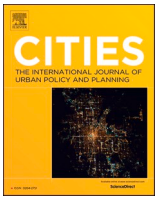
Cities 166 (2025) 106207

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities



Assessing demographic influences on recidivism: A comprehensive analysis through chi-squared tests and machine learning techniques☆

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo 135-8181, Japan

ARTICLE INFO

Keywords:
Chi-squared test
Machine learning
Feature importances
Variable associations
Recidivism analysis

ABSTRACT

This study examines demographic factors influencing recidivism within the context of global urban transitions. Analyzing data from the U.S., where recidivism costs 41 states over $8 billion annually, we employed Chi-squared tests and random forest analyses to identify significant predictors. Our findings reveal that geographic location—particularly in urbanizing areas—and age significantly impact recidivism rates, while gender has minimal influence. These patterns likely extend to both developing and developed nations experiencing similar demographic shifts. Our methodological comparison between statistical and machine learning approaches provides a transferable framework for international researchers. The study contributes novel insights by demonstrating how urbanization's economic, social, and infrastructural dynamics affect reoffending patterns across diverse contexts. We recommend policymakers worldwide implement location-specific and age-targeted interventions while establishing collaborative platforms to share evidence-based strategies, addressing recidivism's universal challenges regardless of national boundaries.

1. Introduction

Many are unaware of the immense financial burden that recidivism places on our communities. According to the Council of State Governments (CSG justice center, n.d.), in 2021 alone, 41 states spent over $8 billion incarcerating more than 19,300 individuals for supervision violations and revocations. In ten of those states, the annual cost of recidivism exceeded $40 per resident. This paper examines the influence of demographic factors on these costs, utilizing statistical analysis and machine learning techniques to assess feature importance.

Researchers in social science and urban governance frequently lack comprehensive understanding of machine learning principles. This knowledge gap significantly impacts their work, particularly regarding the interpretation of data-driven insights informing policy decisions. While machine learning in urban governance aims to predict outcomes based on established indicators, feature importances attempt to clarify relationships between outcomes and features that often lack direct ground truth validation.

A persistent misconception suggests that feature importances derived from machine learning models represent unbiased indicators of genuine associations within urban datasets. This misunderstanding can lead to erroneous conclusions, potentially distorting policy outcomes with detrimental effects on urban planning and community welfare. Over 100 peer-reviewed articles have documented significant biases associated with feature importances across various domains including urban studies (Takefuji, n.d.; Takefuji, 2024a; Takefuji, 2024b; Takefuji, 2024c; Takefuji, 2024d; Takefuji, 2024e; Takefuji, 2024f; Takefuji, 2025a; Takefuji, 2025b; Takefuji, 2025c).

Despite numerous mitigation strategies addressing these biases, no single method completely eliminates bias in feature importances. Each approach provides limited bias reduction, requiring urban researchers and policymakers to remain vigilant and employ multiple strategies to enhance finding reliability. This challenge highlights the complexity of accurately capturing true associations within urban data landscapes, necessitating thoughtful integration of machine learning with traditional statistical methodologies.

Different machine learning models employ diverse methodologies for calculating feature importance, introducing varying degrees of bias. Some models prioritize highly correlated features while others disregard them, potentially resulting in misleading interpretations. Consequently,

☆ **According to ScholarGPS**, Yoshiyasu Takefuji holds notable global rankings in several fields. He ranks 54th out of 395,884 scholars in neural networks (AI), 23rd out of 47,799 in parallel computing, and 14th out of 7,222 in parallel algorithms. Furthermore, he ranks highest in AI tools and human-induced error analysis, underscoring his significant contributions to these domains.

E-mail address: takefuji@keio.jp.

https://doi.org/10.1016/j.cities.2025.106207
Received 13 April 2025; Received in revised form 16 June 2025; Accepted 19 June 2025

researchers may draw erroneous inferences regarding data relationships, compromising finding validity. Addressing this issue requires deeper understanding of machine learning techniques and their implications.

To accurately calculate associations between target variables and features, three essential elements must be considered: data distribution, examination of statistical relationships among variables, and assessment of statistical validity through *p*-values. This paper addresses concerns with over-reliance on feature importances derived from machine learning models and advocates for bias-free rigorous statistical methods such as Chi-squared tests with *p*-values to uncover true associations.

Recidivism is influenced by various factors in urban ecosystems. Cities offer job opportunities and educational resources often inaccessible due to stigma and insufficient support, leading to economic instability—a key recidivism predictor. Strong social support networks can reduce reoffending, yet weakened ties in densely populated areas may contribute to higher rates. Stable housing remains crucial, but high rent and shortages present significant challenges. Although urban areas provide more resources for mental health and substance abuse treatment, barriers including cost and stigma frequently impede access.

