Missing Data Imputation Based on Structural Equation Modeling Enhanced with Self-Attention

Ou Deng*, Qun Jin†

***Abstract*—Addressing missing data in complex datasets like Electronic Health Records (EHR) is critical for ensuring accurate analysis and decision-making in healthcare. This paper proposes Structural Equation Modeling (SEM) enhanced with the Self-Attention method (SESA), an innovative approach for data imputation in EHR. SESA innovates beyond traditional SEM-based methods by incorporating self-attention mechanisms, enhancing the model's adaptability and accuracy across diverse EHR datasets. This enhancement allows SESA to dynamically adjust and optimize imputation processes, overcoming the limitations of static SEM frameworks. Our experimental analyses demonstrate that SESA achieves robust predictive performance, effectively handling missing data in EHR. Moreover, SESA's architecture not only rectifies potential mis-specifications in SEM but also synergizes with causal discovery algorithms, to refine its imputation logic based on underlying data structures. These features highlight SESA's advanced capabilities and its potential for broader application in EHR data analysis and beyond, marking a significant leap forward in the field of data imputation.**

***Index Terms*—SEM, Attention, Deep learning, Missing data, EHR**

I. Introduction

In the contemporary landscape of healthcare research, Electronic Health Records (EHR) serve as a cornerstone, encapsulating a wealth of data pivotal for clinical decision-making, epidemiological studies, and personalized medicine. However, the utility of EHR is often compromised by the pervasive issue of missing data, which can skew analyses, bias results, and ultimately lead to suboptimal patient outcomes. Addressing this challenge necessitates sophisticated imputation methodologies that not only fill the missing gaps but also preserve the inherent data structure and relationships [1].

SESA transcends traditional imputation SEM-based techniques enhanced with the Self-Attention mechanism, a concept inspired by the realm of deep learning, particularly the transformer models renowned for their efficacy in capturing long-term data dependencies. This innovative enhancement allows SESA to capture and utilize the latent structural and relational dynamics within EHR data, facilitating a more nuanced and context-aware imputation process.

SESA operates on the premise that data missingness in EHR is rarely random but often structurally patterned, influenced by underlying health conditions, healthcare processes, and systematic data collection strategies. By employing SEM, SESA meticulously maps out these latent structures, enabling a hypothesis-driven approach to understanding the interdependencies among various health variables. The Self-Attention mechanism further enhances this structure by dynamically weighing the importance of different variables, allowing SESA to adaptively focus on the most relevant factors for each imputation task.

Our empirical analysis of SESA across various datasets and missingness scenarios reveals its superior performance compared to established major imputation methods. Notably, SESA consistently achieves reliable performance, signifying its ability to produce accurate and coherent imputations that resonate with the underlying data fabric.

Furthermore, the application of SESA in diverse EHR datasets, ranging from general health parameters to specific clinical indicators, underscores its versatility and robustness. The method demonstrates reliable efficacy in deciphering complex health data patterns, offering significant improvements in data quality and analytical reliability. These attributes make SESA a compelling choice for researchers and practitioners seeking to mitigate the impacts of missing data in EHR, thereby enhancing the integrity and effectiveness of health data analysis.

The source code for these experiments is publicly available at https://github.com/oudeng/SESA/.

II. Related Work

In the analysis of EHR, addressing missing data remains a longstanding challenge. Traditional methods such as mean imputation, median imputation, or multiple imputation, while effective in certain scenarios, often overlook the complex interrelationships among data, potentially leading to biases in analytical outcomes. In recent years, with the evolution of statistical and machine learning techniques, researchers have begun exploring more advanced imputation methods.

Statistical methods were among the first developed for imputation. Rubin [2] introduced the concepts of Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR), providing a theoretical foundation for understanding the nature and impact of missing data. Fuller [3] discussed multiple imputation methods, estimating parameter uncertainty by creating multiple complete datasets. These methods perform well with MAR data but may be limited when facing complex nonlinear relationships and high-dimensional data.

The approach of SEM has also been extensively developed. Hoyle [4] highlighted the application of SEM in social science research, particularly in understanding the latent structures

*Graduate School of Human Sciences, Waseda University. Email: dengou@toki.waseda.jp

†Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University. Email: jin@waseda.jp

among variables. In recent years, SEM has increasingly been applied in medical research, especially in EHR data analysis, where it helps researchers incorporate medical knowledge and theoretical constructs into their analysis.

Enders and Bandalos [5] investigated the potential of Full Information Maximum Likelihood (FIML) by focusing on its applications to SEM, a complex statistical method used in various scientific disciplines. Through rigorous empirical studies, they demonstrated the unbiased nature of FIML and its clear advantages over other traditional methods such as pairwise and listwise deletions, which often suffer from various limitations including the loss of valuable data and the potential for introducing bias.

Furthermore, Schafer and Graham [6] explored FIML by offering exhaustive reviews that directly compared FIML with other prevalent imputation methods such as multiple and regression imputation. Their analytical insights served to solidify FIML's reputation as the method of choice for dealing with missing data, particularly to preserve the integrity and reliability of datasets.

Subsequent studies by Lazar, Fielding et al., and Graham [7]–[9] delved into the nuances of applying FIML across various data types, from longitudinal studies to complex multivariate analyses. Thus, the body of enriched knowledge surrounding FIML showcases its versatility and adaptability in addressing a wide range of research questions and data challenges.

Modern discourse on missing data has undergone a significant transformation by incorporating a wide array of methodologies that extend well beyond the confines of traditional statistical techniques. Pioneering researchers such as Baraldi and Enders, White et al., Dong and Peng, Mazza, and Enders and Nelson [10]–[16] played a crucial evolutionary role by contributing seminal works that expanded the horizons of missing data analyses. Conversely, the enduring relevance of FIML as a robust and reliable method for missing data imputation was reinforced. However, they ventured into uncharted territories by introducing innovative hybrid models, which ingeniously combined the strengths of FIML with other advanced techniques such as multiple imputation and Bayesian methods to offer more versatile and comprehensive solutions.

Vaswani et al. [17] proposed the transformer model, which effectively captures long-range dependencies through self-attention mechanisms. This mechanism has achieved significant success in fields such as natural language processing and image recognition and is gradually being applied to the analysis of time series and medical data. Recently, Chen et al. [18] innovatively proposed the TriA-BioRE, a novel triangular attention framework, aimed at enhancing the performance of biomedical relation extraction. In the realm of EHRs, self-attention mechanisms have seen noteworthy applications.

