] and parent short form [18] of short forms. Intuitively, we assume the combined information from both child and parent can yield more accurate predictive results than using the information individually. With additional demographic information, the predictive results should be improved. The basic assumptions when developing short forms include: (1) monotone, that is, items have a monotonic relationship with the latent variable; (2) local independence or the conditional independence among the survey items given the latent variable; (3) Uni-dimension of latent

variable, that is the selected items measure the same latent variable; (4) no differential item functioning (DIF), that is, survey items do not function differently among different groups. The properties of short forms link the survey responses of X_i to latent variables θ, which should have high correlation with the clinical exam outcomes Y_i. In GRM, each survey item is parameterized by one discrimination parameter and several difficulty parameters. These parameters determine the characteristic of the survey item and determine whether the survey item is selected in short forms [17, 18].

The input to the fuzzy system includes the two outcomes RFTN and COHSI (score and percentile), the developed short form questions including 12 self-reported survey questions, 8 proxy-reported questions, and 7 demographic questions. The demographic information includes age, gender, race/ethnicity of the child, dental insurance, visit to emergency room for a dental problem, number of children in the household, and relationship with parents.

In Table 3.1 and Table 3.2, the survey questions are listed by either the clinical outcomes or both it endorsed. In children's short form, there are four questions related with RFTN, five questions related with COHSI, and three questions related to both outcomes. The two clinical outcomes can still be separated for overall and long-term oral health status (COHSI) versus the current need for treatment (RFTN). From the parents' short form, there is one question for each clinical outcome and all other six questions are for two outcomes together. As expected, the two clinical outcomes are not obviously different from parent's point of view [18].

3.2 Methods

The original observations were the crisp set with sample size $N = 545$, labeled as $X_i, i = 1, \cdots, N$. We generated 1000 bootstrap samples (total sample size $1000N$). Each bootstrap sample had N observations with independent and identically distributed noise added manually to the sample. In inference layer, as outlined in the previous chapter, the sample was divided into 70% for training the algorithm and 30% for testing. We called this test set internal test because the test set was separated from the original sample and was repeatedly

Table 3.1: Table of short form questions for children

Outcome	Survey Questions
RFTN	It was hard for me to eat because of the pain in my mouth.
RFTN	It was hard for me to pay attention because of the pain.
RFTN	How often do you use dental floss on your teeth?
RFTN	Do other students make jokes about the way your teeth look?
COHSI	It hurts my teeth to chew.
COHSI	My teeth are straight.
COHSI	How much are you afraid to go to a dentist?
COHSI	How often do you brush your teeth?
COHSI	Have you ever avoided laughing because of teeth look?
Both	In general, would you say your overall oral health is:
Both	In the last 4 weeks, how much of the time did you limit food?
Both	How much of the time were you pleased with your teeth look?

used in training modules. Another set of observations (we called external test set) were additionally collected using the same survey instruments and dental examination protocol. This test set was only used to test the generalizability of the algorithms. We called this test set externally because the prevalence of dental disease was higher than the original sample though using the same protocols and methods. Most of the patients in this test set were newly enrolled patients without evidence of preventive care. We reported the prediction results of both two test set. The training algorithms were trained using observations from original data, using bootstrap samples only, and using Fuzzy system.

When training the algorithms, the sensitivity was required to be at least 85%, with the corresponding specificity recorded for purposes of comparison. When training the algorithms based on comparing predicted outcomes with observed clinical outcomes, classification was based on achieving a sensitivity that either exceed 85% or that was as high as possible with specificity at least 20%. In the example, the resulting sensitivity could be either 82% or

Table 3.2: Table of short form questions for parents

Outcome	Survey Questions
RFTN	School miss due to teeth problems in the last year
COHSI	My child's mouth hurts.
Both	It was hard for my child to eat because of tooth pain.
Both	In general, would you say your child's oral health is:
Both	In the last 4 weeks, pleased with teeth look
Both	Worry about problems with his/her tongue, teeth, or gums?
Both	It was hard for my child to pay attention because mouth pain.
Both	In the last 4 weeks, oral health interfere with social activities?

86%, depending how many subjects that have predicted value the same as observed value by choosing the cutoff probabilities in the algorithm. In this case, we always selected the cutoff value with sensitivity as close to 85% as possible, i.e., 86% is sensitivity was selected.

3.2.1 ALgorithms

The categorical outcomes were predicted using the Naive Bayesian (NB) method. The results were reported as sensitivity and specificity. The sensitivity was pre-determined as more than 85% to find the best cut-off value for high specificity.

There are a few parameters that need to be fine-tuned for XGBoost [52, 53]. For example, we use greedy algorithm to search the best parameters (Friedman 2001). Eta controls the weights of subsequent trees is searched from 0.05 to 0.4 (default is 0.3). The maximum depth is searched from 2 to 8 (default is 6). The maximum number of trees is from 5 to 500 (default is 10). Regulation parameter lambda is searched from 0 to 0.4 (default is 1, means to use L_2 regulations on weights), higher value means more conservative model.

3.2.2 Software

The algorithms are available in R packages. For NB, the package *e1071* is used [72]. The visualization of the algorithm is using for Nomogram [73] uses Orange (an open-source data visualization, machine learning and data mining toolkit) [74]. For XGboost, we use *xgboost* developed by Chen in 2015 [75] with some updates in 2018 [76].

