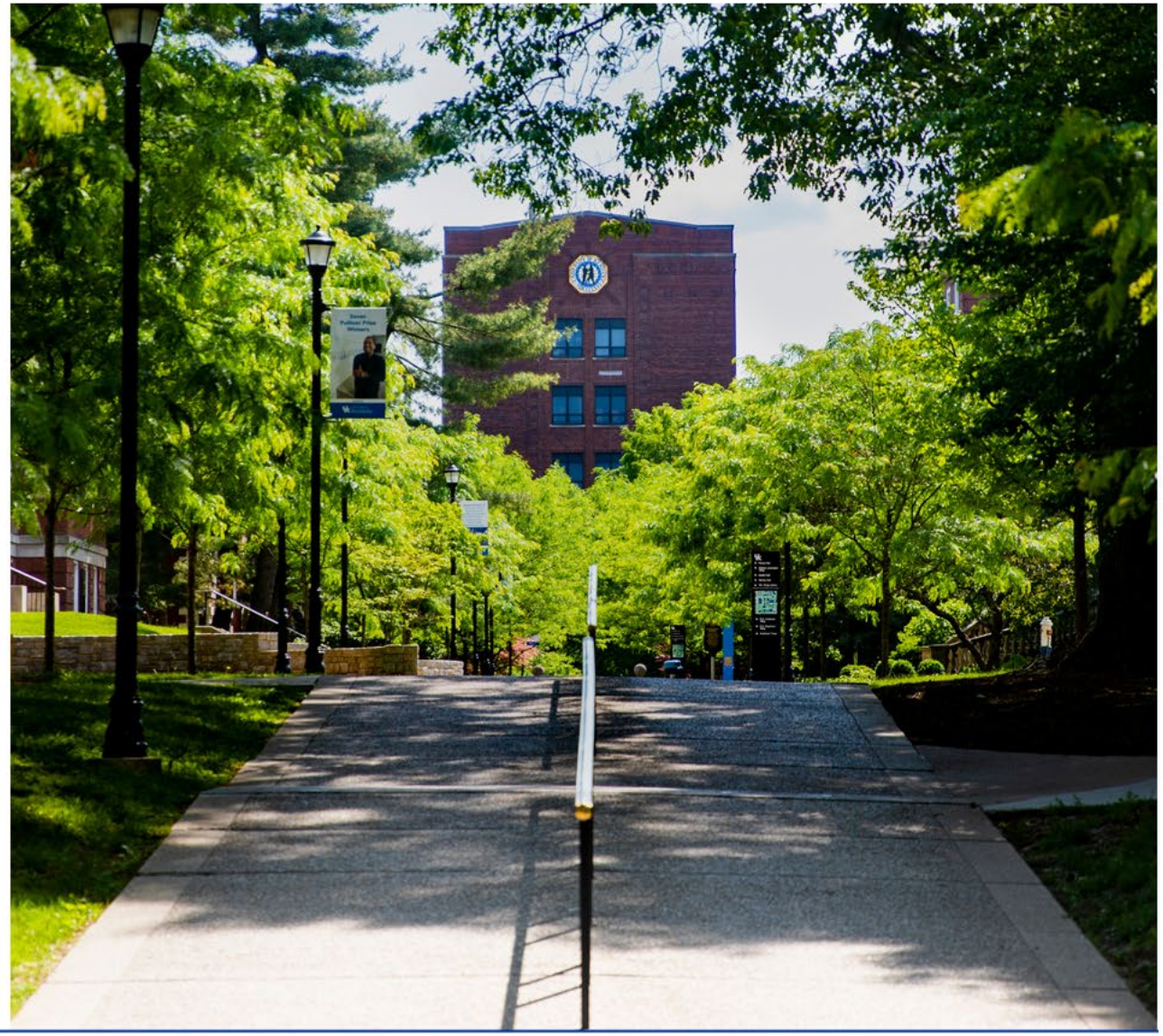
An Introduction to Generative AI in Biomedical Applications, Part 1

Anthony A. Mangino, PhD

UK AI/ML Hub Seminar Series

October 31, 2024



University of Kentucky

Outline

- Who am *I* to talk to *you* about generative AI?
- Introducing Synthetic Data: Generative Adversarial Networks
- Clinical Use Case: Diagnosing Takotsubo Syndrome
- Clinical Prediction Model + Synthetic Data
- Ethical Implications
- Future Directions

Acknowledgments

Taha Ahmed, MD

Vincent Sorrell, MD

Samra Haroon Lodhi, MD

Vedant Gupta, MD

Steven Leung, MD

W. Holmes Finch, PhD

Maria Hernandez Finch, PhD

Kendall Smith, MS

Who am *I* to talk to *you* about generative AI?

Who am *I* to talk to *you* about generative AI?

BALL STATE
UNIVERSITY

- All my degrees are in some flavor of *Psychology*
- Worked as a research consultant throughout graduate school
- My research focused in prediction models…
- …until I discovered generative models.

Who am *I* to talk to *you* about generative AI?

MEASUREMENT AND EVALUATION IN COUNSELING AND DEVELOPMENT
https://doi.org/10.1080/07481756.2021.1906156

Routledge
Taylor & Francis Group

Check for updates

ASSESSMENT, DEVELOPMENT, AND VALIDATION

Improving Predictive Classification Models Using Generative Adversarial Networks in the Prediction of Suicide Attempts

Anthony A. Mangino, Kendall A. Smith, W. Holmes Finch, and Maria E. Hernández-Finch

Ball State University Teachers College, Muncie, IN, USA

American Psychological Association

Technology, Mind & Society
2021 Conference Proceedings

http://dx.doi.org/10.1037/xxxxxxx

ISBN:

Modeling Responsibly

Toward a Fair, Interpretable, and Ethical Machine Learning for the Social Sciences

Anthony A. Mangino[1], Kendall A. Smith[2], and W. Holmes Finch[3]

Current work in the Biostat CIRCL

Since 2021, I have been a biostatistician in the Biostatistics Consulting and Interdisciplinary Research Collaboration Lab (Biostat CIRCL).