Ensemble methods improve imputation accuracy by combining the strengths of various models. For instance, Baraldi and Enders [10], [15] explored the integration of multiple imputation and SEM, as well as how ensemble methods can enhance the flexibility and robustness of missing data imputation. Causal inference plays a crucial role in understanding relationships among variables. Pearl (2000) introduced causal graph models, which help researchers identify potential causal relationships among variables. In EHR data analysis, causal inference can aid in constructing more accurate imputation models, as it accounts for the causal structure among variables.

In addressing the issue of missing values in EHR data, the SESA method proposes an innovative approach that extends the Structural Equation Modeling (SEM) and incorporates the self-attention mechanism. This method capitalizes on the strengths of SEM in modeling complex inter-variable relationships, while dynamically adjusting and optimizing the imputation process through the self-attention mechanism to accommodate various EHR datasets. The introduction of this method is anticipated to propel research advancements in handling medical data with complex structures and relationships.

III. Methodology

The methodology proposed in this study is predicated on an imputation strategy that leverages a deep understanding of the interrelations among data variables, aiming to augment the predictive accuracy for missing data within the original dataset through intricate modeling of the interactions between data variables [7]. SEM offers a precise mathematical framework to encapsulate these relationships, enabling the incorporation of prior medical knowledge and theoretical constructs into our analysis. This initial step is crucial, especially in the realm of EHR research, for it aids in structuring the variable relationships and enhances the interpretability and verification of the imputation results.

Building upon the initial SEM framework, our research employs the FIML method for the preliminary estimation of missing values. FIML, by modeling the joint distribution of observed data, facilitates the derivation of maximum likelihood estimates for model parameters, thus providing an initial imputation grounded in the data's overall distribution. It is important to note that while this step prepares the data for the subsequent self-attention model training, it does not directly involve the self-attention mechanism.

A. Structural Equation Modeling (SEM)

SEM serves as the foundational bedrock of our imputation methodology, offering a robust mathematical framework to articulate and analyze the intricate relationships among observed and latent variables within datasets. SEM excels in its ability to simultaneously model multiple interrelated dependencies, thereby facilitating a comprehensive understanding of the data structure [19]. This modeling approach is particularly adept at encapsulating the complex interplay of variables prevalent in EHR, making it an indispensable initial step in our imputation process. By leveraging SEM, we can incorporate prior medical knowledge and theoretical constructs into our analysis, providing a structured representation of variable relationships that guide subsequent imputation efforts.

B. Full Information Maximum Likelihood (FIML)

FIML is an estimation technique used in SEM to handle missing data [20]. It estimates model parameters by maximizing the likelihood function based on the observed data. The

likelihood function for a dataset with missing values can be expressed as:

$$\ln L(\boldsymbol{\theta}; \boldsymbol{X}_{\text{obs}}) = \ln \sum_{\boldsymbol{X}_{\text{mis}}} P(\boldsymbol{X}|\boldsymbol{\theta}) \tag{1}$$

Here, $\boldsymbol{X}_{\text{obs}}$ denotes the observed data, $\boldsymbol{X}_{\text{mis}}$ the missing data, and $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$ the complete data. FIML aims to estimate the parameter $\boldsymbol{\theta}$ by modeling the joint distribution $P(\boldsymbol{X}|\boldsymbol{\theta})$, thereby obtaining the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. This process involves integrating over the missing data $\boldsymbol{X}_{\text{mis}}$ to compute the expected log-likelihood, which is then maximized to find the optimal parameter estimates.

C. Self-Attention Mechanism

The self-attention mechanism [17] is a pivotal component in modern neural network architectures, particularly in the context of sequence modeling and analysis. It dynamically weighs the importance of different parts of the input data, enabling the model to focus on relevant features for the task at hand.

Given an initial input $\boldsymbol{Y} = \boldsymbol{X}_{\text{FIML}}$, where $\boldsymbol{X}_{\text{FIML}}$ represents the data estimated using the FIML estimation, the self-attention mechanism operates on this input to compute attention scores. The mechanism can be mathematically represented as:

$$\boldsymbol{Y} = \text{SelfAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

Here, $\boldsymbol{Q}$ (queries), $\boldsymbol{K}$ (keys), and $\boldsymbol{V}$ (values) are matrices derived from the initial input $\boldsymbol{Y}$ through separate linear transformations. These matrices facilitate the computation of attention scores, which determine the degree of focus or attention that the model should allocate to various parts of the input data. The dimension d_k represents the scaling factor, typically the dimensionality of the keys and queries, which helps stabilize the gradients during training.

The softmax function applied to the product of queries and keys, normalized by $\sqrt{d_k}$, ensures that the attention weights across the data sum to 1. This normalization allows the model to distribute its focus selectively, emphasizing more relevant information while diminishing the less relevant parts.

The self-attention mechanism's ability to process and relate different positions of the input sequence makes it exceptionally suitable for handling data with complex interdependencies, such as time-series data, sentences, or any sequential information. This mechanism underlies the success of transformer models in various applications, from natural language processing to image recognition, by enabling a nuanced understanding and representation of the input data.

D. Loss Function of Self-Attention

The self-attention neural network employs a composite loss function to guide the training process, ensuring the model's predictions are both accurate and reflective of the underlying data structure. The loss function is formulated as follows:

$$\text{Total\_loss} = \alpha \cdot \text{MSE\_loss} + \beta \cdot \text{cov\_loss} + \gamma \cdot \text{L1\_loss} \tag{3}$$

where α, β, and γ are hyperparameters that balance the contribution of each component to the total loss. This structured approach to loss calculation ensures the model's predictions are not only close to the true values but also respect the covariance structure of the data, and encourage parameter sparsity to prevent overfitting.