3.3 Results

3.3.1 Characteristics of participants

Total number of survey responses (8 questions) from parents is 545, with an additional 363 responses (12 questions) from children eight and older. All 545 children had on site dental examinations (we exclude those observation without exam data). Seven demographic questions are included in addition to the short form questions. The mean score for COHSI is 90 with range from 55 to 100 and median 92 (skewed to the left). Thirty-one percent of children were identified as having a need for treatment or dental clinic visit. Table 3.3 presents the characteristics of the children, parents and the household information [19].

Table 3.3: Characteristics of the sample (children, parents and household)

Variables	Mean (SD) or N (%)
COHSI	90.45 (8.5)
RFTN	169 (31.1%)
Survey items	Mean (SD) or N (%)
Children's age (in years)	9.7 (4.2)
Children's age group	
2 to 7 years old	182 (33.4%)
8 to 12 years old	214 (39.3%)
13 to 17 years old	149 (27.3%)
Children's Gender	

Table 3.3 continued from previous page

Variables	Mean (SD) or N (%)
Male	280 (51.4%)
Female	264 (48.4%)
Male to Female Transgender	1 (0.2%)
Children's Race/Ethnicity	
Caucasian/White	111 (20.4%)
Black/African American	50 (9.2%)
Hispanic/Latino	226 (41.5%)
Asian	59 (10.8%)
Other	99 (18.2%)
Parent's Gender	
Male	160 (29.4%)
Female	385 (70.6%)
Parent's age group (in years)	
Less than 30 years old	67 (12.3%)
30 to 44 years old	302 (55.4%)
45 to 59 years old	161 (29.5%)
60 years old and above	15 (2.8%)
Parent's Primary Language	
English	394 (72.3%)
Other	151 (27.7%)
Number of kids in the family	
1	130 (35.8%)
2	129 (35.5%)
3	54 (14.9%)
≥ 4	50 (13.8%)
Child Dental Insurance	

Table 3.3 continued from previous page

Variables	Mean (SD) or N (%)
No	111 (20.4%)
Yes	434 (79.6%)
Employment	
Full-time job	429 (78.7%)
Part-time job	61 (11.2%)
Not working	55 (10.1%)

3.3.2 Prediction results

The performance of the algorithm is evaluated by accuracy (both sensitivity and specificity). The best prediction (higher accuracy and more stable prediction parameters) is from the algorithm using the most available information, i.e. using short forms responses from both child and parent in addition to demographic information. Whenever there is new information added, the prediction accuracy improved. In the table, we compared the prediction performance of four potential algorithms, using only children's responses; using only parents' responses; using both short forms responses with and without demographic information. For each algorithm, we compare the prediction accuracy by the training samples generated from original sample, bootstrap samples, and Fuzzied samples. The results are shown in Table 3.4.

The product algorithm of NB is visualized in the nomogram as illustrated in Figure 3.3. The nomogram is a convenient output tool from NB prediction. It can be used to predict the probability of treatment needs without using a computer or calculator but circle, lines and rulers. The contribution of each item is directly printed out, which is summed together to transfer to posterior probabilities [74]. The missing response from single item is imputed by average information of peers in the field test sample. The bottom of Figure 3.3 maps the probability of RFTN and total points. The circle represents the prior probability

Table 3.4: Result for categorical outcomes using original sample, bootstrap sample only and fuzzy system

Algorithm Toolkits	Sample (N=545)	Internal test Sensitivity	Internal test Specificity	External test Sensitivity	External test Specificity
Children (C)	Original	86%	20%	79%	19%
	Bootstrap only	85%	18%	86%	48%
	Fuzzy system	85%	24%	71%	29%
Parents (P)	Original	86%	28%	100%	24%
	Bootstrap only	85%	23%	86%	43%
	Fuzzy system	85%	31%	79%	33%
Combine	Original	86%	28%	93%	24%
	Bootstrap only	85%	21%	86%	38%
	Fuzzy system	85%	37%	93%	48%
Add demo	Original	86%	26%	93%	24%
	Bootstrap only	85%	23%	86%	52%
	Fuzzy system	85%	35%	93%	49%

Sample size: Original (N), Bootstrap only ($1000N$), Fuzzy System ($1000N$)

from the field test sample (31%). The total points are the sum of the points from each item in the figure. On top of the figure, it is the point for each item. For example, if the parent responded always to the question "*it is hard for the child to pay attention due to pain in his/her mouth*", then the point for this item is 2.0, corresponding to 78% in need of treatment or dental visit. From the nomogram plot, the demographic items contribute little information to RFTN prediction.

There are two types of continuous outcomes, the actual index score COHSI ranges from 55 to 100 as well as its percentile from 0% to 100%. The classical ways to evaluate the agreement between two continuous variables are the Pearson correlation coefficient for linear trend and root mean square error (RMSE) for absolute difference. We use both statistics to

evaluate the results. Table 3.5 and Table 3.6 show the results.

