ICYM^{2}I: The illusion of multimodal informativeness under missingness

Young Sang Choi*, Vincent Jeanselme*, Pierre Elias, Shalmali Joshi
[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[2]Seymour, Paul, and Gloria Milstein Division of Cardiology, Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA
* equal contribution

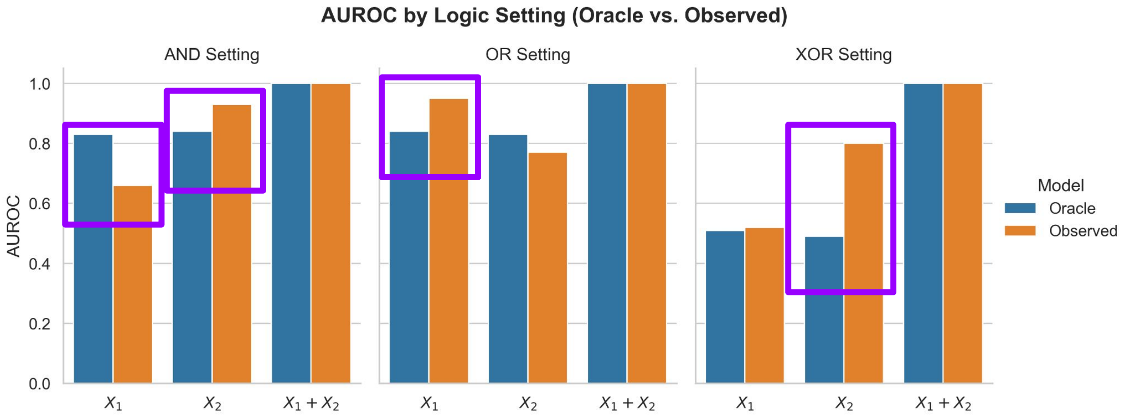
The Illusion of Multimodal Missingness

- **Prediction performance** and **modality-specific information gain** drive decisions to use multimodal or unimodal models.
- Current multimodal methods assume **no modality missingness** or a **stable missingness** process between source and target environments.
- However, the target environment may have different modality availability than the source environment (e.g. sensor failure in self-driving cars due to weather conditions).
- **This shift in modality availability can skew both prediction performance and information gain metrics.**

The Illusion in Bitwise Operators (Prediction)

X_1, X_2: bits drawn from a Bernoulli distribution with p=0.5.
Y: bitwise-logic outcome from X_1 and X_2 (AND, OR, and XOR).
X_2 has 50% missingness: $M_2 \sim Bern(0.6X_1 + 0.2)$

Oracle: train on all data.
Observed: train on observed data only.

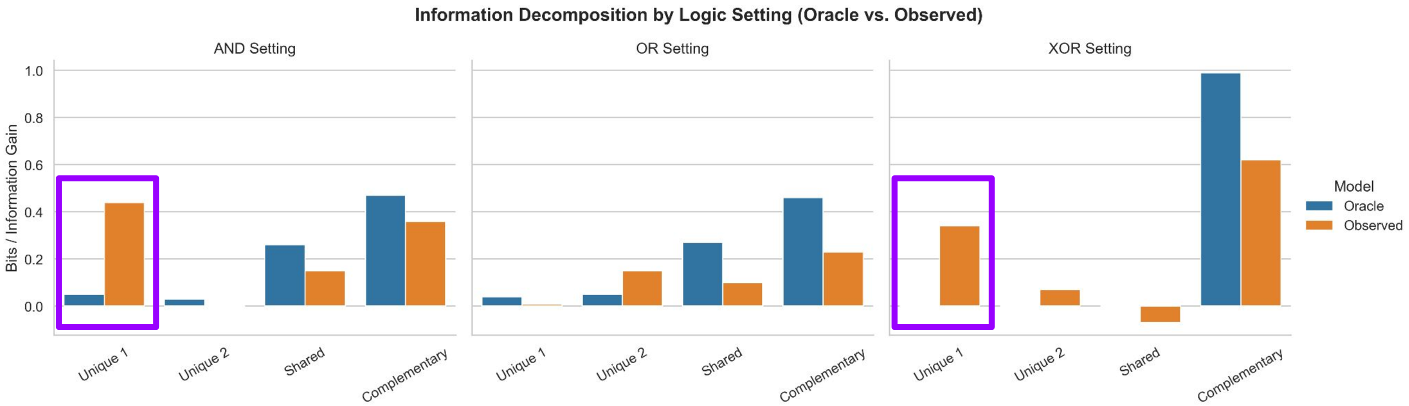


Even in this simple setting, missingness can cause skewed predictive performance estimates.