The MSE_loss guides the iterative refinement of the model's predictions, denoted as $\hat{y}$, to closely approximate the baseline values, y, across all observations. Through successive iterations, $\hat{y}$ evolves to minimize the discrepancy with y, acknowledging that y represents the underlying but unknowable true values:

$$\text{MSE\_loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

The cov_loss component aligns the covariance matrix of the model's iteratively refined predictions, $\hat{Y}$, with the baseline covariance matrix, representing an approximation of the unknowable true dataset, Y. This alignment ensures that the model's predictions not only achieve accuracy at the individual level but also reflect the underlying distribution's overall variability and the intricate relationships among features. Essentially, it enhances the model's capacity to grasp and replicate the inherent structure of the data:

$$\text{cov\_loss} = \left\| \text{Cov}(\hat{Y}) - \text{Cov}(Y) \right\|_F \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm[1], quantifying the degree of disparity between the two covariance matrices.

The L1_loss component imposes sparsity on the model parameters, $\boldsymbol{w}$, reducing the risk of overfitting:

$$\text{L1\_loss} = \lambda \sum_{i=1}^{n} |w_i| \tag{6}$$

where λ is a regularization parameter.

This loss function facilitates the self-attention model's learning by not only minimizing prediction error but also ensuring that the predictions adhere to the data's structural properties and maintain model simplicity.

E. Algorithmic Implementation

The SESA mechanism is algorithmically implemented as Algorithm 1, integrating SEM, FIML, and the self-attention mechanism into a cohesive process for missing data imputation. The algorithm begins with SEM to model variable relationships, followed by FIML for initial missing value estimation and concludes with the self-attention mechanism for data refinement.

[1]The Frobenius norm is defined as the square root of the sum of the squares of all elements in a matrix. It serves as a metric to quantify the magnitude or complexity of a matrix, commonly employed in assessing the discrepancies between matrices and analyzing errors within optimization problems.

Algorithm 1 SESA Mechanism

Require: Observed dataset X with missing values.
Ensure: Refined imputed values for X_{mis}.

Step 1: SEM Initialization
Utilize SEM to model the relationships among variables based on prior medical knowledge.
Step 2: FIML Estimation
Initialize model parameters $\boldsymbol{\theta}$ based on SEM output.
Employ FIML to estimate missing values, generating preliminary imputed data X_{FIML}.
Step 3: Self-Attention Refinement
Set self-attention model parameters $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$.
for each iteration until convergence **do**
    Apply self-attention to X_{FIML}, refining imputation based on learned data patterns.
    Update model parameters to minimize the loss function:

$$\mathcal{L} = \alpha \cdot \text{MSE\_loss}(Y, \hat{Y}) + \beta \cdot \text{cov\_loss}(\hat{Y}, Y) + \gamma \cdot ||\boldsymbol{W}||_1$$

end for
Output: Refined imputed dataset Y_{imputed}.

F. Summary of SESA

The SESA methodology represents a novel approach to addressing the challenge of missing data imputation, particularly within the context of EHR. By integrating the statistical rigor of SEM and FIML with the dynamic adaptability of the self-attention mechanism, SESA offers a comprehensive solution that not only enhances imputation accuracy but also respects the inherent structure of the data. Key components of the SESA framework include:

- **SEM** provides a structured representation of the relationships among observed and latent variables, leveraging prior medical knowledge to guide the imputation process.
- **FIML** offers an initial estimation of missing values by maximizing the likelihood function based on the observed data, utilizing all available information without bias.
- **Self-Attention Mechanism** refines the initial imputations by dynamically focusing on the most relevant parts of the data, capturing long-distance dependencies and complex patterns within EHR.

The sequential integration of these components ensures that SESA not only provides a statistically sound initial estimate for missing values but also leverages deep learning techniques to refine these estimates, resulting in imputations that are both accurate and reflective of the data's complex relationships. This innovative combination addresses the limitations of traditional imputation methods, offering a robust and efficient solution for the complex datasets characteristic of EHR.

In essence, SESA embodies the synergy between statistical modeling and machine learning, paving the way for advanced imputation techniques that can adapt to the unique challenges presented by healthcare data. The framework's ability to incorporate domain-specific knowledge through SEM and dynamically adjust to the data's structure through self-attention makes it a powerful tool for researchers and practitioners alike, seeking to mitigate the impact of missing data in their analyses.

IV. Experiment

A. Experiment Design

Data source and Experimental dataset

Given the authoritative nature and quality of EHR data sources, we opted for the renowned public dataset, "Indicators of Heart Disease (CDC2022)[2]." This dataset originates from the Centers for Disease Control and Prevention (CDC) and constitutes a significant portion of the Behavioral Risk Factor Surveillance System (BRFSS)[3], which conducts annual telephone surveys to gather data on the health status of over 400,000 U.S. residents.

Like most EHR datasets, the CDC2022 dataset comprises both numerical and categorical variables, with the latter being more prevalent. We extracted a subset of this extensive dataset that contained no missing data, encompassing over 246,000 entities, to serve as our foundational dataset. This dataset size is sufficiently large to meet our experimental needs.

From this foundational dataset, we randomly selected 300/1000/3000 entities to form our experimental dataset. Given the common practice in sociomedical and EHR-related data analysis of grouping data by gender, age, or specific characteristics, this dataset size is typical for most related research and is suitable for most statistical analysis and testing methods to yield valuable conclusions. This scale is also appropriate for testing our methods.

Experimental methodology

We treat this dataset without missing values as the "Ground truth" and then artificially remove a certain percentage of data randomly. The experiments apply both baseline imputation methods and our method to impute missing data in the dataset, subsequently evaluating the accuracy of the imputed data points relative to the Ground Truth using a series of metrics.

Variable selection

We identify and select variables commonly encountered in EHR datasets, encompassing both numerical and categorical data types. The variables chosen from the CDC2022 dataset include 'AgeCategory', 'GeneralHealth', 'HadDiabetes', 'BMI', 'SmokerStatus', and 'SleepHours'.

Figure 1 presents our experimental dataset, illustrating the characteristic features of numerical and categorical variables typical in EHR data. Numerical variables were utilized in their original form, while categorical variables underwent ordinal label encoding to facilitate analysis.

Evaluation metrics

The anticipated outcome for the imputation process is that the distribution of the imputed EHR data aligns as closely as

[2]This public dataset of CDC2022 can be accessed at https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

[3]The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The official website of BRFSS is located at https://www.cdc.gov/brfss/.

