# Synthetic Multimodal Data Modelling for Data Imputation

Paper Authors: *Francisco Carrillo-Perez, Marija Pizurica, Kathleen Marchal, and Olivier Gevaert*

Discussed by **Yikun Zhang**

AI Health Reading Group
January 31, 2025

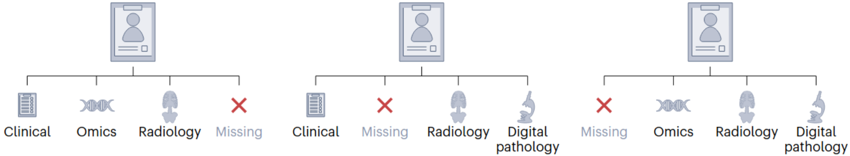
Department of STATISTICS

UNIVERSITY *of* WASHINGTON

## Paper Background

**Problem:** Missing data is a persistent problem in biomedical research.

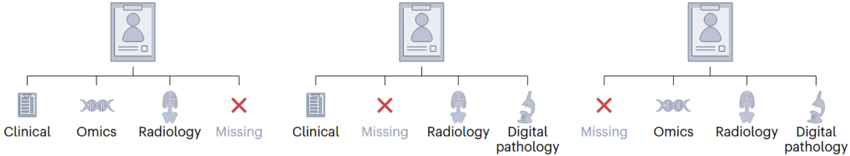
- Acquiring multiple data modalities for each patient is often expensive.



# Paper Background

**Problem:** Missing data is a persistent problem in biomedical research.

- Acquiring multiple data modalities for each patient is often expensive.

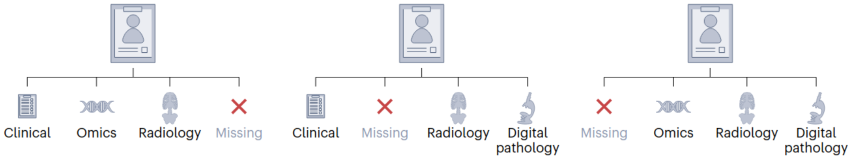



**Challenge:** Most of the existing data-imputation techniques can only handle a single data modalities.

- Their predictions rely heavily on “similarities” between data points.

**Problem:** Missing data is a persistent problem in biomedical research.

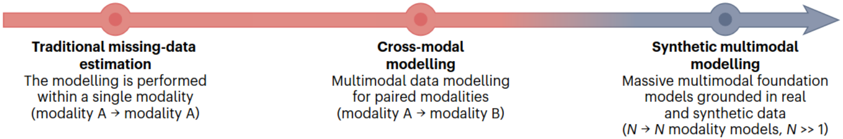
- Acquiring multiple data modalities for each patient is often expensive.



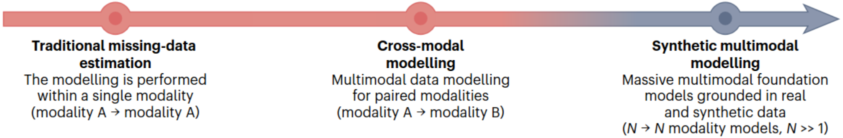

**Challenge:** Most of the existing data-imputation techniques can only handle a single data modalities.

- Their predictions rely heavily on “similarities” between data points.

- Diverse test and data modalities supply complementary insight into a distinct facet of the patient’s health or disease state.

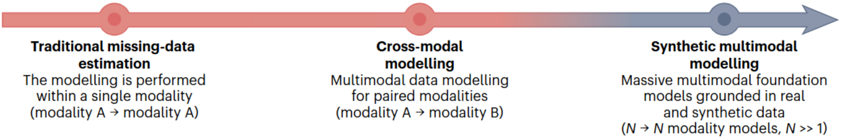


# Paper Exposition



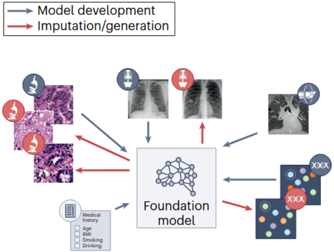

**Prospective Solution:** Synthetic multimodal data modeling.

- This framework utilizes foundation models to impute missing data and to generate realistic synthetic samples (Carrillo-Perez et al., 2024).




**Prospective Solution:** Synthetic multimodal data modeling.

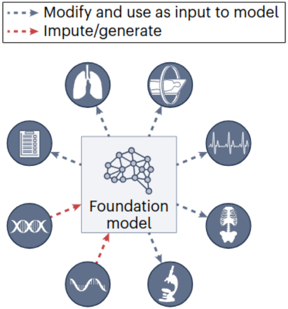
- This framework utilizes foundation models to impute missing data and to generate realistic synthetic samples (Carrillo-Perez et al., 2024).
- Foundation models integrate multimodal information into (low-dim) embeddings so as to capture complex interactions between modalities.