Table 3.5: Result for COHSI from XGboost Algorithm

	Pearson correlation		RMSE	
	Internal test	External test	Internal test	External test
Original				
Children	0.39	0.29	8.21	9.16
Parent	0.32	0.34	8.37	8.71
Combine	0.49	0.21	7.78	11.28
Add demo	0.41	0.37	8.23	9.06
Bootstrap				
Children	0.39	0.30	7.79	9.14
Parent	0.29	0.20	8.09	9.33
Combine	0.50	0.20	7.32	9.29
Add demo	0.92	0.17	3.27	9.8
Fuzzied sample				
Children	0.59	0.58	7.03	7.31
Parent	0.28	0.29	8.39	8.61
Combine	0.65	0.66	6.65	6.74
Add demo	0.90	0.88	3.89	4.19

We compare the results from fuzzy system with the results using bootstrap procedure without adding noise. The training model is not generalizable to test results as shown in the middle of Table 3.5 and Table 3.6. The correlation and RMSE is smaller in the test set of training module but blows up when applied to new data.

Table 3.5 and Table 3.6 report the four versions of algorithms derived by children's short form, parents' short form, combined two versions with and without demographic information. The training module developed algorithms using original sample, bootstrap sample,

and Fuzzied sample. In this illustration, the algorithm from the fuzzy system performed much better than original sample and the bootstrap sample only. To obtain pre-determined sensitivity of RFTN, the algorithm from fuzzy system good level of specificity in external test set. The Pearson product-moment correlations are high for all four versions of algorithms when using fuzzy system. The correlation increased as more information was used. The apparent interpretation for why the performance was better in the noise-added condition is that the additional sample size beyond more than the original 545 compensated for the added noise and because the training sample and test sample were similar. Also the repeated use of demographic information generating more similar samples led to high correlation and smaller RMSE. This is under the assumption that people with the same demographics answer surveys similarly.

In Table 3.5 (Table 3.6), the correlation of predicted COHSI (percentile), with raw COHSI (percentile) was 0.41 (0.43), but with fuzzy system output COHSI was 0.90 (0.92) in the training results. In testing results, the correlation coefficient of predicted COHSI (percentile), with raw COHSI was 0.37 (0.39), but with fuzzy output COHSI was 0.88 (0.91). The performance in Table 3.6 was more stable and generalizable to new data because the test results were comparable to the training results. The performance is more stable in the test data. In the results from the fuzzy system, with information source added, the correlation increased and the RMSE decreased. The best performance was achieved when all information was used, i.e. both short forms and demographic information. The RMSE for predicted COHSI score (percentile) was 3.89 (1.12) in fuzzy system training results, with 4.19 (1.26) in testing data from additional source. These results indicated the generalizability of the algorithm trained by Fuzzied sample.

3.4 Conclusion

In this example, we applied the PAC learnable fuzzy system developed from previous chapters to predict categorical outcome RFTN and continuous outcome COHSI with its percentile. The training algorithms are derived using NB method and XGboost method from samples

generated using original sample, bootstrap samples, and bootstrap samples with manually introduced noise (Fuzzied sample). We manually picked the algorithms that performed the best in both training and in the additional collected data (external test set). The association between the survey outcomes and clinical exam results are not related to previous exams. We expected the algorithm can be generalized to this test data and performed better because COHSI score of this additional site was lower and the prevalence of RFTN was higher than training data. We selected the most generalizable algorithm to report in Table 3.5 (Table 3.6). Though we maintained the high level of similarity between the test set and training set, the data in test set was collected independently. The clinical outcomes (COHSI and RFTN) were directly collected from dental examinations performed by the same dentists. We can treat the test set as an external test.