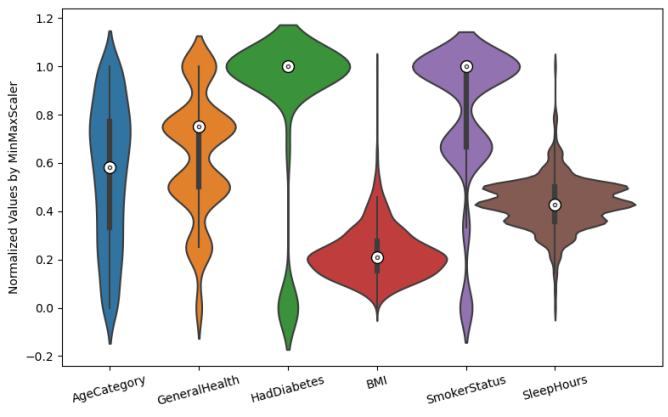

Fig. 1. Distribution of the selected variables ("Ground truth") from the CDC2022 dataset. The data normalized within the respective ranges of variables.

possible with the distribution of the actual data, ensuring minimal discrepancies in both numerical and categorical variables. Thus, we selected the following metrics:

- **Root Mean Square Error (RMSE)** represents the square root of the average squared differences between predicted and observed values. A smaller value indicates a higher precision of imputation.
- **Mean Absolute Percentage Error (MAPE%)** represents the average absolute percentage error of imputed values relative to true values. A smaller value indicates a higher accuracy of imputation.
- **Coefficient of Determination (R^2)** reflects the degree to which the imputation model accounts for the variability in the data. Theoretically, R^2 values range from 0 to 1, with values closer to 1 indicating higher model explanatory power. Note that R^2 can also be negative when the model performs poorly.
- **Wasserstein Distance (W-Distance)** measures the difference between the distributions of imputed data and the true data. A smaller value indicates a closer match to the true distribution.
- **Wilcoxon signed-rank test** measures the statistical significance between the imputed data and the true data. Given that certain variables, particularly categorical ones, may not adhere to a normal distribution, we utilize the non-parametric Wilcoxon signed-rank test. This method assesses whether data, following deletion and subsequent imputation, differ significantly from the original true data. The p-value threshold is set at 0.05 ordinarily, and the analysis includes an evaluation of the Effect Size and the 95% Confidence Interval (CI).

Experimental baseline methodologies of imputation

The experimental baselines incorporated the major imputation methodologies: Mean, Median, K-Neighbors, Random Forest, Bayesian Ridge, Gradient Boosting, Epsilon-Support Vector Regression (SVR), and Multilayer Perceptron (MLP) Regression. These methodologies represent the cornerstone of data imputation practices, duly integrated within the **Python scikit-learn**[4] **library**, which is a paramount toolkit in the data science domain and extensively leveraged across a multitude of medical research studies.

- **K-Neighbors** is a non-parametric, instance-based learning method that predicts a value based on the average of the k-nearest neighbors, capturing local data patterns which can be particularly effective for imputation when the data has a complex structure not well-represented by global trends.
- **Random Forest** builds multiple decision trees and merges them to get a more accurate and stable prediction in missing data imputation, particularly effective in capturing complex nonlinear relationships between features.
- **Bayesian Ridge** applies Bayesian statistics to ridge regression, providing the advantage of automatic regularization parameter tuning and incorporating prior information, which can lead to more reliable imputation results under uncertainty and in cases with relatively small datasets or datasets with high multicollinearity.
- **Gradient Boosting** builds sequential models to minimize loss. It imputes missing values due to its ability to handle diverse data distributions and complex nonlinear relationships, ensuring precise and stable imputation even in datasets with significant patterns of missingness. Gradient Boosting is one of the core estimator engines in the Multiple Imputation by Chained Equations (MICE) method.
- **Epsilon-Support Vector Regression (SVR)** is a support vector machine-based regression, which handles nonlinearities using kernel functions, to manage outliers and work with high-dimensional data, especially in datasets where the relationship between features is complex and not easily modeled by traditional linear methods.
- **Multilayer Perceptron (MLP)** is a neural network model capable of learning non-linear models, offering the advantage of high flexibility and the capacity to capture complex patterns in the data, making it suitable for imputing missing values in datasets where relationships between variables are nonlinear and intricate.

Experimental environment

The experimental setup employs Python 3.8 for data preprocessing utilizing Pandas, and baseline methods are implemented using the scikit-learn library. Causal analysis is conducted via the CausalNex library, operating in a CPU environment. The SESA algorithm is developed by the PyTorch framework and executed on an NVIDIA RTX A6000 GPU.

B. Experiment results

Experiment assesses various data imputation methods across different sample sizes (300, 1000, 3000) under a fixed 30% missing rate condition. Evaluation metrics include RMSE, MAPE,R^2, Wasserstein distance, and Wilcoxon Rank Test

[4]Scikit-learn is an open-source machine learning library for Python, providing a wide array of supervised and unsupervised learning algorithms, tools for model fitting, data preprocessing, model evaluation, and many other utilities. The official website of scikit-learn is located at https://scikit-learn.org/.

comprehensively reflect the overall performance and accuracy of the imputation methods.

Results from Table I indicate that the SESA method often excels in key performance indicators, paralleling or surpassing other mainstream imputation techniques. Notably, at sample sizes of 300 and 1000, SESA demonstrates robust performance across RMSE, MAPE, and R^2 metrics. At a sample size of 300, it reveals the best imputation results, with lower RMSE and MAPE values and a higher R^2, signifying its sensitivity to data variations and imputation accuracy. As the sample size expands to 1000 and 3000, SESA maintains consistent imputation performance, even though its advantage in Wasserstein distance slightly diminishes. However, considering all evaluation metrics, SESA's performance still surpasses the majority of baseline methods.

Table I also identifies the top three performing methods for each data set's imputation assessment metrics, marked with asterisks. SESA leads with six stars, followed by Gradient Boosting with five, and other machine learning methods and the Mean method each receiving four stars.

While the mean and median methods exhibit practicality in some cases, especially at a sample size of 1000 where the Mean method ranks in the top three for RMSE, MAPE, and R^2, these simple imputation methods often struggle to maintain consistency across all evaluation metrics. Notably, they exceed a 0.5 Wasserstein distance across all sample sizes, significantly higher than other methods. Wilcoxon Rank Test further corroborates these findings, indicating statistical discrepancies between these simple imputation methods and the actual data.