My collaborative work has been with researchers in units including:

- Internal Medicine; Division of Hospital Medicine
- KIPRC; Overdose Data to Action (OD2A)
- College of Social Work
- Department of Surgery; Division of Surgical Oncology
- Department of Otolaryngology – Head & Neck Surgery

Current work in the Biostat CIRCL

- This collaborative work has provided me with a broader sense of…
  - Why we do clinical and epidemiological research;
  - What responsibilities we hold in doing this work;
  - How quantitative scientists contribute to these projects.

- A couple of these collaborative projects have led to [statistics] conference presentations on the use of generated data/generative models to augment/replicate existing datasets.

- Collectively, my work has led me to this conclusion…

Axiom: As statisticians, we are also ethicists.

Introducing Synthetic Data: Generative Adversarial Networks

I might not call this AI, but you might. I am very much an AI skeptic. Caveat emptor.

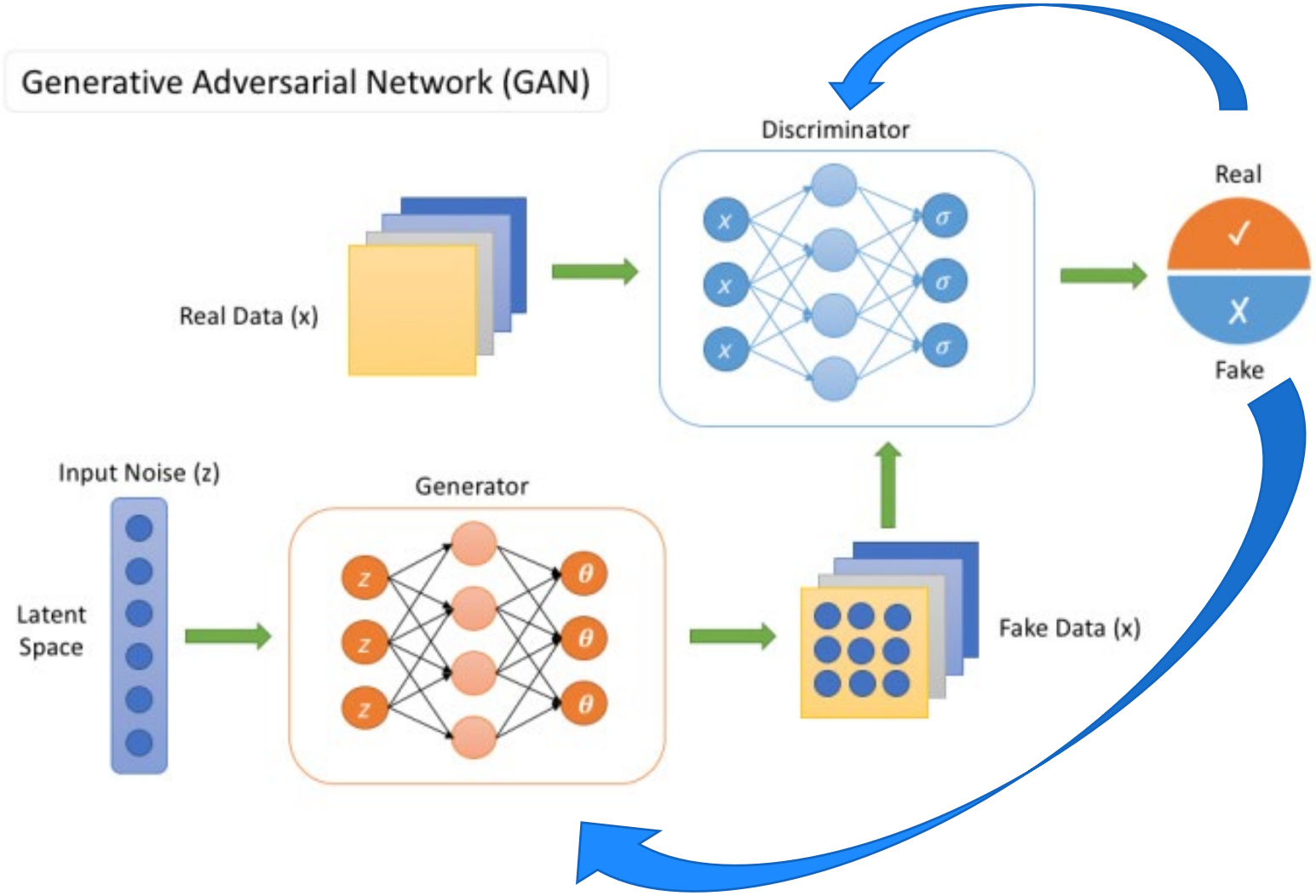
What are Generative Adversarial Networks?

Generative Adversarial Networks (GANs) were created (or at least published!) in 2014 by Ian Goodfellow while he was a student at Université de Montréal (he subsequently worked at Google Brain).

The objective of GANs is to create synthetic data that look and behave like real/source data.

Originally created for use with image data, GANs have a variety of possible architectures that are relevant to and useful for image, video, audio, and tabular data.

They can also be used in reinforcement learning or computer vision tasks.



"We train D to maximize the probability of assigning the correct label to both training examples and samples from G. We simultaneously train G to minimize log(1 – D(G(z)))."

Goodfellow et al., 2014

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

Diagram credited to Abhishek Verma: https://www.linkedin.com/pulse/generative-adversarial-network-abhishek-verma/

≠

University of Kentucky

Images from Goodfellow et al., 2014

Our First Dive into Generative Modeling

Goal: Create more effective prediction models for adolescents at risk of attempting suicide (2.6%) using generated data.

Data: 2017 CDC Youth Risk Behavior Survey

Comparing data balanced using:

- GAN-generated cases
- Synthetic minority oversampling technique
- Random over-sampling examples
- Bootstrapped oversampling

Models trained using GAN data performed surprisingly well, often outperforming all other methods.