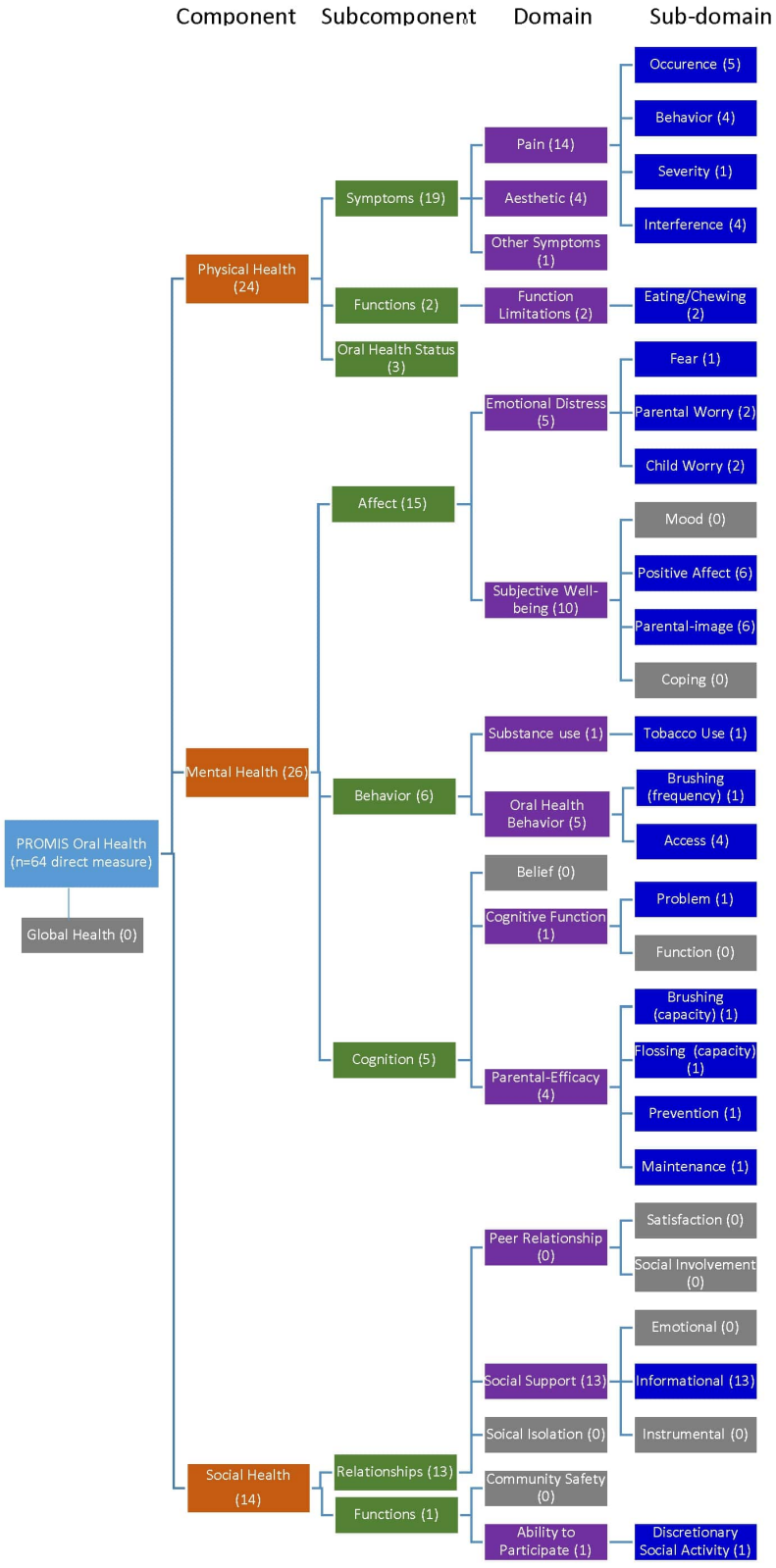

Figure 3.1: Conceptual model of PROMIS items

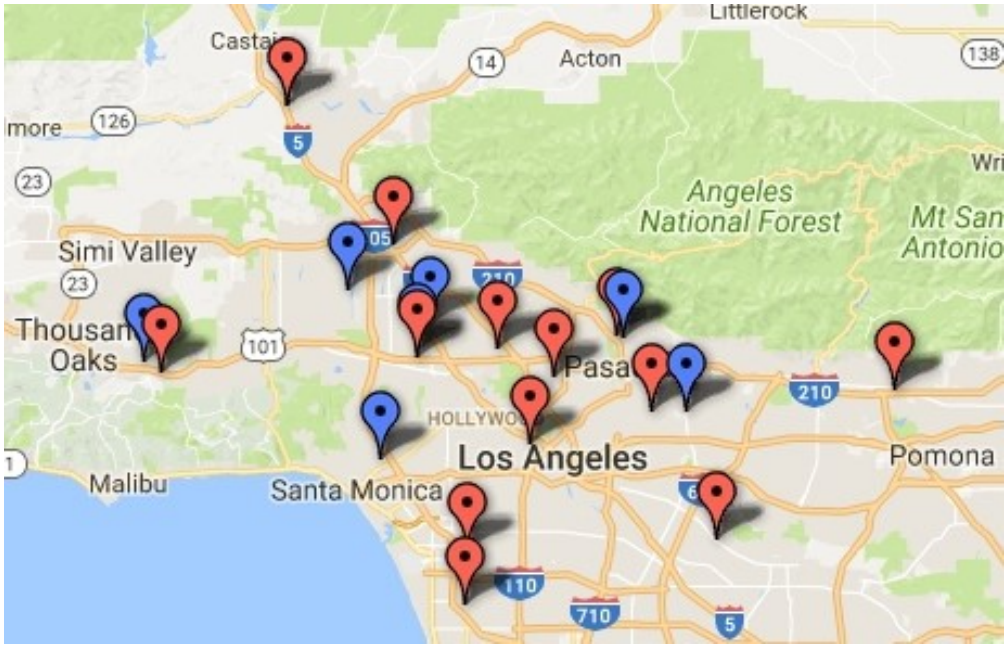

Figure 3.2: Field test samples collected from dental clinics in Los Angeles County

