Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data

袁至誠
Chih-cheng Rex Yuan
rexyuan.com

Institute of Information Science, Academia Sinica

Tuesday 3rd June, 2025

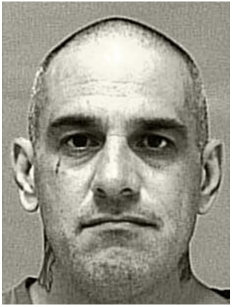
Why Fairness Audits Matter

- AI systems are influencing justice, health, finance.
- Bias in AI means real-world discrimination.
- Example: COMPAS audit by ProPublica.

COMPAS

recidivism *noun*

the tendency of a convicted criminal to reoffend.



- Prior Offenses: 2 armed robberies, 1 attempted armed robbery
- Subsequent Offenses: 1 grand theft
- Risk Score: 3

- Prior Offenses: 4 juvenile misdemeanors
- Subsequent Offenses: None
- Risk Score: 8

COMPAS

- COMPAS is an algorithm used by U.S. courts for predicting recidivism based on a questionaire.
- In 2016, ProPublica found that the algorithm is biased.

  > *Black defendants were often predicted to be at a higher risk of recidivism than they actually were. White defendants were often predicted to be less risky than they were.*

- The false-positive rates vary significantly across black people and white people, violating *equalized odds*.

[1] (Link) ProPublica - How We Analyzed the COMPAS Recidivism Algorithm
[2] (Link) Vsauce2 - The Dangerous Math Used To Predict Criminals

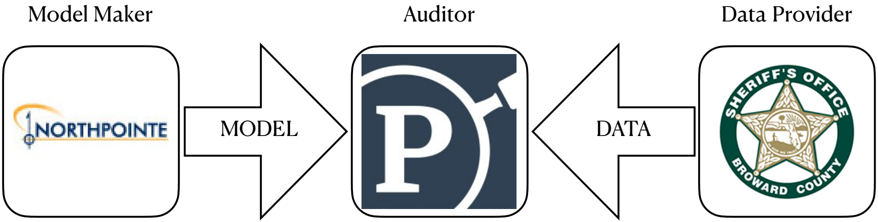
Fairness Metrics: Equalized Odds

- Equalized odds measures whether a model's error rates are balanced across groups.
- In other words, it asks: *Does the model make mistakes equally, regardless of group membership?*
- Formally, it requires that

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \epsilon$$
$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \epsilon$$

  where $\hat{Y}$ is the predicted outcome, Y is the ground truth, and S is the sensitive attribute.
- It ensures that both false positive and true positive rates are similar across groups.

Auditing Framework



The Privacy Problem

The COMPAS audit was possible only because the U.S. law made the data publicly accessible. Otherwise, such audits would be infeasible.

- Auditors need sensitive data to test fairness.
- But holding this data introduces security and privacy risks.
- Security risk: hackers could steal the data.
  Example: (Link) 23andMe data breach.
- Privacy risk: published stats could expose insights.
  Example: (Link) All of Us inference attack.

Solution: Privacy-Preserving Audits

- Use synthetic data generated from real data.
- Apply differentially private synthetic data generation techniques to ensure individual info stays hidden.
- Auditors only keep the synthetic data and not the real data.

Data Marginals

Original Data

ID	Sex	Race
1	Male	White
2	Female	White
3	Female	Black
4	Male	Black
5	Female	Black
6	Female	White

(Sex,Race) Marginal

Sex	Race	Count
Male	White	1
Female	White	2
Male	Black	1
Female	Black	2

Differentially Private Synthetic Data

- Fake data that mimic the real data's patterns.
- Generated with differential privacy, which adds noise to marginals to protect privacy.
- Lets auditors analyze fairness without exposing personal data.

Noising Marginal

Original Marginal

Sex	Race	Count
Male	White	1
Female	White	2
Male	Black	1
Female	Black	2

Noised Marginal

Sex	Race	Count
Male	White	0.8
Female	White	2.1
Male	Black	0.4
Female	Black	1.9

How The Synthetic Data Are Generated

- We use a proven method that won a U.S. government competition: (Link) NIST 2018.
- It adds noise to protect privacy while preserving overall patterns.
- The result: data that looks real but contains no real individuals.

Synthetic Data

Original Data

ID	Sex	Race
1	Male	White
2	Female	White
3	Female	Black
4	Male	Black
5	Female	Black
6	Female	White

Synthetic Data

ID	Sex	Race
1	Male	White
2	Female	Black
3	Female	Black
4	Male	White
5	Female	White
6	Female	Black

Does It Work?

- We experimented on real datsets: Adult, COMPAS, Diabetes.
- We compared fairness metrics of original and synthetic data.
- Most metrics are within negligible difference.

Equalized Odds on COMPAS: Original vs. Synthetic

Metric	Original	Synthetic	Difference
Equalized Odds (False Positive)	0.0249	0.0802	0.0553
Equalized Odds (True Positive)	0.0177	0.0825	0.0648

Limitations

- Results are not guaranteed to be accurate.
- Complex patterns may be missed.
- Privacy vs precision trade-off.

Policy Implications

- Enables safer third-party audits under privacy guarantees.
- Avoids liability for storing sensitive datasets.

Conclusion

- Synthetic data can support fairness audits with privacy.
- Our framework is practical and provably private.
- Opens new pathways for legal oversight of AI systems.

Slides: `https://github.com/RexYuan/Eunectes`.