MEASUREMENT AND EVALUATION IN COUNSELING AND DEVELOPMENT
https://doi.org/10.1080/07481756.2021.1906156

ASSESSMENT, DEVELOPMENT, AND VALIDATION

Improving Predictive Classification Models Using Generative Adversarial Networks in the Prediction of Suicide Attempts

Anthony A. Mangino, Kendall A. Smith, W. Holmes Finch, and Maria E. Hernández-Finch

Ball State University Teachers College, Muncie, IN, USA

Our First Dive into Generative Modeling

Methods:

- N = 203,663 adolescents in 9th– 12th grade
- **Predictors (p = 22):** Demographics; Drug use; Sexual behaviors; Physical health & exercise; Feelings of sadness/hopelessness; Physical/sexual abuse & violence

- Split source dataset into 75%/25% training/test.
- Subset of 10,000 random cases from each, the training and test datasets.
- Feed training set into the GAN to obtain synthetic training set

- Prediction Models:
  - Logistic Regression
  - Random Forest
  - Boosting
  - Bayesian additive regression trees (BART)

All prediction accuracy rates were obtained using the *source test set*.

Our First Dive into Generative Modeling

Goal: Create more effective prediction models for adolescents at risk of attempting suicide using generated data.

Data: 2017 CDC Youth Risk Behavior Survey

Comparing data balanced using:

- GAN-generated cases
- Synthetic minority oversampling technique
- Random over-sampling examples
- Bootstrapped oversampling

Models trained using GAN data performed surprisingly well, often outperforming established methods.

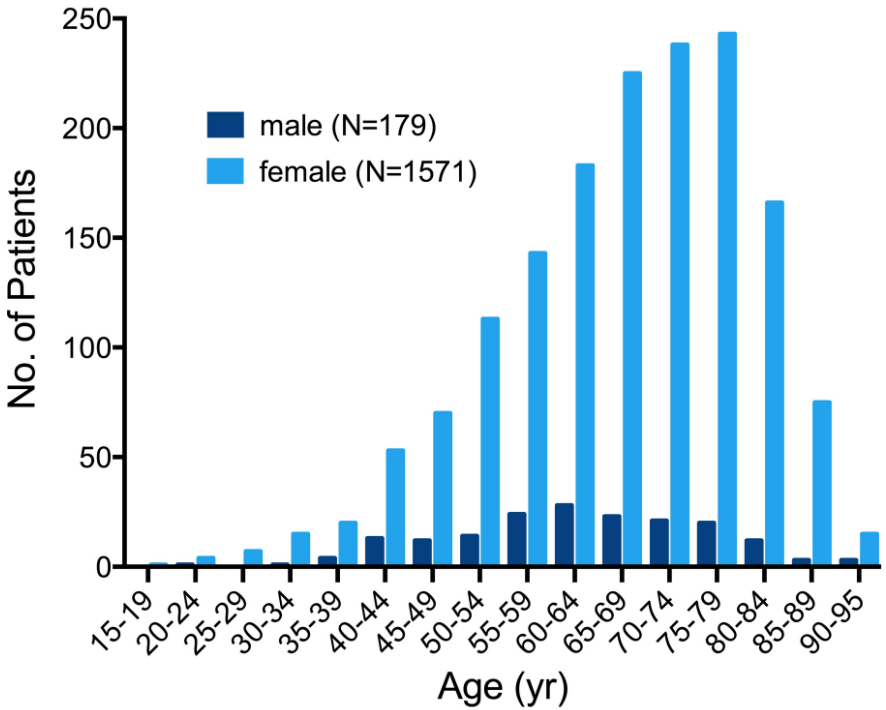
Table 2. Overall Classification Accuracy, Specificity, Sensitivity, AUC, and CE by Sampling Method and Estimation Model.[a,b]

Method	Logistic Regression	Random Forest	Boosting	BART
Overall Classification Accuracy				
Standard	0.891	0.959	0.956	0.948
Combined	**0.966**	0.976	**0.965**	**0.974**
GAN	0.557	0.941	0.941	0.924
SMOTE	0.817	***0.979***	0.956	0.878
ROSE	0.864	0.954	0.964	0.867
Over-sampling	0.873	0.903	0.947	0.892
Specificity				
Standard	0.890	0.984	0.977	0.997
Combined	**0.996**	**0.999**	**0.992**	***1.000***
GAN	0.558	0.955	0.954	0.994
SMOTE	0.683	0.984	0.980	0.976
ROSE	0.780	0.940	0.986	0.970
Over-sampling	0.789	0.920	0.967	0.940
Sensitivity				
Standard	0.087	0.110	0.149	0.030
Combined	0.892	0.879	0.889	0.788
GAN	0.516	0.924	**0.924**	**0.836**
SMOTE	0.952	***0.975***	0.877	0.779
ROSE	0.949	0.967	0.901	0.762
Over-sampling	**0.957**	0.854	0.852	**0.836**
AUC				
Standard	**0.891**	0.943	0.935	0.893
Combined	0.541	0.931	0.833	0.515
GAN	0.537	0.955	**0.985**	**0.915**
SMOTE	0.818	***0.998***	0.840	0.877
ROSE	0.864	0.990	0.851	0.866
Over-sampling	0.873	0.917	0.836	0.892

Clinical Use Case: Diagnosing Takotsubo Syndrome

Takotsubo Syndrome

- **Takotsubo Syndrome (TTS; a.k.a. 'broken heart syndrome')** is a relatively rare and **reversible** condition with symptoms that mimic specific clinical presentation of **left anterior descending acute coronary syndrome** (LAD-ACS):
  - severe pressure and/or pain in chest,
  - shortness of breath,
  - sudden onset fatigue,
  - cold sweats,
  - lightheadedness.
- Rarely reported prior to the early 2000s.
- Often preceded by great emotional and/or physical stress.
- TTS is far more prevalent in women than men.

