MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

# Uncertainty in breast cancer risk prediction: A conformal prediction study of race stratification


Julio C. Facelli

Distinguished Professor

*Department of Biomedical Informatics, The University of Utah*

*Julio.facelli@utah.edu*




AIDH IMIA

# Acknowledgments

•My co-authors: Alexander S. Millar, John Arnn, Sam Himes

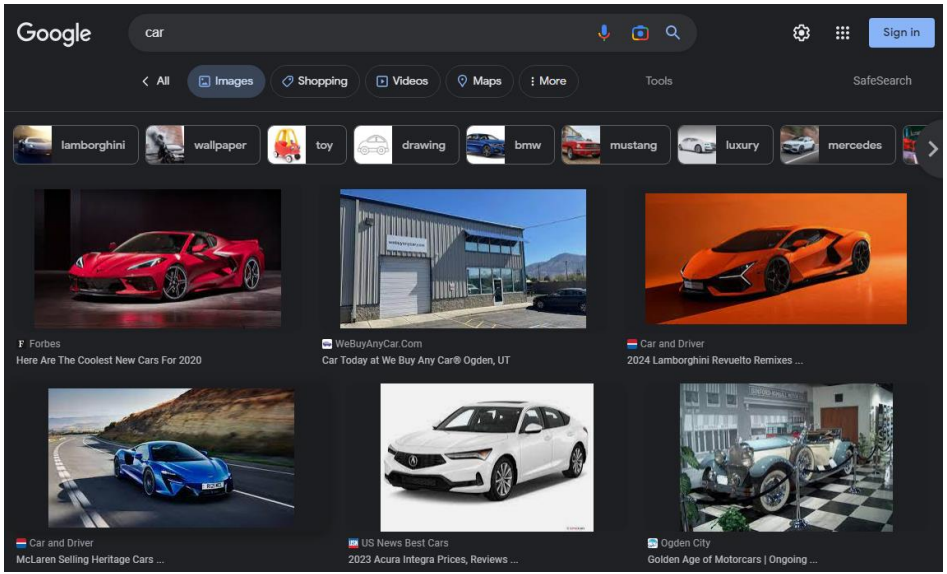
- This work has been partially supported by National Center for Advancing Translational Sciences (UL1TR002538). Computer resources were provided by the University of Utah Center for High-Performance Computing, which has been partially funded by the NIH Shared Instrumentation Grant 1S10OD02164401A1.

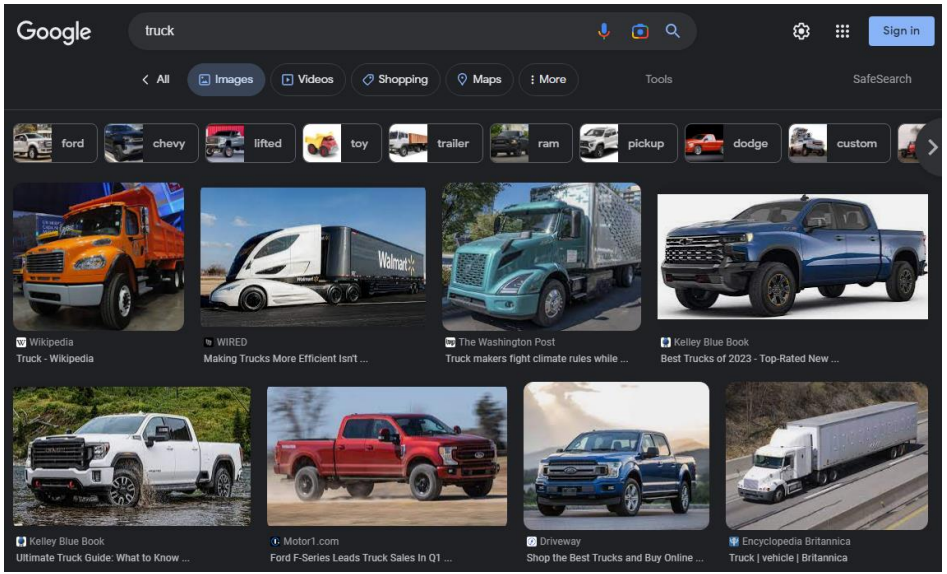
# Background

- **Predictions**
  - Inherent Uncertainty
    - Uncertainty in the data (i.e., first BP read 145/90, second BP read 135/85)
    - Data discretization (i.e., 129/79 is normal BP, but 130/80 is not)
    - Models can not fit all data equally well and disregard (by design) outliers
  - Bias
    - The data represents a subset of the “real space”
    - The represented sub-set is conditioned availability of data (particularly true when reusing data)
    - The data distributions are known a priori, so there always may be outliers that were not considered

# MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA




- https://www.google.com/search?q=truck&tbm=isch&ved=2ahUKEwjS_8y42qz-AhVuMUQIHdJbD8MQ2-cCegQIABAA&oq=truck&gs_lcp=CgNpbWcQAzIKCAAQigUQsQMQQzIKCAAQigUQsQMQQzIKCAAQigUQsQMQQzIKCAAQigUQsQMQQzIKCAAQigUQQzIFCAAQgAQyCggAEIoFELEDEEMyBwgAEIoFEEMyCAgAEIAEELEDUIQIWMIXYJMZaABwAHgAgAE1iAGaApIBATaYAQCgAQGqAQtnd3Mtd2l6LWltZ8ABAQ&sclient=img&ei=cQY7ZJKDJu7ikPIP0re9mAw&bih=1329&biw=2560&rlz=1C1CHZN_enUS948US948


Source: https://www.google.com/search?q=car&tbm=isch&ved=2ahUKEwjY_4Op2qz-AhUFPkQIHd3UDFQQ2-cCegQIABAA&oq=car&gs_lcp=CgNpbWcQAzIKCAAQigUQsQMQQzIHCAAQigUQQzIHCAAQigUQQzIHCAAQigUQQzIHCAAQigUQQzIHCAAQigUQQzIHCAAQigUQQzIHCAAQigUQQzoFCAAQgAQ6CAgAEIAEELEDUOICWJIFYMlPaABwAHgAgAEwiAGpAZIBATSYAQCgAQGqAQtnd3Mtd2I6LWltZ8ABAQ&sclient=img&ei=UAY7ZNiXO4X8kPIP3amzoAU&bih=1329&biw=2560&rlz=1C1CHZN_enUS948US948

# Conformal Prediction

"Conformal prediction can be used with any method of point prediction for classification or regression, including support-vector machines, decision trees, boosting, neural networks, and Bayesian prediction. Starting from the method for point prediction, we construct a nonconformity measure, which measures how unusual an example looks relative to previous examples, and the conformal algorithm turns this nonconformity measure into prediction regions."

https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf

# BCRAT and BOADICEA Models



**Baseline-risk function approximates NCI algorithm**

- 0.3*T1 (age) +
- 0.2*N_Biop (# of biopsies) +
- 0.1*N_Rels (# first-degree relatives with breast cancer) +
- 0.1*AgeMen (age at menarche) +
- 0.3*rnorm(20,3) (random noise)

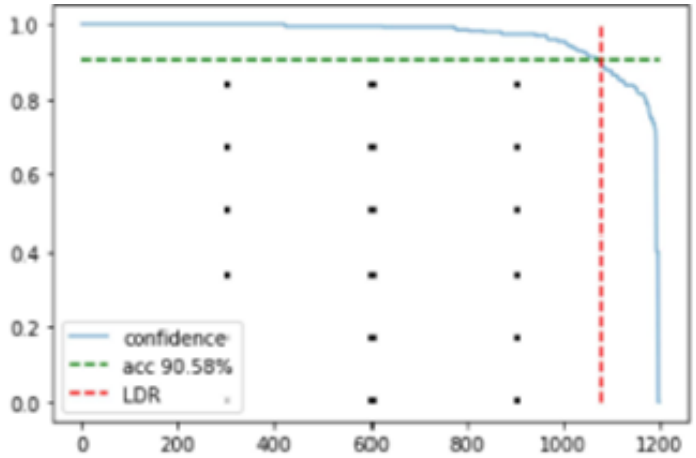
# Model Race-Differential Risk

- **Baseline Risk**
  - 0.3*T1 + 0.2*N_Biop + 0.1*N_Rels + 0.1*AgeMen + 0.3*rnorm(20,3)