The Illusion in Bitwise Operators (Information Gain)

Leveraging **Partial Information Decomposition (PID)**:

- Unique 1 and Unique 2: Only X_1 or X_2: have information about Y.
- Shared: Both X_1 or X_2: have the same information about Y.
- Complementary: The combination of X_1 and X_2 gives us information about Y.



Similar to prediction, information gain metrics are skewed.

ICYM^2I: In Case You Multimodal Missed It

Goal:

Recover unbiased estimates of **predictive performance** and **information gain** under modality missingness.

Assumptions:

Modalities are missing at random (MAR) and positivity.

Proposed Method:

A double Inverse Probability Weighted (IPW) correction-based framework:

ICYM^2I=ICYM^2I-Learn (predictive performance)+ICYM^2I-PID (information gain)

ICYM^{2}I-Learn (Prediction)

Lemma 1 (IPW Training) *The loss function computed on the observed data $l_{\Omega_{obs}}(x_1, x_2, y)$ can be reweighted to approximate the target loss $l_{\Omega}(x_1, x_2, y)$ as follows:*

$$l_{\Omega}(x_1, x_2, y) = \frac{1}{1 - p(m_1, m_2, m_y \mid C)} \, l_{\Omega_{obs}}(x_1, x_2, y)$$

where $p(m_1, m_2, m_y \mid C)$ is the probability of missingness, given the covariates C.

- Using the MAR and Positivity assumptions, we can use Inverse Probability Weighting (IPW) to upweight under-observed samples as outlined above.
- This method is **model agnostic** and **requires only data and binary missingness labels** to train the propensity models.

ICYM^{2}I-PID (Information Gain)

Lemma 2 (Corrected mutual information)

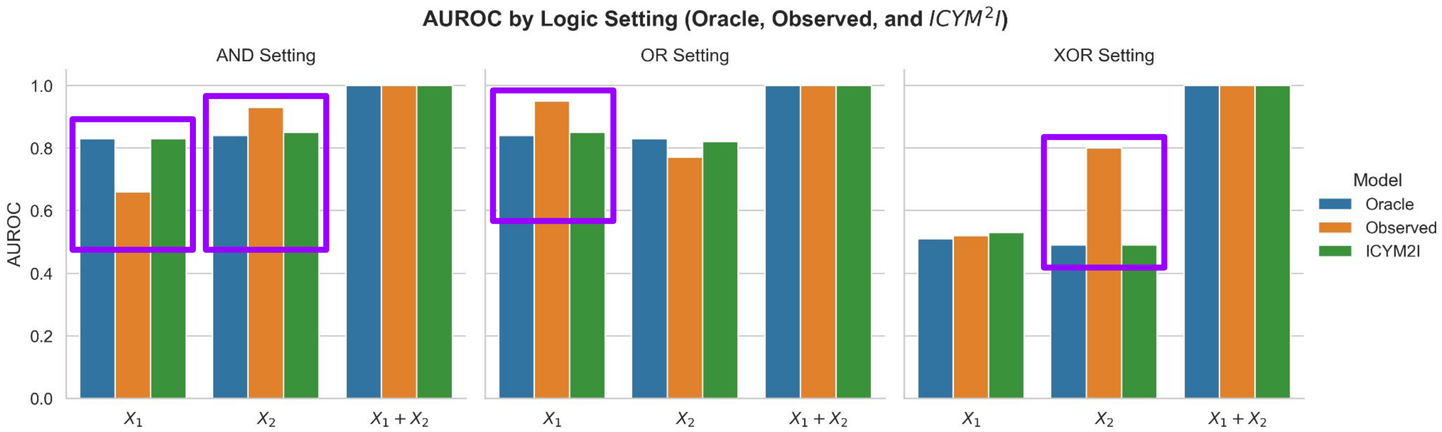
$$I_\Omega^{\mathrm{IPW}}(Y:(X_1,X_2)) = \mathbb{E}_{\substack{x_1,x_2\sim p_{\Omega_{\mathrm{obs}}}(x_1,x_2)\\ y\sim p_\Omega(y|x_1,x_2)}}\left[\frac{1-p(m_1,m_2)}{1-p(m_1,m_2|x_1,x_2,y)}\log\left(\frac{p_\Omega(x_1,x_2,y)}{p_\Omega(x_1,x_2)p_\Omega(y)}\right)\right]$$