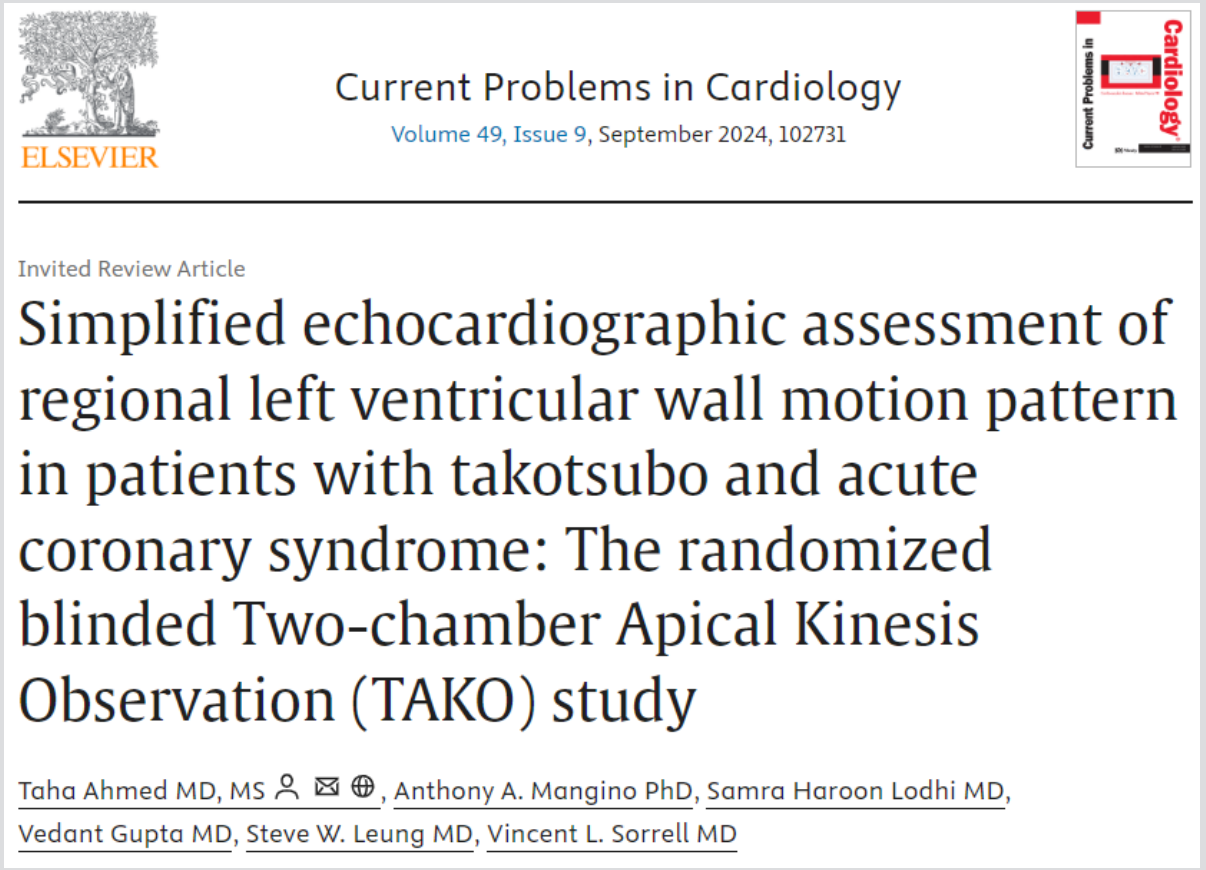


Sharkey et al., (2018)
Templin et al., (2015).

Our Original Study

- Our goal was to assess the capability of an **echocardiogram** to provide sufficient information for a clinician to diagnose TTS compared to ACS **in the absence of the conventional coronary angiography**.

- **N = 102 patients** (complete cases) fulfilling the Mayo Clinic criteria (Madhavan & Prasad, 2010) for TTS presenting to University of Kentucky Healthcare hospitals between 2011 and 2021.



ELSEVIER

Current Problems in Cardiology

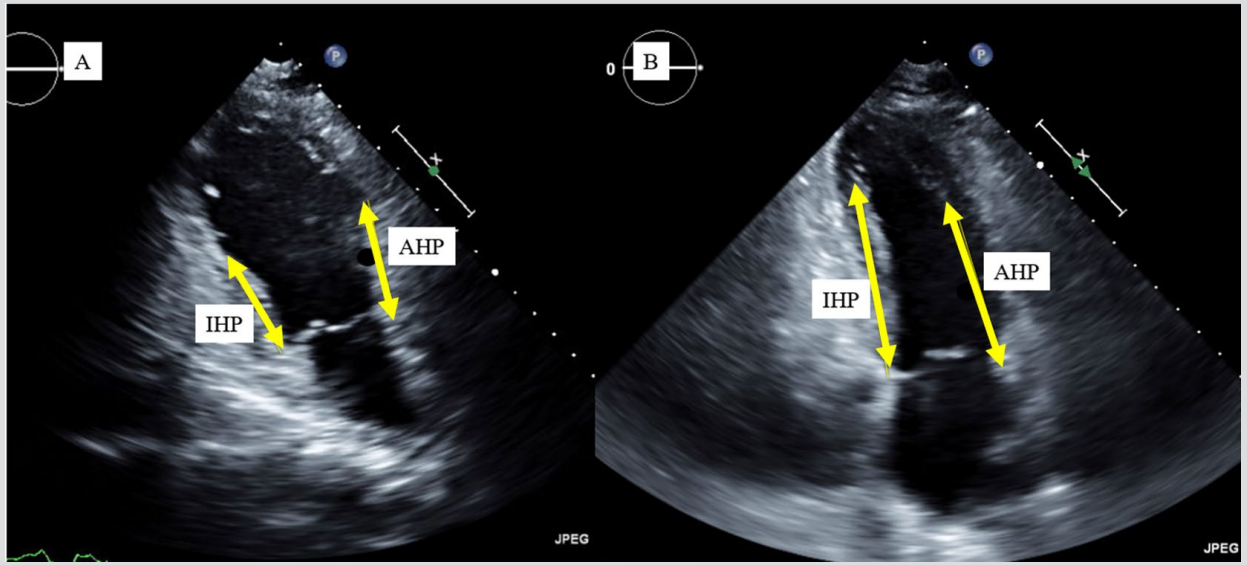
Volume 49, Issue 9, September 2024, 102731