The remaining seven machine learning imputation methods show relatively consistent performance across the metrics, with their superiority becoming more apparent as sample size increases. At a sample size of 3000, Bayesian Ridge, SVR, and MLP slightly outperform others in RMSE, MAPE, and R^2. In terms of Wasserstein distance, K-Neighbors, Random Forest, and Gradient Boosting consistently show significantly lower values across all sample sizes, suggesting better alignment with the original data distribution.

In Table II, Wilcoxon Rank Statistic and p-value provide insight into the statistical significance of the differences between the imputed and original data. The superiority of machine learning imputation methods becomes evident with increased data volume.

Table II also employs a star-ranking system to identify the top three performing methods for each variable's imputation result, based on p-values close to 1.0 and evaluated in conjunction with the Effect Size. SESA and Bayesian Ridge Regression each secure four stars, indicating superior performance.

SESA and Bayesian Ridge show comparable imputation performance, especially notable in variables like GeneralHealth, HadDiabetes, and BMI, reflecting their robust performance across different variable types. Specifically, for BMI, the minor differences in Wilcoxon Rank Test results among these methods suggest that SESA and neural network-based MLP might have a certain advantage in imputing numerical variables.

For the SmokerStatus variable, Gradient Boosting performs best, followed by SESA and Random Forest. The larger differences in Wilcoxon Rank Test results for other methods may relate to SmokerStatus's data characteristics, indicating potentially complex associations with other variables.

In summary, SESA demonstrates certain advantages over other mainstream methods, especially in small to medium data scales. With favorable Wilcoxon Rank Test results for most variables, indicating no significant difference with the original data, SESA underscores the effectiveness and accuracy of its imputation method. It maintains high accuracy and stability, not only in small-sample conditions but also as the sample size increases, evidencing its strong generalizability and adaptability to various scales of missing data challenges.

V. Discussion

A. Fitting and Analysis of Structural Equation Modeling (SEM) Models

In the SESA imputation method, the input of the SEM plays a foundational role [19]. For EHR data, although the relationships between some health variables are relatively clear, many relationships remain unclear or require further exploration. Due to the complexity of variable relationships in EHR data, conventional correlation analysis and SEM model testing struggle to reveal the deeper connections within the data.

This study sets forth two SEM models for analysis: one is the BMI model, in which BMI is the dependent variable, determined by independent variables such as age category, sleep duration, diabetes status, smoking status, and general health status; the other is the GeneralHealth model, treating general health status as the dependent variable. Using the Python semopy library [21], these two models were fitted and analyzed to assess the relationships between the dependent variable and other variables, and the model fit was calculated. The detailed results are presented in Table III.

In the BMI model, GeneralHealth, AgeCategory, SleepHours, and HadDiabetes significantly influence BMI, with GeneralHealth and HadDiabetes exerting a larger negative impact, SleepHours having a lesser negative effect, and SmokerStatus having a positive impact on BMI. The estimates and standard errors for each predictor variable are reported, and the p-values indicate statistically significant impacts of all variables on BMI. The autocorrelation of BMI is significant, with a large estimate and a very small p-value.

In the GeneralHealth model, BMI, AgeCategory, SleepHours, HadDiabetes, and SmokerStatus significantly influence GeneralHealth, with BMI and AgeCategory negatively correlated with GeneralHealth, while SleepHours, HadDiabetes, and SmokerStatus are positively correlated. The estimates, standard errors, and statistical significance for each independent variable are precisely reported.

The optimization results are minimal for both models, indicating a good model fit. The Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA) provide excellent fit indices for both models, with CFI close to 1 and RMSEA at 0, indicating both models have a good

TABLE I
EXPERIMENT-I: IMPUTATION EVALUATION ON THE SAMPLE SIZE OF 300/1000/3000 AND FIXED MISSING RATE OF 30%.

Imputation methods	Sample size	RMSE		MAPE%		R^2		Wasserstein-Dist	
Mean	300	1.2487		11.2723		0.6945		0.5723	
	1000	1.3007	•	10.8440	•	0.6639	•	0.5337	
	3000	1.3485		10.4587		0.6861	•	0.5536	
Median	300	1.2474	•	12.2662		0.6762		0.5101	
	1000	1.3217	•	11.4416		0.6487		0.5159	
	3000	1.3686		11.0255		0.6712		0.5348	
K-Neighbors	300	1.3925		11.0388	•	0.6621		0.2581	•
	1000	1.4498		11.2692		0.6011		0.2341	•
	3000	1.4586		10.5035		0.6414		0.2068	•
Random Forest	300	1.4311		10.9684	•	0.6260		0.2242	•
	1000	1.4628		10.9906		0.5956		0.2321	•
	3000	1.4889		10.3550		0.6330		0.2272	•
Bayesian Ridge	300	1.3101		11.4200		0.6809	•	0.5087	
	1000	1.3299		11.1335		0.6504		0.4643	
	3000	1.3188	•	9.8909	•	0.7117	•	0.3766	
Gradient Boosting	300	1.4214		11.1105		0.6364		0.2344	•
	1000	1.4013		10.9240	•	0.6177		0.2718	•
	3000	1.4234		10.0576	•	0.6570		0.2503	•
Epsilon-Support Vector	300	1.2406	•	11.9456		0.6871		0.4356	
	1000	1.3113		11.1278		0.6576	•	0.4006	
	3000	1.3304	•	10.4246		0.6916	•	0.3969	
Multilayer Perceptron	300	1.6714		13.5015		0.5215		0.2606	
	1000	1.3920		10.9748		0.6453		0.2995	
	3000	1.3660	•	9.9485	•	0.6890	•	0.3144	
SESA (BMI model)	300	1.2120	•	10.9471	•	0.7172	•	0.3831	
	1000	1.2822	•	10.4313	•	0.6828	•	0.3671	
	3000	1.4085		10.6284		0.6663		0.4415	
SESA (GeneralHealth model) (for Section V only)	300	1.2543		12.3369		0.6521		0.4309	
	1000	1.3345		10.6692		0.6822		0.4695	
	3000	1.3682		10.9475		0.6611		0.4831	

Note: (1) The metrics of RMSE, MAPE%, R^2, and Wasserstein Distance represent the mean values aggregated across all assessed variables, providing a comprehensive view of the imputation performance.
(2) The bullet (•) denotes the methods that ranked within the top three for performance, considering the specified sample size, thereby highlighting the most effective imputation strategies in the given context.

fit. The p-values here are used to assess the statistical significance of the relationships between BMI (or GeneralHealth) and other variables in the model. Lower p-values (¡ 0.05) generally indicate that there is strong evidence against the null hypothesis, showing a significant association between the variables in the model. Despite the clear fit results, it remains challenging to determine which model is better for initializing SESA imputation.