Then, ICYM^{2}I-PID uses the following projection set to operationalize PID-bound estimation while accounting for missingness:

$$\begin{aligned}\Delta_\Omega{}^{\mathrm{ICYM^2I}} &\approx \{q \propto \exp(f_1(x_1)\cdot f_2(x_2)) : q(x_i,y) = p_{\Omega_\phi}(y,x_i)\, \forall x_i \in \mathcal{X}_i, y\in\mathcal{Y}, i\in\{1,2\}\}\\ &= \{q \propto \exp(f_1(x_1)\cdot f_2(x_2)) : q(x_i,y) = \mathrm{IPW}_{p_{\Omega_\phi}}(p_{\Omega_{\mathrm{obs}_\phi}}(y,x_i))\, \forall x_i \in \mathcal{X}_i, y\in\mathcal{Y}, i\in\{1,2\}\}\end{aligned}$$

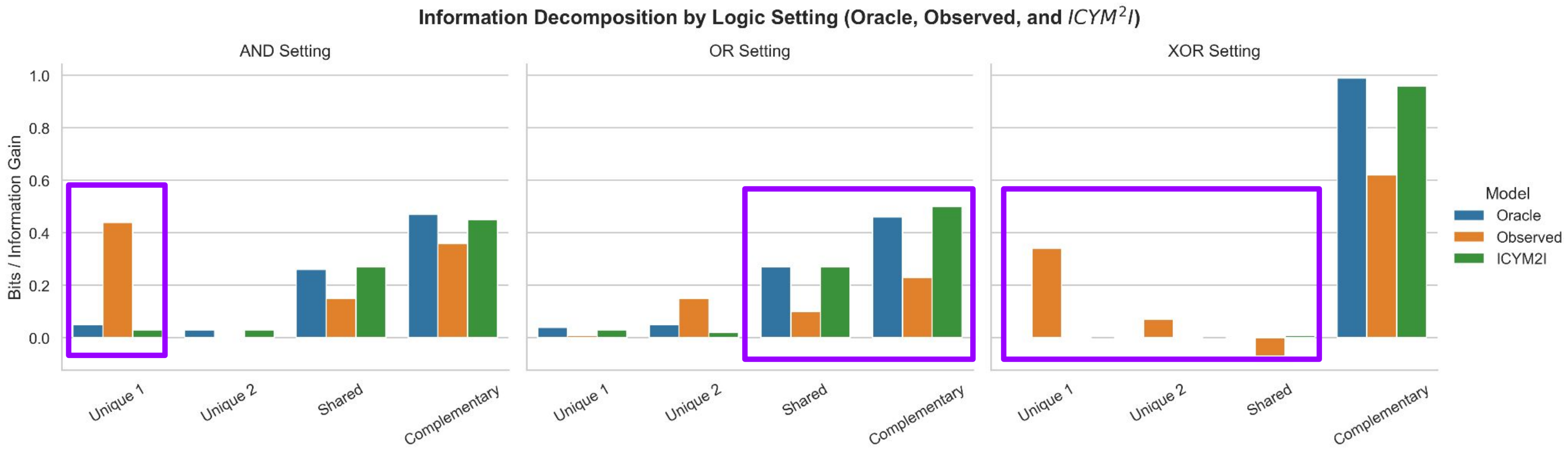
- Recovering the partial information bounds introduced by Bertschinger et al. requires minimizing the three-way mutual information between X_1, X_2, and Y.
- ICYM2I-PID corrects for the shift caused by modality missingness with the mutual information formulation above.

Bitwise Operators Revisited (Prediction)



While a model naïvely trained on the observed data has incongruent performance across unimodal settings with the oracle model, **ICYM^{2}I-Learn is able to correct for this bias**.

Bitwise Operators Revisited (Information Gain)



Similarly, ICYM^{2}I-PID corrects for partial information decomposition metrics across the three bitwise operator settings.

Semi-Synthetic Experiments

Table 3. Impact of 70% missingness on multimodality information for UR-FUNNY (Hasan et al., 2019) and Hateful Memes (Kiela et al., 2020). Parentheses denote standard deviation across batches.