  **+**
- **Race-Differential Risk**
  - Baseline Risk * Risk-Scaling Factor

| Race | Percentage of the Population | Risk Scaling Factor |
|---|---|---|
| 0 | 0.50 | -0.0025 |
| 1 | 0.20 | 0.00375 |
| 2 | 0.20 | -0.00125 |
| 3 | 0.08 | 0.00125 |
| 4 | 0.015 | 0.00875 |
| 5 | 0.003 | 0.0175 |
| 6 | 0.002 | 0.005 |

**Figure 1.** Prediction confidence distribution for the two data sets used in this study.



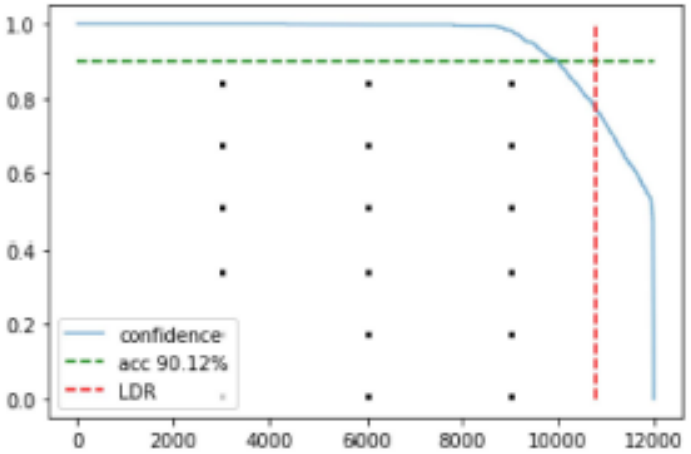



**Figure 2.** Comparison of the distribution of **race groups in the population (blue)** and the corresponding prevalence among those in the **LDR of the confidence distribution (orange).**