Invited Review Article

Simplified echocardiographic assessment of regional left ventricular wall motion pattern in patients with takotsubo and acute coronary syndrome: The randomized blinded Two-chamber Apical Kinesis Observation (TAKO) study

Taha Ahmed MD, MS, Anthony A. Mangino PhD, Samra Haroon Lodhi MD, Vedant Gupta MD, Steve W. Leung MD, Vincent L. Sorrell MD

Our Original Study



- Echocardiogram used apical 2-chamber view to assess anterior and inferior wall segments to hinge points.
- Because raw hinge point measurements would be affected by patient sex, the **ratio** of the measurements was used.
- **Ratio of anterior to inferior hinge points** was hypothesized to generally be:
  - Near or greater than 1 in TTS patients
  - Less than 1 in ACS patients

Our Original Study: Model Derivation Cohort Results

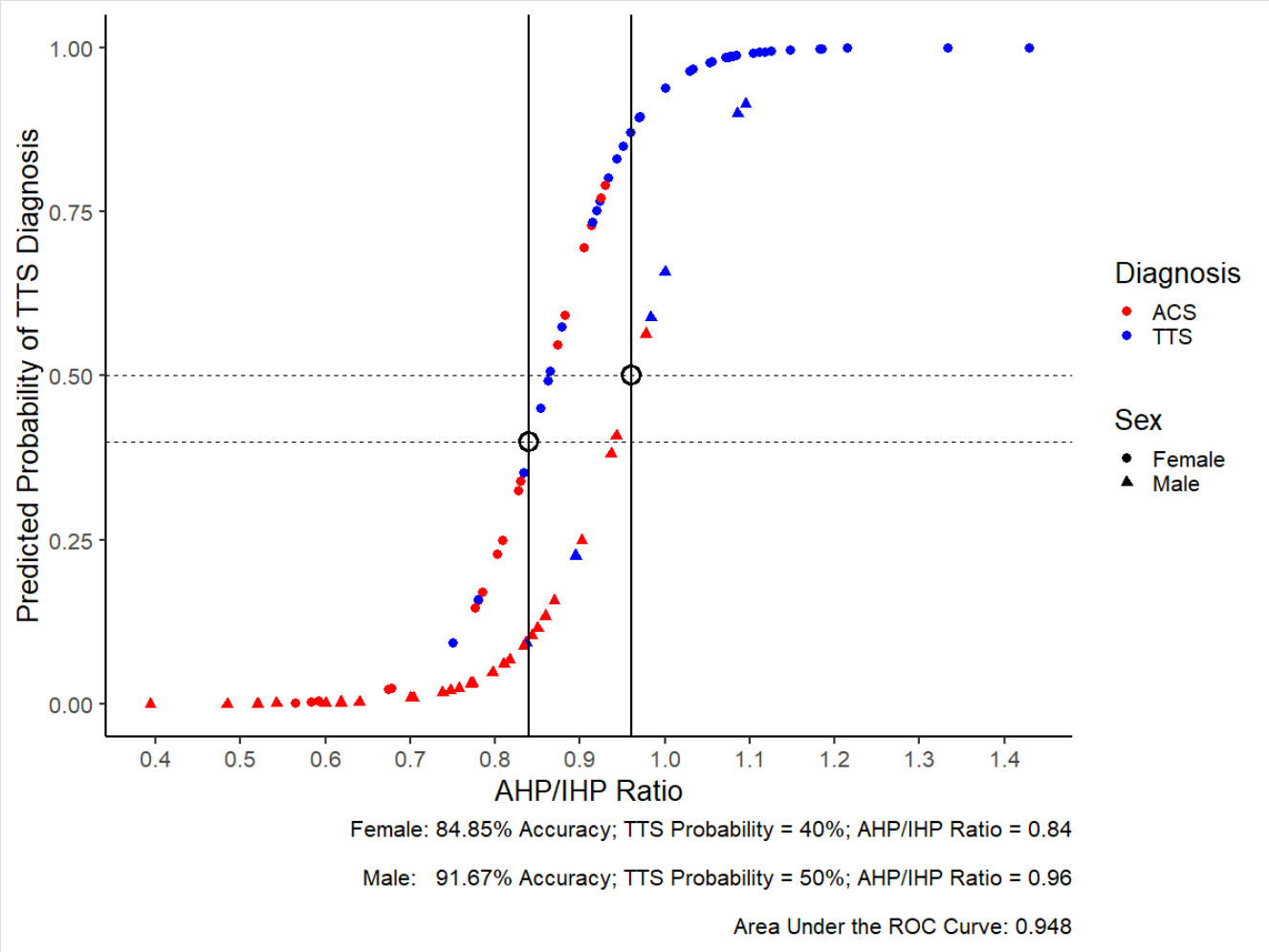
- **Logistic regression model** was fit with the following specification:

*diagnosis ~ AHP/IHP Ratio * Sex*

Sensitivity = Correct TTS Diagnosis

Specificity = Correct ACS Diagnosis

	ACS	TTS
Female	20	46
Male	30	6



Validation Cohort

Because of the rarity of TTS, we used our derivation cohort (n = 102) as an **internal validation cohort** by:

- Recruiting 8 readers to review derivation cohort echocardiographs
  - One fellow
  - Four assistant professors
  - Two associate professors
  - One full professor
- Randomly assigning between 3 and 5 readers to each patient record (29 or 30 records per reader)
- Collect new AHP & IHP measurements, and predicted diagnosis

- Concordance between charted diagnosis and clinician-predicted diagnosis was 70.6%

Validation Cohort: Results

Training Set		Source	Synthetic
Source Validation Set	Training	X	
	Validation		

We assessed our logistic regression model using actual charted diagnosis as outcome, and clinician-predicted diagnosis as the outcome.