		AUROC			Information Decomposition			
		Text	Image/Video	Image + Text	$\text{Unique}_{\text{text}}$	$\text{Unique}_{\text{image}}$	Shared	Complementary
UR-FU.	Oracle	0.68 (0.01)	0.60 (0.02)	0.69 (0.02)	0.10 (0.00)	0.02 (0.00)	0.00 (0.00)	0.00 (0.00)
	Observed	0.61 (0.03)	0.54 (0.04)	0.63 (0.03)	0.05 (0.00)	0.00 (0.00)	0.03 (0.00)	0.00 (0.00)
	ICYM^{2}I	0.66 (0.03)	0.57 (0.04)	0.62 (0.04)	0.07 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Memes	Oracle	0.71 (0.01)	0.57 (0.01)	0.72 (0.01)	0.09 (0.01)	0.00 (0.00)	0.04 (0.00)	0.05 (0.01)
	Observed	0.68 (0.02)	0.61 (0.02)	0.71 (0.02)	0.13 (0.00)	0.04 (0.00)	0.01 (0.00)	0.00 (0.00)
	ICYM^{2}I	0.67 (0.02)	0.61 (0.02)	0.71 (0.02)	0.10 (0.00)	0.01 (0.00)	0.02 (0.03)	0.03 (0.01)

- We induce 70% missingness on the image wmodality with UR-FUNNY (Hasan et al., 2019) and Hateful Memes (Kiela et al., 2020).
- Boxed regions demonstrate the ability of ICYM^{2}I-Learn and ICYM^{2}I-PID to correct the bias.

Real World Clinical Example

Table 4. Informativeness of ECG and CXR modalities on model-based structural heart disease detection. Parentheses denote standard deviation across batches ($n = 1024$).

	AUROC			Information Decomposition			
	ECG	CXR	ECG + CXR	$\text{Unique}_{\text{ECG}}$	$\text{Unique}_{\text{CXR}}$	Shared	Complementary
Observed	0.83 (0.01)	0.72 (0.02)	0.82 (0.01)	0.11 (0.00)	0.01 (0.00)	0.10 (0.00)	0.00 (0.00)
ICYM^2I	0.82 (0.01)	0.73 (0.02)	0.83 (0.01)	0.07 (0.00)	0.01 (0.00)	0.48 (0.00)	0.01 (0.00)

- We collected real-world clinical data of electrocardiograms (ECGs), chest X-rays (CXRs) and associated structural heart diesease labels (SHD).
- The **corrected PID values suggest that CXRs are not independently useful for this task.**
- Note that since this is real-world data we do not report oracle model metrics.

Limitations

- ICYM^{2}I is designed for settings where **modalities are paired** (e.g. a patient and their associated medical exams) and does not extend to unpaired settings.
- Our method is designed for Missing at Random (MAR) missingness and is stable under missing completely at random (MCAR). However, identifying and addressing modality missingness under **Missing not at Random (MNAR) remains an open question**.
- The Partial Information Decomposition (PID) bounds **only cover two input modalities**. Consequently, for practitioners will have to utilize a one-versus-rest to perform ICYM^{2}I-PID with three or more input modalities.

References

Rubin, Donald B. "Inference and missing data." *Biometrika* 63.3 (1976): 581-592.

Williams, Paul L., and Randall D. Beer. "Nonnegative decomposition of multivariate information." *arXiv preprint arXiv:1004.2515* (2010).

Bertschinger, Nils, et al. "Quantifying unique information." *Entropy* 16.4 (2014): 2161-2183.

Liang, Paul Pu, et al. "Quantifying & modeling multimodal interactions: An information decomposition framework." *Advances in Neural Information Processing Systems* 36 (2023): 27351-27393.

Hasan, Md Kamrul, et al. "UR-FUNNY: A multimodal language dataset for understanding humor." *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2019.

Kiela, Douwe, et al. "The hateful memes challenge: Detecting hate speech in multimodal memes." *Advances in neural information processing systems* 33 (2020): 2611-2624.

Paper Link

Conclusion and Takeaways

- Assuming **stable missingness** between source and target distributions can lead to **suboptimal modeling decisions**.
- We **formalize the issue of missing modalities** in multimodal settings.
- **ICYM^{2}I** is a framework for the unbiased estimation of **predictive performance** and **information gain** under modality missingness.
- Our method is **model-agnostic** and requires only data and binary missingness labels to train the propensity models.

Please visit our poster on Thursday, April 23 at 2:15 PM – 4:45 PM!

"ICYM^{2}I: The illusion of multimodal informativeness under missingness"

Authors:
Young Sang Choi
Vincent Jeanselme
Pierre Elias
Shalmali Joshi

Contact:
young.sang.choi@columbia.edu

COLUMBIA | COLUMBIA UNIVERSITY IRVING MEDICAL CENTER