B. Improving SEM Initialization Based on Causal Discovery Directed Acyclic Graph (DAG)

Following the above discussion, when the optimal initialization model cannot be determined through SEM fitting and analysis, we consider employing causal discovery methods to seek better solutions. Using the NOTEARS algorithm for DAG analysis of selected variables, we aim to gain a more rational SEM initialization input from the causal structure. The NOTEARS algorithm [22] is an optimization-based method designed to learn Directed Acyclic Graphs (DAGs) from data, ensuring the learned graph structure is acyclic.

With the same dataset and experimental setup, the results from the NOTEARS algorithm reveal potential causal relationships between variables, uncovering possible hierarchical structures and providing estimates of causal direction and strength. Particularly, the DAG analysis reinforces the rationale behind the BMI model as an initialization input for SEM.

In TablesI and II, the results of SESA (BMI model) and SESA (GeneralHealth model) are presented. The experimental findings indicate that the former, which aligns more closely with the DAG analysis, slightly outperforms the latter across almost all evaluation metrics. Two interesting observations emerge from the analysis: Firstly, in the Wilcoxon Rank Test, the latter model shows better imputation performance for GeneralHealth than the former, indirectly validating the impact of initial SEM input on the outcome of SESA. Secondly, at a large data scale (3000 entities), the discrepancy between the two different initial SEM models in terms of RMSE, MAPE, R^2, and Wasserstein Distance is significantly reduced compared to smaller and medium data scales. This phenomenon suggests the enhanced role of the Self-Attention mechanism; as the dataset size increases, Self-Attention is capable of more substantially mitigating the disparities caused by different SEM model initializations.

Comparing the GeneralHealth model as an initialization,

TABLE II
EXPERIMENT-I: SAMPLE SIZE:1000; MISSING RATE: 30%.

Variables	**Imputation methods**	**Wilcoxon Rank Statistic**	**p-value** (0.05)		**Effect Size**	**95% CI** lower	upper
AgeCategory	Mean	20992.5	0.7828	•	663.8411	8.0	9.0
	Median	20992.5	0.7828	•	663.8411	8.0	9.0
	K-Neighbors	17117.5	0.0068		541.3029	8.0	9.0
	Random Forest	16607.0	0.0144		525.1595	8.0	9.0
	Bayesian Ridge	17104.5	0.1063		540.8918	8.0	9.0
	Gradient Boosting	19090.5	0.2231		603.6946	8.0	9.0
	Epsilon-Support Vector	19910.0	0.6615	•	629.6095	8.0	9.0
	Multilayer Perceptron	19613.0	0.2969		620.2175	8.0	9.0
	SESA (BMI model)	20067.0	0.1936		634.5743	8.0	9.0
	SESA (GeneralHealth model)	12531.5	0.0281		396.2808	8.0	8.0
GeneralHealth	Mean	4631.5	< 0.0001		146.4609	4.0	4.0
	Median	4161.0	< 0.0001		131.5824	4.0	4.0
	K-Neighbors	10283.5	0.8308	•	325.1928	4.0	4.0
	Random Forest	7430.5	0.0130		234.9730	4.0	4.0
	Bayesian Ridge	9020.0	0.6376	•	285.2374	4.0	4.0
	Gradient Boosting	8600.0	0.2466		271.9560	4.0	4.0
	Epsilon-Support Vector	6935.0	0.0487		219.3040	4.0	4.0
	Multilayer Perceptron	7728.0	0.0997		244.3808	4.0	4.0
	SESA (BMI model)	6355.0	0.3877	•	200.9627	4.0	4.0
	SESA (GeneralHealth model)	7963.0	0.5013		251.8122	4.0	4.0
HadDiabetes	Mean	0.0	< 0.0001		0.0000	4.0	4.0
	Median	0.0	< 0.0001		0.0000	4.0	4.0
	K-Neighbors	1972.5	0.1390	•	62.3759	4.0	4.0
	Random Forest	3726.0	0.0269		117.8265	4.0	4.0
	Bayesian Ridge	4918.0	0.5265	•	155.5208	4.0	4.0
	Gradient Boosting	3798.0	0.0037		120.1033	4.0	4.0
	Epsilon-Support Vector	0.0	< 0.0001		0.0000	4.0	4.0
	Multilayer Perceptron	5738.5	0.0588		181.4673	4.0	4.0
	SESA (BMI model)	8163.0	0.2064	•	258.1367	4.0	4.0
	SESA (GeneralHealth model)	4085.5	0.1495		129.1949	4.0	4.0
BMI	Mean	20493.0	0.6016		648.0456	27.11	27.98
	Median	17371.5	0.0091		549.3351	27.10	27.97
	K-Neighbors	20317.0	0.5192		642.4800	27.12	27.94
	Random Forest	20388.0	0.5518		644.7252	27.10	27.98
	Bayesian Ridge	20520.0	0.6148	•	648.8994	27.11	27.98
	Gradient Boosting	19474.0	0.2182		615.8220	27.12	27.98
	Epsilon-Support Vector	17672.0	0.0129		558.8378	27.11	27.98
	Multilayer Perceptron	20918.0	0.8210	•	661.4852	27.11	27.98
	SESA (BMI model)	21746.0	0.7221	•	687.6689	27.10	27.98
	SESA (GeneralHealth model)	20335.0	0.0215		643.0492	27.21	28.13
SmokerStatus	Mean	7440.0	< 0.0001		235.2735	4.0	4.0
	Median	0.0	< 0.0001		0.0000	4.0	4.0
	K-Neighbors	5166.0	0.2096		163.3633	4.0	4.0
	Random Forest	7539.5	0.4336	•	238.4199	4.0	4.0
	Bayesian Ridge	8078.5	0.2464		255.4646	4.0	4.0
	Gradient Boosting	8084.0	0.9258	•	255.6385	4.0	4.0
	Epsilon-Support Vector	0.0	< 0.0001		0.0000	4.0	4.0
	Multilayer Perceptron	7758.0	0.3907		245.3295	4.0	4.0
	SESA (BMI model)	6921.0	0.3946	•	218.8612	4.0	4.0
	SESA (GeneralHealth model)	7024.5	0.5907		222.1342	4.0	4.0
SleepHours	Mean	8894.0	0.8044	•	281.2530	7.0	7.0
	Median	8894.0	0.8044	•	281.2530	7.0	7.0
	K-Neighbors	9906.0	0.5033		313.2552	7.0	7.0
	Random Forest	9471.0	0.3961		299.4993	7.0	7.0
	Bayesian Ridge	8356.0	0.6269	•	264.2399	7.0	7.0
	Gradient Boosting	8920.5	0.5580		282.0910	7.0	7.0
	Epsilon-Support Vector	9223.0	0.9553	•	291.6569	7.0	7.0
	Multilayer Perceptron	8614.5	0.3170		272.4144	7.0	7.0
	SESA (BMI model)	6921.0	0.3946		218.8612	7.0	7.0
	SESA (GeneralHealth model)	9251.0	0.0404		292.5423	7.0	7.0