# Advantages of Synthetic Multimodal Data Modeling

1. Dive deeper into the joint data distribution of modalities, and thus enhance imputation quality.

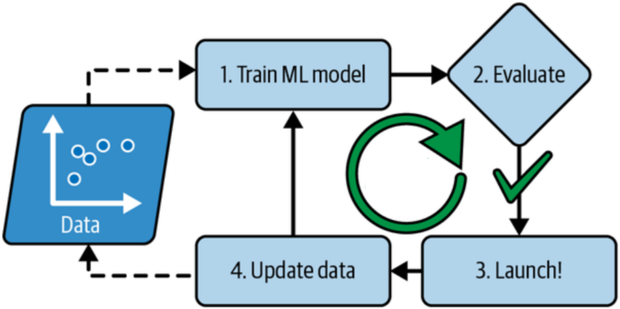
2. Explore multi-faceted knowledge through in silico hypothesis testing (*i.e.*, via computer simulations).
   - Perform interventions and ablation studies into certain data modalities or study the effect on generated synthetic modalities (Roohani et al., 2024).




   - Synthetic data from the model can be recycled, facilitating self-supervised learning (Krishnan et al., 2022).

# Advantages of Synthetic Multimodal Data Modeling

3. Offer unique flexibility when handling evolving patient data.
   - Dynamically update the model representation of all modalities available, *i.e.*, online learning mechanism.




   - This is achievable due to the gradient descent updates of modern ML model training scheme.

# Adoption Challenges of Synthetic Multimodal Data Modeling

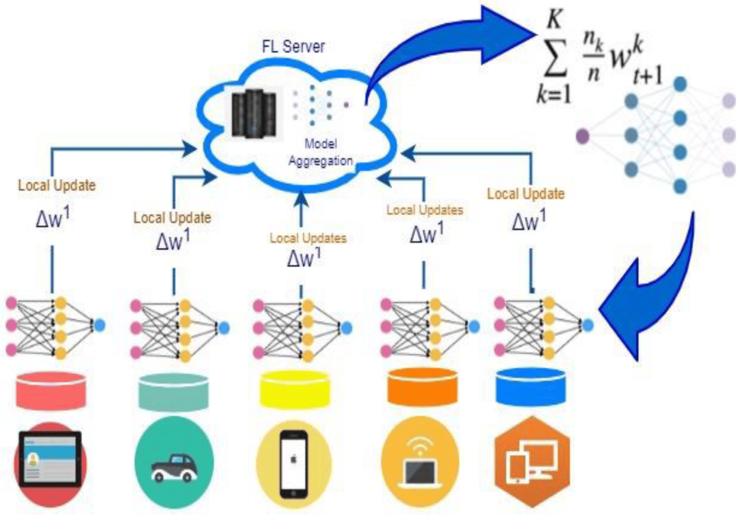
1. How can we evaluate the quality of generated data?
   - Existing metrics, such as Fréchet inception distance, may contain flaws (Stein et al., 2024).

$$d_F(\mu, \nu) = \left[ \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} ||x - y||^2 \, d\gamma(x, y) \right],$$

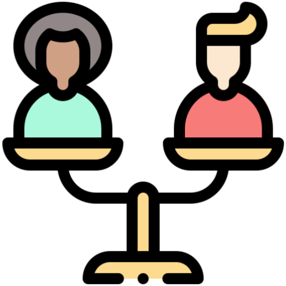
     where $\Gamma(\mu, \nu)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals $\mu$ and $\nu$ on the first and second factor, respectively.

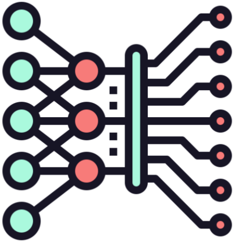
2. How can we address maliciously generated data, *i.e.*, deepfake?
   - Introduce visually imperceptible yet computationally detectable watermarks.

3. How can foundation models comply with data privacy regulations?
   - Perhaps we can use federated learning (Kairouz et al., 2021).



4. How can we maintain the algorithmic fairness of foundation models?
   - Current data sources are often skewed towards developed countries and male patients.





5. How can foundation models handle missing-not-at-random data?
   - Models may be overfitting to specific missingness patterns in the training data.

# Thank you!

More details can be found in

Carrillo-Perez, F., Pizurica, M., Marchal, K. and Gevaert, O. "Synthetic Multimodal Data Modelling for Data Imputation." *Nature Biomedical Engineering* (2024): 1-5. https://www.nature.com/articles/s41551-024-01324-1.

# Reference


F. Carrillo-Perez, M. Pizurica, K. Marchal, and O. Gevaert. Synthetic multimodal data modelling for data imputation. *Nature Biomedical Engineering*, pages 1–5, 2024.

P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

R. Krishnan, P. Rajpurkar, and E. J. Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

Y. Roohani, K. Huang, and J. Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.