Metric	Overall Accuracy	Sensitivity (TTS)	Specificity (ACS)
Training Set	0.853	0.865	0.840
Validation: Actual Charted Diagnosis	0.68	0.78	0.586
Validation: Clinician-Predicted Diagnosis	0.704	0.816	0.608

- In both cases, we used the ratio of reader-derived AHP & IHP measurements.
- Can we do better? This is not [yet] good enough for use in a clinical setting!

Clinical Use Case: Takotsubo Syndrome + Synthetic Training Cases

Synthetic Training Cohort

Because TTS is a) a rare phenomenon, and b) a sex-imbalanced diagnosis, we used a GAN to **create a larger synthetic training sample**.

The GAN was fit with a batch size of 50 across 1000 epochs with a Wasserstein value/loss function to yield 1020 cases.

The GAN was fit using the RGAN package in R (Neunhoeffer, 2022).

Significant differences between datasets found only for patient sex (p = 0.006).

	Source Data	GAN-Generated Data
Female, n (%)	66 (64.7%)	510 (50.0%)
Male, n (%)	36 (35.3%)	510 (50.0%)
Age, M (s)	59.6 (12.7)	58.8 (6.09
Inferior Hinge Point, M (s)	4.71 (1.47)	4.58 (1.95)
Anterior Hinge Point, M (s)	4.02 (0.995)	4.10 (1.28)
AHP/IHP Ratio, M (s)	0.891 (0.184)	0.889 (0.215)
ACS Diagnosis, n (%)	50 (49.0%)	510 (50.0%)
TTS Diagnosis, n (%)	52 (51.0%)	510 (50.0%)
Male TTS Cases, n (%)	6 (5.9%)	23 (2.2%)

Synthetic Cohort: Concordance with Source Training Cohort

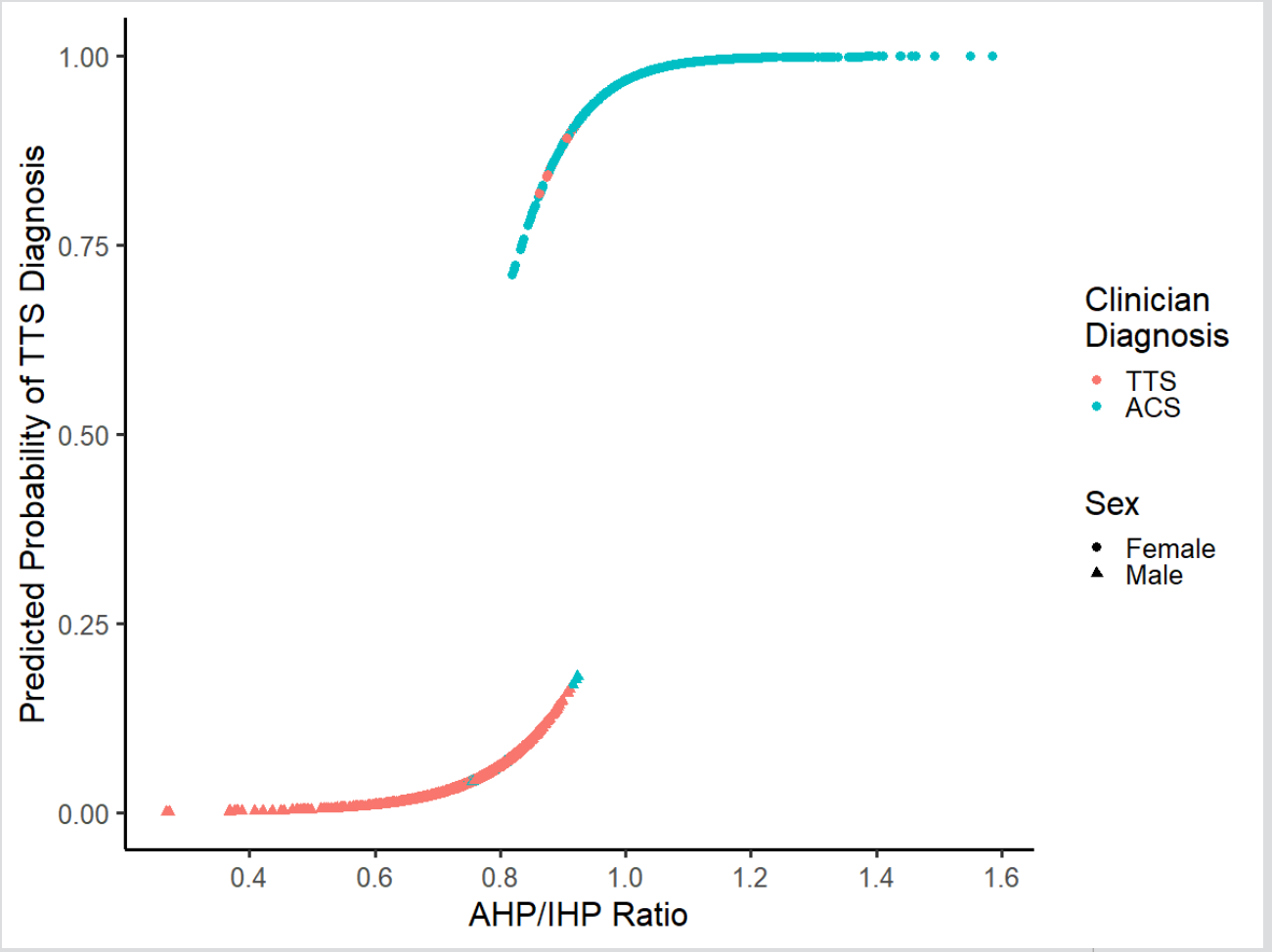
Training Set		Source	Synthetic
Source Validation Set	Training		X
	Validation		

We fit a logistic regression model with the same specification, but with **only the 1020 synthetic cases**.