Note: The bullet (•) symbol identifies the imputation methods ranked within the top three. This ranking implies insufficient evidence to reject the null hypothesis, suggesting no statistically significant differences between the imputation results and the original data.

TABLE III
PARAMETER ESTIMATES FOR SEM (BMI MODEL) AND SEM (GENERALHEALTH MODEL)

Dependent Variable		Independent Variable	Estimate	Std. Err	p-value
BMI	$\sim$	GeneralHealth	-1.3283	0.1150	< 0.0001
BMI	$\sim$	AgeCategory	-0.1060	0.03355	0.0016
BMI	$\sim$	SleepHours	-0.2146	0.0769	0.0052
BMI	$\sim$	HadDiabetes	-1.1532	0.1128	< 0.0001
BMI	$\sim$	SmokerStatus	0.2503	0.1266	0.0480
BMI	$\sim\sim$	BMI	38.4322	0.9923	< 0.0001
GeneralHealth	$\sim$	BMI	-0.0320	0.0028	< 0.0001
GeneralHealth	$\sim$	AgeCategory	-0.0238	0.0052	< 0.0001
GeneralHealth	$\sim$	SleepHours	0.0352	0.0119	0.0032
GeneralHealth	$\sim$	HadDiabetes	0.1784	0.0175	< 0.0001
GeneralHealth	$\sim$	SmokerStatus	0.2001	0.0193	< 0.0001
GeneralHealth	$\sim\sim$	GeneralHealth	0.9280	0.0240	< 0.0001
Results for BMI model:		Optimization result: 7.5608e-08; CFI: 1.0188; RMSEA: 0			
Results for GeneralHealth model:		Optimization result: 1.1269e-07; CFI: 1.0188; RMSEA: 0			

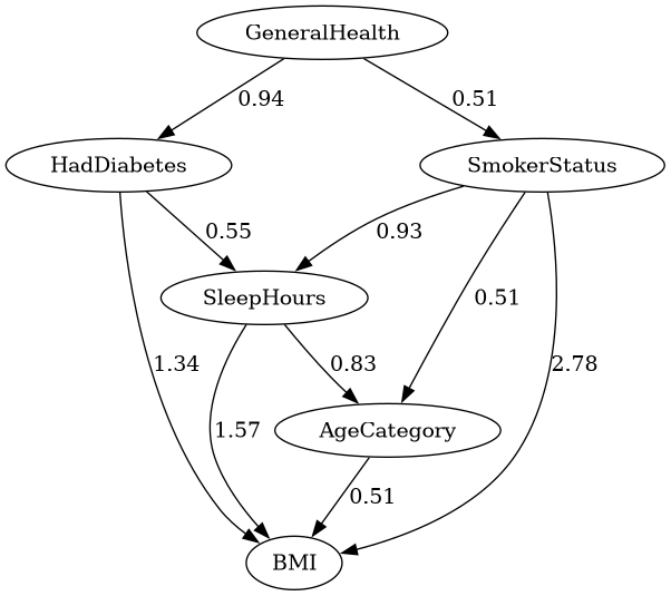

Fig. 2. Causal discovery analysis results of for the selected variables group by NOTEARS algorithm. In this context, nodes represent the variables included in the dataset, while directed edges indicate potential causal directions. The weight of each directed edge reflects the strength of the causal effect.

results indicate that the BMI model is closer to the causal relationships parsed by the NOTEARS algorithm, thus providing better SEM fitting and FIML estimation outcomes. Additionally, the SelfAttention mechanism has shown positive optimization effects on the FIML estimation results under both SEM configurations, making certain corrections even under the less reasonable SEM setup.

In summary, the statistical-based FIML estimation and the neural network-based SelfAttention mechanism demonstrate complementary advantages. FIML estimation proves robust in small data environments, providing a solid foundation for SelfAttention, which, in medium-sized data environments, leverages the learning strengths of neural networks to uncover potential deep relationships in the data, thereby optimizing imputation results.

However, it is essential to acknowledge that improvements in SEM models are not without constraints due to data heterogeneity and locality. Data heterogeneity refers to the diversity and variability among individuals or groups within a dataset, leading to a variety of characteristics and behaviors. Locality implies that data analysis or model construction may only consider specific variables or subsets, potentially overlooking broader contexts or other relevant variables. Inductive analyses of SEM and DAG are based on population-scale generalizations, unable to cover individual differences and all potential variable relationships comprehensively. Therefore, selecting appropriate variable groups for data imputation is crucial, based on prior knowledge and research objectives, to aptly reflect the problem structure while minimizing bias introduction.

VI. CONCLUSION

The findings from this research underscore the effectiveness of the SESA method in addressing the perennial challenge of missing data in EHR. Through the innovative integration of SEM and Self-Attention mechanisms, SESA not only outperforms traditional imputation methods but also offers a dynamic, adaptive approach to understanding and reconstructing the intricate patterns of health data.

The empirical validation of SESA across multiple datasets and missing scenarios establishes its efficacy, as evidenced by its superior performance metrics, including RMSE, MAPE, and R-squared values. These results highlight SESA's capability to produce accurate, reliable imputations that are in harmony with the intrinsic data structure, thereby enhancing the quality and utility of EHR for clinical and research purposes.