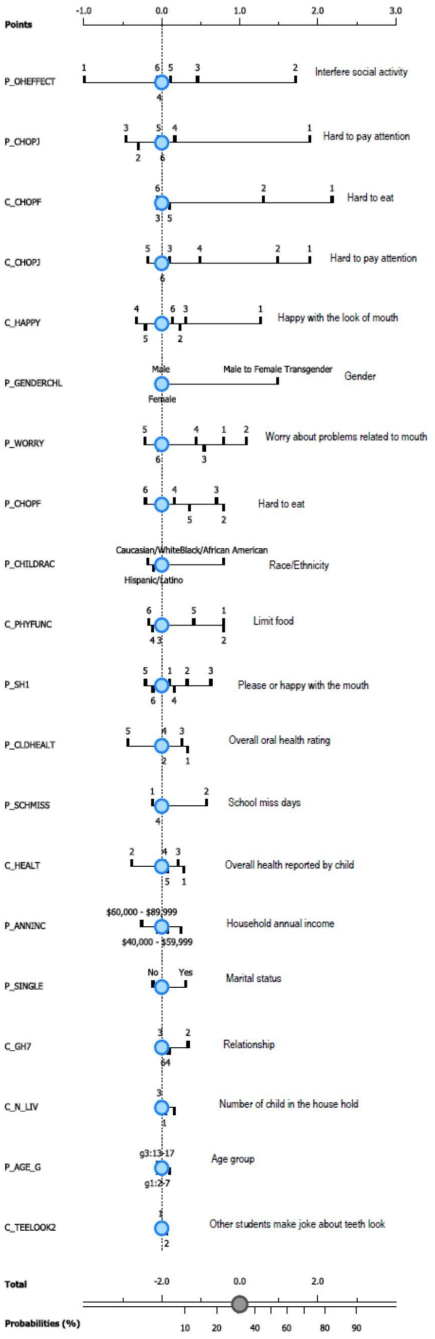

Figure 3.3: Nomogram of Naive Bayesian Model

Table 3.6: Result for rank (percentile) of COHSI from XGBoost Algorithm

	Pearson correlation		RMSE	
	Internal test	External test	Internal test	External test
Original				
Children	0.41	0.21	2.78	3.27
Parent	0.31	0.29	2.88	3.09
Combine	0.51	0.29	2.59	3.16
Add Demo	0.43	0.39	2.73	2.90
Bootstrap				
Children	0.40	0.25	2.71	3.20
Parent	0.30	0.17	2.82	3.29
Combine	0.49	0.24	2.56	3.13
Add demo	0.91	0.26	1.24	3.53
Fuzzied sample				
Children	0.58	0.60	2.38	2.41
Parent	0.26	0.27	2.85	2.91
Combine	0.66	0.67	2.21	2.22
Add Demo	0.92	0.91	1.12	1.26

CHAPTER 4

Discussion

This dissertation presents the development of a fuzzy system based on the input value of survey responses to predict clinical outcomes. The vagueness of the harsh options when completing the survey motivated us to use fuzzy set membership functions to describe the input values. The input observations from survey sampling are usually not large and complete enough to cover all the possible combinations in the design space. In most models, e.g. IRT, require each option endorsed by at least three observations to have a stable estimate of the threshold value. The goal for the fuzzy system is to grow enough sample size to derive the machine learning algorithms based on the available resources to predict the new input. The main aim for the algorithm is to predict the outcome with some level of uncertainty from new input from the survey, which is Probably Approximately Correct (PAC) learning. The new input may or may not come from the same sample.

There are various bootstrap re-sampling methods developed to handle complex survey structure, weighting, imputing and small area estimation [29]. In this dissertation, we only focused on simple re-sampling methods that is, simple random sampling with replacement. We do not consider the weighting schema when we grow our sample size, but some weighting strategies or re-sampling methods (e.g. stratified simple random sample, stratified multi-stage sampling, balanced repeated replications, and mirror match bootstrap) have great potential in representing new population [77, 78].

The membership function is a step function that is in between the Trapezoid and standard normal distribution, though it is generated by introducing the standard normal noise. The distribution of the noise means we treat each response as normally distributed according to its own observed values. The variance introduced to each option equals 1. In the first step of

data-level defuzzification, the noisy observations are fuzzified at the grid level with grid point determined by the survey. This step ensures the prediction is only from the response space. Mathematically, this step changes the membership function from a Gaussian distribution to a step function with pre-determined probabilities. The center has highest probability among all noisy observations. In the future fuzzy process, we may consider the original observation with raw noise (continuous over response space). This may generate more unique observations that may smooth the prediction.

The combination of fuzzy set theory (FST) and the bootstrap method has grown significantly during the last decade, for example, fuzzy regression (least square) based on bootstrap [79], and fuzzy random variables with bootstrap used in decision making and neural network [80].

In the inference layer, we used two types of very popular machine learning algorithms, Naive Bayesian (NB) and Extreme Gradient Boost (XGBoost). We examined three types of outcomes: binary outcome for classification, and continuous outcome for actual score, and rank score for percentile. NB is selected due to its easy interpretation and by survey item level prediction results. The explicit format of NB for classification based on probability and 0-1 loss function is PAC learnable. In the real example, we use sensitivity and specificity to evaluate the prediction. We may estimate the required sample size based on this estimation because the fuzzy learning process is PAC learnable. XGBoost is the most popular machine algorithm recently because of its effectiveness and accurateness. The differentiable loss function is required for the Taylor expansion at 2nd order. We use squared-error loss for this function, where the 2nd derivative is a constant. The two algorithms (XGboost and Naive Bayesian) are used to implement the inference layer. In this part, the algorithm is developed, and parameters are trained. The other algorithms can be applied in fuzzy system in prediction and classification too.

The independence of observations has been a curse for most statistical models. This assumption is hard to verify and impossible to avoid for most of the models. In Item Response Theory (IRT), the assumption is conditional independence, i.e. given the latent variable, the survey responses are not related to one another. The IRT models and Naive

Bayesian methods share the same conditional independence assumption.

In the final stage before output, we separated the predictions into four type of values. Based on the closeness to the original sample, a different weighted function is added to the related prediction components. Type I used most of the available information to predict. Type IV used up to the most available resources due to the distance from the original sample. The output values are evaluated using sensitivity and specificity for binary outcomes with 0-1 loss function and using mean square error and correlation for continuous variables and rank percentiles.