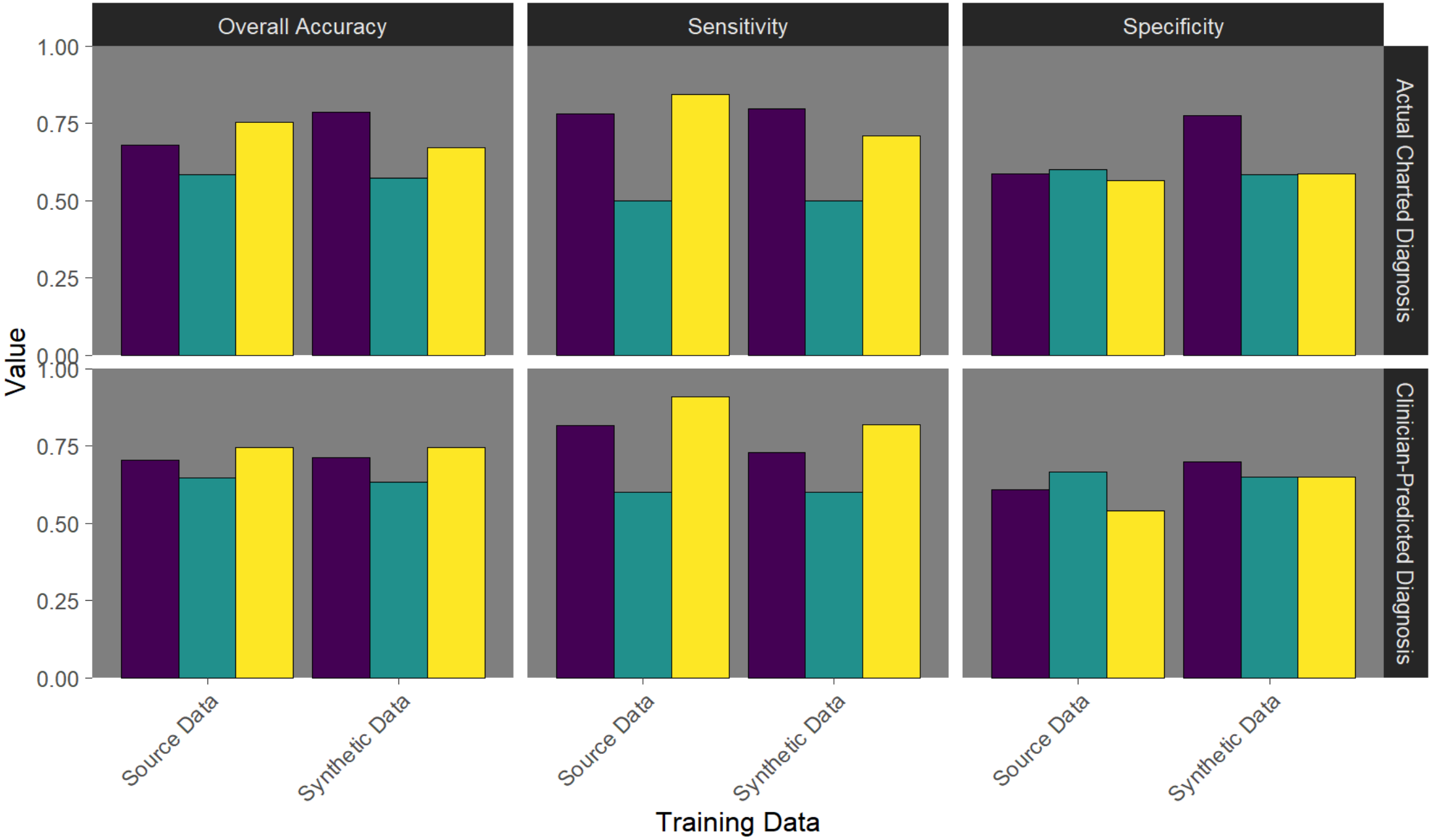
*diagnosis ~ AHP/IHP Ratio * Sex*

Agreement with source training cohort was assessed as if the source cohort were a novel set of cases (i.e., the validation set).

Cutoff = 0.6
Overall Accuracy = 83.33%
Sensitivity = 86.00%
Specificity = 80.77%



Validation Cohort: Results with Synthetic Training Set



Cutpoints

Overall = 0.7

Male = 0.2

Female = 0.9

Conclusions

- Using a logistic regression trained on GAN-generated data **did not yield more precise predictions** in our validation cohort than from a logistic regression trained on our source data.

- This **contradicts our previous work** using both clinical (Mangino, 2023) and educational (Mangino et al., 2021) data.

- Previous research indicates **the classifier itself** (LR vs RF vs Boosting vs etc.) has an appreciable effect on the utility of GAN-generated data (Mangino et al., 2021).

- It is possible that as the GAN creates data that more closely match the source data, our secondary model results more closely match those obtained from source data.

Ethical Implications & Future Directions

Axiom: As statisticians, we are also ethicists.

Ethical Implications

- We've determined that our synthetic data are only as informative as our source data.
- Without a well-behaved and comprehensive dataset, our GAN can create a simulacrum **faithful to** the source data, but not necessarily **better than**.
- What does this mean for practice?...
- Using **good** generated data does not automatically beget **better clinical decision-making**.

Like any other model, generative models are only as good as their source data.

The Embedded Ethical Problem

Just because we **can** build better models (**assuming we can**)doesn't directly entail that we **must** use them.

- But when is it ethical to do so?
- When is it our responsibility to do so?
- When is it our responsibility not to?

Example: We build a classification model to identify students at risk of dropping out of high school using generated data; our predictions are more accurate than the same model using real data. A student is identified as being at risk of dropping out. How do you explain this to the student's parents?

Responsibility to Constituency

Responsibility to Science

One Partial Solution: Interpretability

One major area absent in research on generative modeling is interpretability.