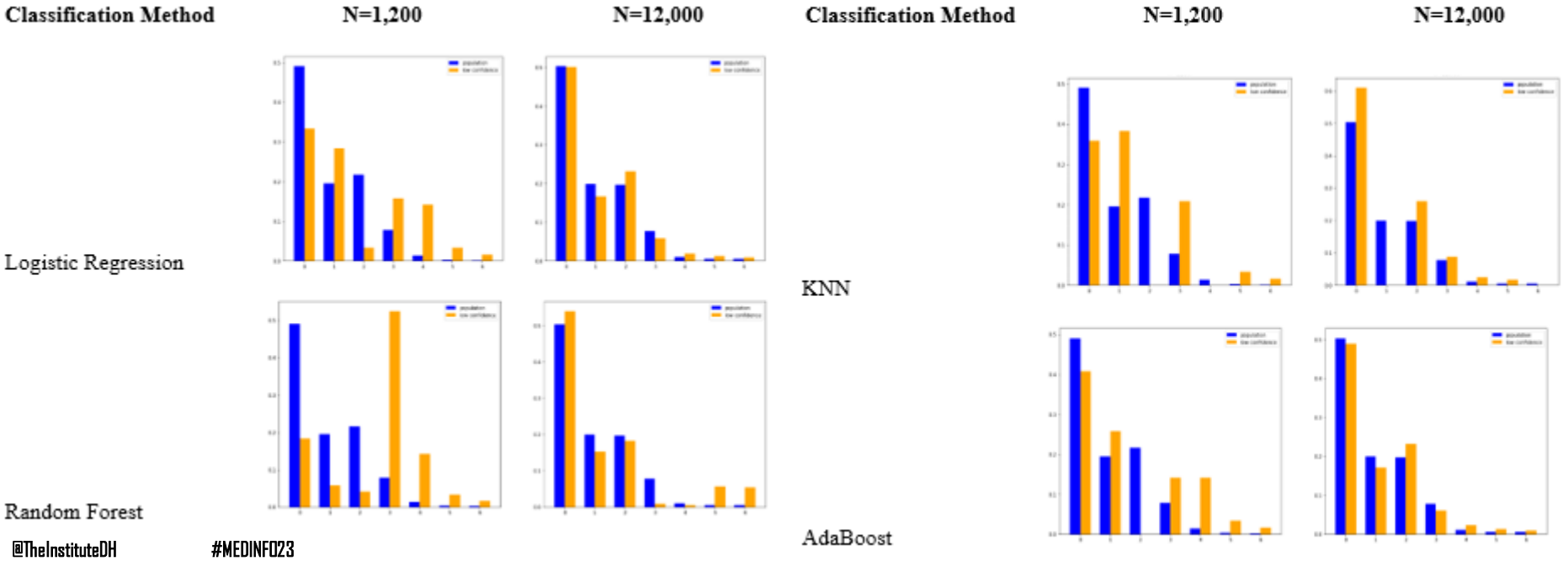


Figure 7, inspired by [6, 7], presents individual predictions (corresponding to a single individual) made by an ensemble of Conformal Predictors. The credibility threshold (upper red line) corresponds to the 75th percentile among Conformal Predictors' maximum credibility value in their respective low-credibility regions (LDR). Similarly, the confidence threshold (lower red line) corresponds to the 75th percentile among Conformal Predictors' maximum confidence value in their respective low-confidence regions (LDR). In fact, Figure 6 corresponds to the first (far left) Conformal Predictor's prediction and Figure 5 corresponds to the second Conformal Predictor's prediction.

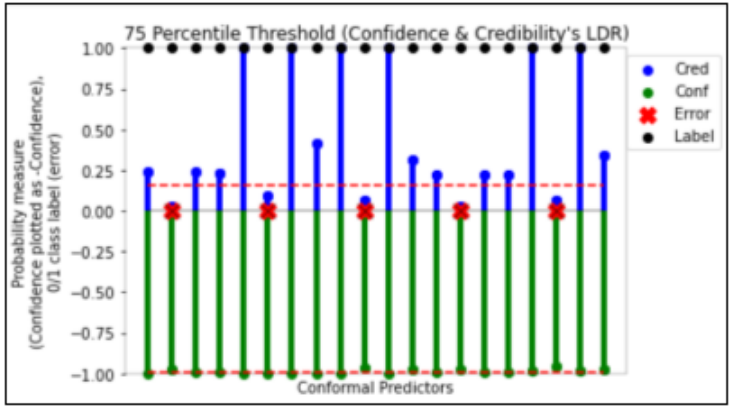



**Figure 7.** Individual Prediction – Multiple Conformal Predictors.

Figure 5 presents an individual prediction where the prediction does not meet one or more threshold criteria. In this case, the prediction credibility (red bar) is below the credibility threshold (red region). Figure 6 presents an individual prediction where the prediction credibility (green bar) exceeds the credibility threshold (reg region), the prediction confidence (green gauge) exceeds the confidence threshold (grey region), and the specified acceptable error (eps., 0.1) produces a non-empty prediction set.

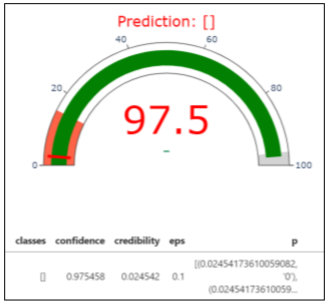



**Figure 5.** Individual Prediction Indicator – Below Threshold.

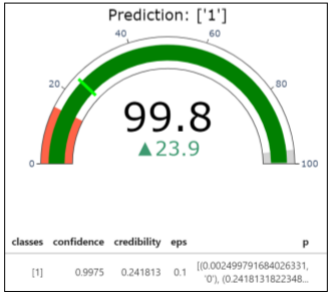



**Figure 6.** Individual Prediction Indicator – Above Threshold.

# Conclusions

- The method presented here show that CP can be used effectively to find the race stratification of individuals in the LDR of a predictive model.
- The results presented here cannot be used for clinical applications because the race-differential risks were arbitrarily assigned for the purpose of these demonstrations, but the methods demonstrated here can be used with any risk model derived from ML approaches.
- We also show that the undesirable stratification associating larger uncertainties to minority races can be corrected when using larger overall samples.
- Future work will include investigations to better understand the sensitivity of under/over representation of races in the LDR; classes upon their differential risks; and to find out if increased target, instead of overall, sample size could remediate stratification with a smaller enlargement of the sample size.