A recent paper showed improvement in PAC learning into a new stage by releasing the criteria of the difference with a probability, called Predictive-PAC (P-PAC) learning [81], where the expectation is restricted on a σ–field of invariant events. The manually introduced noise in the fuzzy system may also improve some other algorithms that could be P-PAC. In this model, there is triple uncertainty in the estimation.

In the entire dissertation, the fundamental assumption is that the sample observed distribution $\mathcal{D}$ is not too different from the population distribution. However, this assumption may not hold when working with convenience samples. Due to factors such as non-response, diverse locations of clinics, differing availability of the patients, any sample is apt to have sample unavoidable biases. In this case, during splitting the samples for training set, we can introduce the stratified sample, instead of using a simple random sample. In this way, each training sample may represent the distribution of the population with respect to the stratified parameter, e.g. age, race and ethnicity, or access to care. Eventually, the sample could be expanded to better represent the population and hence improve the performance of the derived algorithms or models.

The toolkit using short forms and algorithms to predict clinical outcomes in Figure 4.1 could be used in oral health surveillance. While the screening by short forms can never fully replicate the examination by dental professionals, it can provide a cost-effective way to conduct oral health screening for large populations of children and adolescents. It can be used in practices at local, state and even national level for tracking large child and adolescent

populations with oral health needs and setting the priorities within populations with varying urgency for dental and oral health care. This is especially true with school-aged children and adolescents who may need immediate care, and therefore contribute to gaining more timely access to needed care. The algorithms derived from PAC learnable fuzzy systems offer potential to serve this purpose.

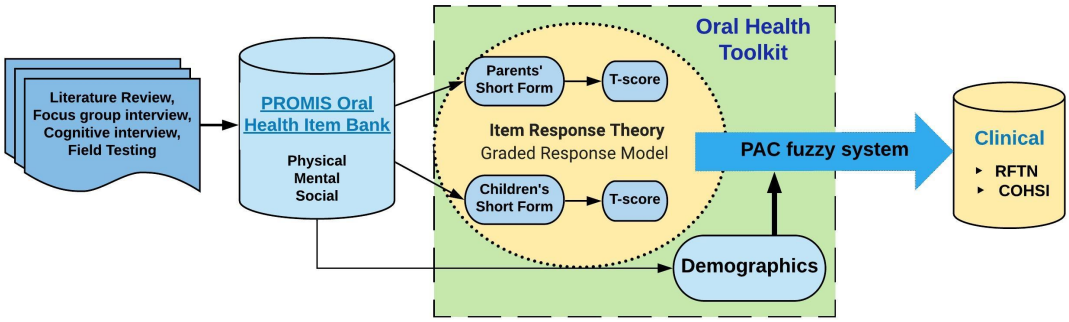

Figure 4.1: Process of developing oral health toolkit

REFERENCES

[1] G. J. Huba and P. M. Bentler, "On the usefulness of latent variable causal modeling in testing theories of naturally occurring events (including adolescent drug use): a rejoinder to martin.," 1982.

[2] R. D. Hays, K. F. Widaman, M. R. DiMatteo, and A. W. Stacy, "Structural-equation models of current drug use: Are appropriate models so simple (x)?," *Journal of Personality and Social Psychology*, vol. 52, no. 1, p. 134, 1987.

[3] H. M. Hersh and A. Caramazza, "A fuzzy set approach to modifiers and vagueness in natural language.," *Journal of Experimental Psychology: General*, vol. 105, no. 3, p. 254, 1976.

[4] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoningi," *Information sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[5] C. I. Mosier, "A psychometric study of meaning," *The journal of social psychology*, vol. 13, no. 1, pp. 123–140, 1941.

[6] M. Rose, J. B. Bjorner, B. Gandek, B. Bruce, J. F. Fries, and J. E. Ware Jr, "The promis physical function item bank was calibrated to a standardized metric and shown to improve measurement efficiency," *Journal of clinical epidemiology*, vol. 67, no. 5, pp. 516–526, 2014.

[7] S. H. Paz, K. L. Spritzer, S. P. Reise, and R. D. Hays, "Differential item functioning of the patient-reported outcomes information system (promis®) pain interference item bank by language (spanish versus english)," *Quality of Life Research*, vol. 26, no. 6, pp. 1451–1462, 2017.

[8] J. C. Cappelleri, J. J. Lundy, and R. D. Hays, "Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures," *Clinical therapeutics*, vol. 36, no. 5, pp. 648–662, 2014.

[9] M. L. Puri, D. A. Ralescu, and L. Zadeh, "Fuzzy random variables," in *Readings in Fuzzy Sets for Intelligent Systems*, pp. 265–271, Elsevier, 1993.

[10] A. F. Shapiro, "Fuzzy random variables," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 307–314, 2009.

[11] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[12] E. Hüllermeier, "Fuzzy methods in machine learning and data mining: Status and prospects," *Fuzzy sets and Systems*, vol. 156, no. 3, pp. 387–406, 2005.

[13] E. Hüllermeier, "Fuzzy sets in machine learning and data mining," *Applied Soft Computing*, vol. 11, no. 2, pp. 1493–1505, 2011.

[14] D. Dubois and H. Prade, "Articles written on the occasion of the 50th anniversary of fuzzy set theory," *FUZZY LOGIC IN ITS 50TH YEAR*, 2015.