Interpretation of secondary models (e.g., logistic regression) is paramount in most of our research.

[Why] Are generative Models not held to the same standard?

Technology, Mind & Society
2021 Conference Proceedings

http://dx.doi.org/10.1037/xxxxxxx

ISBN:

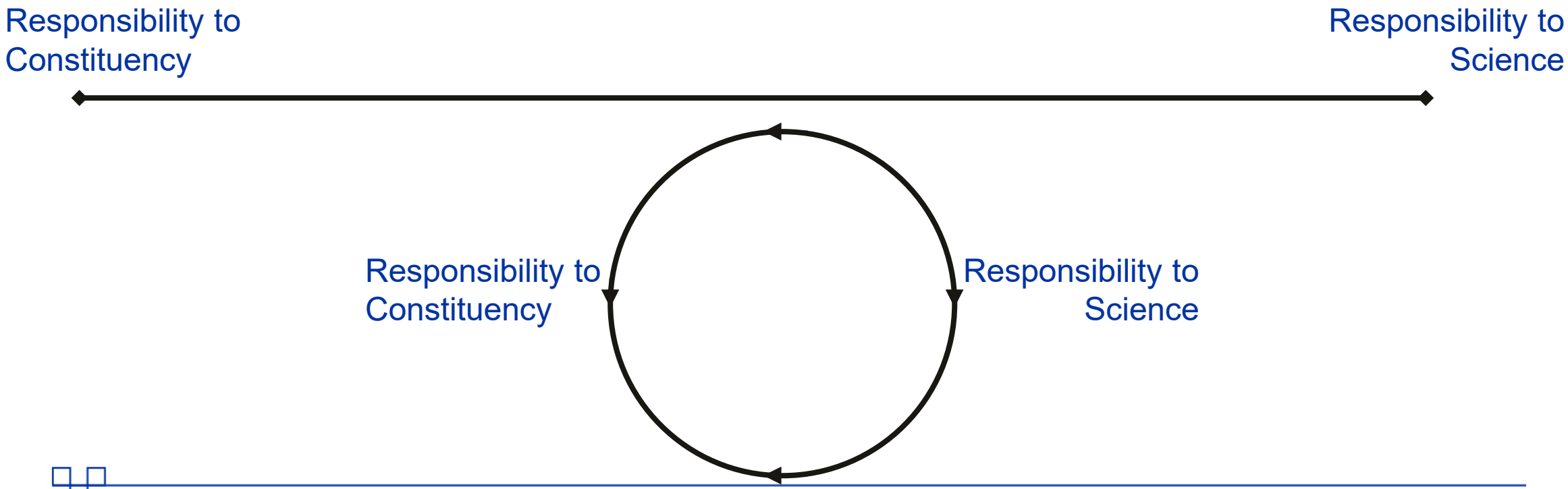
Modeling Responsibly

Toward a Fair, Interpretable, and Ethical Machine Learning for the Social Sciences

Anthony A. Mangino[1], Kendall A. Smith[2], and W. Holmes Finch[3]

One Partial Solution: Interpretability

If we are able to develop standardized metrics to quantify the synchrony between source and generated data, we can begin building **informed trust** in generative models.



Future Directions

- More fully examine synthetic datasets to determine optimal GAN hyperparameters.
- Assess whether these results hold in small, complete datasets or if we can obtain greater precision in larger, more complex datasets.
- Determine whether second-order bias is introduced in synthetic data, whether through mechanisms like imputation or through the very process of generating synthetic data.
- Devise methods for quantifying the synchrony between source and synthetic data. Identifying interpretable metrics for generative models.

References

Ahmed, T., Mangino, A.A., Lodhi, S.H., Gupta, V., Leung, S.W., & Sorrell, V.L. (2024). Simplified echocardiographic assessment of regional left ventricular wall motion pattern in patients with takotsubo and acute coronary syndrome: The Randomized Blinded Two-chamber Apical Kinesis Observation (TAKO) Study. Current Problems in Cardiology, 102731. DOI: 10.1016/j.cpcardiol.2024.102731

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27-36*.

Madhavan, M., & Prasad, A. (2010). Proposed Mayo Clinic criteria for the diagnosis of Tako-Tsubo cardiomyopathy and long-term prognosis. Herz, 35(4), 240–243. https://doi.org/10.1007/s00059-010-3339-x

Mangino, A.A. (2023, August). Case Studies in Data Emulation and Augmentation Using Generative Adversarial Networks in Psychoeducational Data. Presented at the 2023 Joint Statistical Meetings; Toronto, ON, CA.

Mangino, A.A., Smith, K.A., Finch, W.H., & Hernández-Finch, M. E., (2021). Improving Predictive Classification Models Using Generative Adversarial Networks in the Prediction of Suicide Attempts. Measurement and Evaluation in Counseling and Development, 55(2), 116-135. https://doi.org/10.1080/07481756.2021.1906156

Neunhoeffer, N., (2022). RGAN: Generative Adversarial Nets (GAN) in R. R package version 0.1.1. https://CRAN.R-project.org/package=RGAN

Sharkey SW, Maron BJ. Survival After Takotsubo, Revisited. J Am Coll Cardiol. 2018 Aug 21;72(8):883-884. doi: 10.1016/j.jacc.2018.06.022. PMID: 30115227.

Templin, C., Ghadri, J. R., Diekmann, J., Napp, L. C., Bataiosu, D. R., Jaguszewski, M., ... & Lüscher, T. F. (2015). Clinical features and outcomes of takotsubo (stress) cardiomyopathy. New England Journal of Medicine, 373(10), 929-938.

Verma, A. (2019, July). Generative adversarial network. *Linkedin*. Https://www.Linkedin.Com/pulse/generative-adversarial-network-abhishek-verma/

Thank you!

Contact Information

Anthony A. Mangino, PhD

Department of Biostatistics

Biostatistics Consulting and Interdisciplinary Research Collaboration Lab (Biostat CIRCL)

Anthony.Mangino@uky.edu

Any questions?