Moreover, the discussions revolving around the SESA method illuminate its theoretical and practical implications, particularly in the realm of healthcare data analysis. The method's ability to integrate causal inference through DAG analysis provides a deeper insight into the underlying variable relationships, enriching the SEM initialization process. This aspect of SESA not only refines the imputation outcomes but also contributes to a more nuanced understanding of the health data ecosystem.

In conclusion, the SESA method represents a significant advancement in the field of data imputation, particularly within

the context of EHR. Its ability to accurately capture and reflect the complex, multifaceted nature of health data makes it a valuable tool for researchers and practitioners alike. Future endeavors will aim to further refine SESA, exploring its applicability and scalability across various domains and data types, to fully harness its potential in improving data completeness and reliability in EHR and beyond.

VII. Limitation and Future work

The SESA method, though effective in EHR data imputation, encounters limitations related to its linear SEM assumption, which may not align with the complex, nonlinear relationships in healthcare data. Addressing this requires exploring nonlinear SEM models or integrating nonlinear dynamics within SESA to better capture the data's intricacies. Overfitting poses a challenge, especially in high-dimensional EHR datasets. The method's Self-Attention component, while adept at identifying complex patterns, can overfit the training data, reducing generalizability. Future versions of SESA should incorporate regularization or advanced validation techniques to prevent overfitting and improve model robustness.

Data quality significantly influences SESA's performance. Inaccurate or biased input data can skew imputation outcomes, necessitating stringent data preprocessing and validation to ensure data integrity. Methodologically, integrating causal discovery, like NOTEARS, into SESA offers promising research directions. Enhancing this integration could improve the method's ability to generate informed and context-aware imputation models, offering deeper insights into healthcare data's causal structures.

Future work will likely involve incorporating advanced analytical techniques, such as deep learning, to enhance SESA's data interpretation capabilities. Broadening SESA's applicability to diverse data types and healthcare areas will increase its utility and impact in medical research.

In conclusion, while SESA stands as a significant advancement in data imputation, continuous research is vital to optimize its efficacy and broaden its application spectrum in healthcare analytics.

Acknowledgments

The work was supported in part by 2020-2025 JSPS A3 Foresight Program (Grant No. JPJSA3F20200001), 2022–2024 Japan National Initiative Promotion Grant for Digital Rural City, 2022-2024 Masaru Ibuka Foundation Research Project on Oriental Medicine, 2023 and 2024 Waseda University Grants for Special Research Projects (Nos. 2023C-216 and 2024C-223), 2023-2024 Waseda University Advanced Research Center Project for Regional Cooperation Support, and 2023-2024 Japan Association for the Advancement of Medical Equipment (JAAME) Grant.

References

[1] K. J. M. Janssen, A. R. T. Donders, F. E. Harrell, Y. Vergouwe, Q. Chen, D. E. Grobbee, and K. G. M. Moons, "Missing covariate data in medical research: to impute is better than to ignore." *Journal of clinical epidemiology*, vol. 63 7, pp. 721–7, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:38075961

[2] D. B. Rubin, "Inference and missing data," *Psychometrika*, vol. 1975, p. 19, 1975. [Online]. Available: https://api.semanticscholar.org/CorpusID:120971461

[3] W. A. Fuller, *Measurement Error Models*. Wiley, June 1987.

[4] R. H. E. Hoyle, *Structural equation modeling: Concepts, issues, and applications*. Sage Publication, 1995.

[5] C. K. Enders and D. L. Bandalos, "The relative performance of full information maximum likelihood estimation for missing data in structural equation models," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 8, pp. 430 – 457, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:33752188

[6] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art." *Psychological methods*, vol. 7 2, pp. 147–77, 2002. [Online]. Available: https://api.semanticscholar.org/CorpusID:7745507

[7] N. A. Lazar, "Statistical analysis with missing data," *Technometrics*, vol. 45, pp. 364 – 365, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:40372606

[8] S. Fielding, P. M. Fayers, J. H. Loge, M. S. Jordhøy, and S. Kaasa, "Methods for handling missing data in palliative care research." *Palliat Med*, vol. 20, no. 8, pp. 791–798, Dec 2006.

[9] J. W. Graham, "Missing data analysis: making it work in the real world." *Annual review of psychology*, vol. 60, pp. 549–76, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:8417694

[10] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses." *J Sch Psychol*, vol. 48, no. 1, pp. 5–37, Feb 2010.

[11] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine*, vol. 30, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:37379599

[12] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:17840217

[13] P. Li, E. A. Stuart, and D. B. Allison, "Multiple Imputation: A Flexible Tool for Handling Missing Data," *JAMA*, vol. 314, no. 18, pp. 1966–1967, 11 2015. [Online]. Available: https://doi.org/10.1001/jama.2015.15281

[14] G. L. Mazza, C. K. Enders, and L. S. Ruehlman, "Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation," *Multivariate Behavioral Research*, vol. 50, no. 5, pp. 504–519, 2015, pMID: 26610249. [Online]. Available: https://doi.org/10.1080/00273171.2015.1068157

[15] C. K. Enders, "Multiple imputation as a flexible tool for missing data handling in clinical research." *Behav Res Ther*, vol. 98, pp. 4–18, Nov 2017.

[16] T. D. Nelson, R. L. Brock, S. Yokum, C. C. Tomaso, C. R. Savage, and E. Stice, "Much ado about missingness: A demonstration of full information maximum likelihood estimation to address missingness in functional magnetic resonance imaging data," *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.746424

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] L. Chen, J. Su, T. W. Lam, and R. Luo, "Exploring pair-aware triangular attention for biomedical relation extraction," *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263623066

[19] P. D. Allison, "Missing data techniques for structural equation modeling." *J Abnorm Psychol*, vol. 112, no. 4, pp. 545–557, Nov 2003.

[20] D. L. Schminkey, T. von Oertzen, and L. F. C. Bullock, "Handling missing data with multilevel structural equation modeling and full information maximum likelihood techniques." *Research in nursing & health*, vol. 39 4, pp. 286–97, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:3353670

[21] A. A. Igolkina and G. Meshcheryakov, "semopy: A python package for structural equation modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 27, no. 6, pp. 952–963, 2020. [Online]. Available: https://doi.org/10.1080/10705511.2019.1704289

[22] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," in *Neural Information Processing Systems*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53217974