[15] H.-J. Zimmermann, "Fuzzy set theory," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 317–332, 2010.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[17] H. Liu, R. Hays, Y. Wang, M. Marcus, C. Maida, J. Shen, D. Xiong, S. Lee, V. Spolsky, I. Coulter, *et al.*, "Short form development for oral health patient-reported outcome evaluation in children and adolescents," *Quality of Life Research*, vol. 27, no. 6, pp. 1599–1611, 2018.

[18] Y. Wang, R. Hays, M. Marcus, C. Maida, J. Shen, D. Xiong, S. Lee, V. Spolsky, I. Coulter, J. Crall, and H. Liu, "Development of a parents short form survey of their children's oral health," *International Journal of Paediatric Dentistry*, vol. 29, no. 3, pp. 332–344, 2019.

[19] Y. Wang, R. Hays, M. Marcus, C. Maida, J. Shen, D. Xiong, S. Lee, V. Spolsky, I. Coulter, J. Crall, *et al.*, "Developing childrens oral health assessment toolkits using machine learning algorithm," *JDR Clinical & Translational Research*, Under Revision.

[20] W. Labov, "The boundaries of words and their meanings," *New ways of analyzing variation in English*, 1973.

[21] G. Lakoff, "Hedges: A study in meaning criteria and the logic of fuzzy concepts," *Journal of Philosophical Logic*, pp. 458–508, 1973.

[22] L. V. Jones and L. L. Thurstone, "The psychophysics of semantics: an experimental investigation.," *Journal of Applied Psychology*, vol. 39, no. 1, p. 31, 1955.

[23] E. Portmann and A. Meier, "A fuzzy grassroots ontology for improving weblog extraction.," *JDIM*, vol. 8, no. 4, pp. 276–284, 2010.

[24] L. A. Zadeh, "Fuzzy logic and approximate reasoning," *Synthese*, vol. 30, no. 3-4, pp. 407–428, 1975.

[25] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 01 1979.

[26] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*, pp. 569–593, Springer, 1992.

[27] M. R. Chernick, *Bootstrap methods: A guide for practitioners and researchers*, vol. 619. John Wiley & Sons, 2011.

[28] P. C. Austin and J. V. Tu, "Bootstrap methods for developing predictive models," *The American Statistician*, vol. 58, no. 2, pp. 131–137, 2004.

[29] P. Lahiri, "On the impact of bootstrap in survey sampling and small-area estimation," *Statistical Science*, pp. 199–210, 2003.

[30] M. R. Chernick, "Bootstrap methods: A guide for practitioners and researchers. hoboken," 2008.

[31] A. L. Koch, J. A. Gershen, and M. Marcus, "A children's oral health status index based on dentists' judgment.," *Journal of the American Dental Association (1939)*, vol. 110, no. 1, pp. 36–42, 1985.

[32] J. Machen, P. Hagan, and R. Mercer, "Using children's oral health status index (cohsi) for longitudinal assessments," *J Dent Res*, vol. 64, no. Spec Issue, p. 292, 1985.

[33] P. Hagan, S. Levy, and J. Machen, "Validation of the children's oral health status index (cohsi).," *ASDC journal of dentistry for children*, vol. 54, no. 2, pp. 110–113, 1987.

[34] J. D. Bader, R. C. Graves, J. A. Disney, H. M. Bohannan, J. W. Stamm, J. R. Abernathy, and R. L. Lindahl, "Identifying children who will experience high caries increments," *Community dentistry and oral epidemiology*, vol. 14, no. 4, pp. 198–201, 1986.

[35] A. J. Nowak and P. S. Casamassimo, "The dental home: a primary care oral health concept," *The Journal of the American Dental Association*, vol. 133, no. 1, pp. 93–98, 2002.

[36] M. Marcus, C. Maida, Y. Wang, D. Xiong, R. Hays, I. Coulter, S. Lee, V. Spolsky, J. Shen, J. Crall, *et al.*, "Child and parent demographic characteristics and oral health perceptions associated with clinically measured oral health," *JDR Clinical & Translational Research*, vol. 3, no. 3, pp. 302–313, 2018.

[37] L. A. Zadeh, "Fuzzy logic and its applications," *New York, NY, USA*, 1965.

[38] R. J. Kuo, C. Chen, and Y. Hwang, "An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network," *Fuzzy sets and systems*, vol. 118, no. 1, pp. 21–45, 2001.

[39] L. A. Zadeh, "Probability measures of fuzzy events," *Journal of mathematical analysis and applications*, vol. 23, no. 2, pp. 421–427, 1968.

[40] M. Mizumoto and K. Tanaka, "Fuzzy sets and their operations," *Information and Control*, vol. 48, no. 1, pp. 30–48, 1981.

[41] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, pp. 54–75, 1986.

[42] H. Kwakernaak, "Fuzzy random variables–I. definitions and theorems," *Information sciences*, vol. 15, no. 1, pp. 1–29, 1978.

[43] H. Kwakernaak, "Fuzzy random variables–II. algorithms and examples for the discrete case," *Information Sciences*, vol. 17, no. 3, pp. 253–278, 1979.

[44] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive Bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[45] R. Kohavi, "Scaling up the accuracy of naive–Bayes classifiers: A decision-tree hybrid.," in *Kdd*, vol. 96, pp. 202–207, Citeseer, 1996.

[46] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers," in *Icml*, vol. 1, pp. 609–616, Citeseer, 2001.

[47] K. J. D'souza and Z. Ansari, "Big data science in building medical data classifier using naïve bayes model," in *2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 76–80, IEEE, 2018.

[48] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.

[49] H. Das, B. Naik, and H. Behera, "Classification of diabetes mellitus disease (dmd): a data mining (dm) approach," in *Progress in Computing, Analytics and Networking*, pp. 539–549, Springer, 2018.

[50] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., 1995.

[51] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, pp. 41–46, 2001.

[52] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[53] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, ACM, 2016.

[54] R. A. Taylor, C. L. Moore, K.-H. Cheung, and C. Brandt, "Predicting urinary tract infections in the emergency department with machine learning," *PloS one*, vol. 13, no. 3, p. e0194085, 2018.

[55] A. Tahmassebi, A. Gandomi, I. McCann, M. Schulte, A. Goudriaan, and A. Meyer-Baese, "Deep learning in medical imaging: fmri big data analysis via convolutional neural networks," *Proc. Pract. Exp. Adv. Res. Comput. ACM*, 2018.

[56] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001.

[57] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[58] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, pp. 1473–1480, 2006.

[59] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital signal processing*, vol. 17, no. 4, pp. 694–701, 2007.

[60] W. Van Leekwijck and E. E. Kerre, "Defuzzification: criteria and classification," *Fuzzy sets and systems*, vol. 108, no. 2, pp. 159–178, 1999.

[61] L. G. Valiant, "A theory of the learnable," in *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pp. 436–445, ACM, 1984.

[62] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*, pp. 11–30, Springer, 2015.

[63] L. Valiant, *Probably Approximately Correct: NatureÕs Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ), 2013.

[64] D. Haussler, *Probably approximately correct learning*. University of California, Santa Cruz, Computer Research Laboratory, 1990.

[65] H. Liu, R. D. Hays, M. Marcus, I. Coulter, C. Maida, F. Ramos-Gomez, J. Shen, Y. Wang, V. Spolsky, S. Lee, *et al.*, "Patient-reported oral health outcome measurement for children and adolescents," *BMC oral health*, vol. 16, no. 1, p. 95, 2016.

[66] C. A. Maida, M. Marcus, R. D. Hays, I. D. Coulter, F. Ramos-Gomez, S. Y. Lee, P. S. McClory, L. V. Van, Y. Wang, J. Shen, *et al.*, "Child and adolescent perceptions of oral health over the life course," *Quality of Life Research*, vol. 24, no. 11, pp. 2739–2751, 2015.

[67] C. A. Maida, M. Marcus, R. D. Hays, I. D. Coulter, F. Ramos-Gomez, S. Y. Lee, P. S. McClory, L. V. Van, Y. Wang, J. Shen, *et al.*, "Qualitative methods in the development of a parent survey of childrens oral health status," *Journal of patient-reported outcomes*, vol. 2, no. 1, p. 7, 2018.

[68] G. Chen, P. Faris, B. Hemmelgarn, R. L. Walker, and H. Quan, "Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa," *BMC medical research methodology*, vol. 9, no. 1, p. 5, 2009.

[69] S. P. Reise and J. Yu, "Parameter recovery in the graded response model using multilog," *Journal of educational Measurement*, vol. 27, no. 2, pp. 133–144, 1990.

[70] S. Jiang, C. Wang, and D. J. Weiss, "Sample size requirements for estimation of item parameters in the multidimensional graded response model," *Frontiers in psychology*, vol. 7, p. 109, 2016.

[71] S. E. Embretson and S. P. Reise, *Item response theory*. Psychology Press, 2013.

[72] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M. F. Leisch, "Package e1071," *R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html*, 2009.

[73] W. Kim, K. S. Kim, and R. W. Park, "Nomogram of naive bayesian model for recurrence prediction of breast cancer," *Healthcare informatics research*, vol. 22, no. 2, pp. 89–94, 2016.

[74] M. Možina, J. Demšar, M. Kattan, and B. Zupan, "Nomograms for visualization of naive bayesian classifier," in *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 337–348, Springer, 2004.

[75] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[76] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "xgboost: Extreme gradient boosting. r package version 0.6. 4.1," 2018.

[77] J. N. Rao and C. Wu, "Resampling inference with complex survey data," *Journal of the american statistical association*, vol. 83, no. 401, pp. 231–241, 1988.

[78] R. R. Sitter, "Comparing three bootstrap methods for survey data," *Canadian Journal of Statistics*, vol. 20, no. 2, pp. 135–154, 1992.

[79] W.-J. Lee, H. Y. Jung, J. H. Yoon, and S. H. Choi, "The statistical inferences of fuzzy regression based on bootstrap techniques," *Soft Computing*, vol. 19, no. 4, pp. 883–890, 2015.

[80] Y. Wang, L. Tian, and Z. Chen, "A reputation bootstrapping model for e-commerce based on fuzzy dematel method and neural network," *IEEE Access*, vol. 7, pp. 52266–52276, 2019.

[81] C. Shalizi and A. Kontorovich, "Predictive pac learning and process decompositions," in *Advances in neural information processing systems*, pp. 1619–1627, 2013.