This paper examines relationships between recidivism and independent variables including gender, county, and age using two analytical approaches: Chi-squared tests with *p*-values and feature importances derived from random forests. While machine learning focuses on prediction accuracy, feature importances aim to represent associations between target and independent variables. However, model-specific characteristics cause feature importances to vary significantly between models, introducing potential bias.

Feature importances in machine learning are model-specific because they derive from the internal workings of particular models (Saarela and Jauhiainen, 2021). Different models calculate importance through unique methods, leading to variations in scores assigned to identical features. Decision trees calculate importance based on impurity reduction, while linear models use feature coefficients. Models like random forests capture complex interactions affecting importance scores (Cava et al., 2020).

These scores can be influenced by model bias and variance (Michelucci, 2024) and don't necessarily reflect causal relationships but rather how features contribute to predictions. They reflect correlations rather than causation, with confounding variables potentially distorting importance scores. Different models make different assumptions about data, affecting importance scores and making them specific to models rather than reflecting true associations (Michelucci, 2024; Theng and Bhoyar, 2024). Although feature importances have been used in numerous studies (Erion et al., 2021; Nichols et al., 2024; Slack et al., 2023; Tang et al., 2024; Wan et al., 2024; Ziegenfeuter et al., 2024), they don't represent true associations between variables.

True associations between variables can be computed using Chi-squared tests or similar methods (Ko et al., 2024; McCarthy et al., 2008). The Chi-squared test determines significant associations between categorical variables by comparing observed frequencies with expected frequencies, identifying meaningful deviations and assessing variable independence. When testing shows significance, one variable's value depends on another's. Designed specifically for categorical data, the test provides *p*-values indicating statistical significance, with low values suggesting strong associations.

This paper investigates associations between recidivism (y) and features (X = (x1, x2, x3)): county, gender, and age. Chi-squared tests quantify these associations with *p*-values, evaluating statistical significance between the dependent variable and each predictor. Additionally, random forest algorithms generate feature importances for all variables. This dual approach combines statistical significance with machine learning insights for comprehensive recidivism analysis, while acknowledging that no universal tool exists to calculate true variable associations.

2. Methods

This paper presents a concise analysis of recidivism trends using both the Chi-squared test and Random Forest Classification. The analysis is conducted based on a dataset, Recidivism_Beginning_2008.csv, from the State of New York, released on July 26, 2024, which comprises 287,139 instances and 5 variables which may be the largest dataset on recidivism in the world (Data.Gov, n.d.). This study aims to identify significant patterns and predictors of recidivism, providing valuable insights for policymakers and researchers. The dependent variable, 'Return Status' of recidivism, is selected as the target, while 'County', 'Gender', and 'Age' are considered independent variables. This paper analyzes the associations between the target and these independent variables.

Generative AI is used and demonstrated to assist novices and non-programmers in generating Python code. Due to the inherent imperfections of generative AI, multiple interactions and user verification are necessary to achieve successful and desired outcomes. Before crafting queries, users should be familiar with the dataset and variables. The following initial inputs (queries) are provided to the Copilot of generative AI: single-quoted strings represent variables, while double-quoted strings indicate values. The final Python code was verified by experts to ensure the outcomes and conclusions.

Chi-squared query: use Recidivism_Beginning_2008.csv file. 'Release Year' indicates year. 'County of Indictment' indicates county names. Remove "UNKNOWN" from 'County of Indictment' values. 'Gender' indicates "MALE" or "FEMALE". 'Age at Release' indicates age values. 'Return Status' indicates outcomes in strings. Calculate chi-squared and *p*-value of 3 associations between 'Return Status' and 'County of Indictment', that between 'Return Status' and 'Gender', and that between 'Return Status' and 'Age at Release' for individual years from 2008 to 2020. Plot the trends of 3 black lines of chi-squared and *p*-value with 4 linestyles and 2 widths (1,2). The graph should have 6 distinct lines with rotating xticks with 90°. Show Python full code.

Random forest query: use Recidivism_Beginning_2008.csv file. 'Release Year' indicates year. 'County of Indictment' indicates county names. Remove "UNKNOWN" from 'County of Indictment' values. 'Gender' indicates "MALE" or "FEMALE". 'Age at Release' indicates age values. 'Return Status' indicates outcomes in strings. Use randomforest model where y is 'Return Status' and X are 'Release Year', 'County of Indictment', 'Gender' and 'Age at Release'. Calculate Chi-squared statistics, *p*-values and accuracies for individual years from 2008 to 2020. Plot a graph of three values such as Chi-squared value, p-value and accuracy from 2008 to 2020 where right Y-axis indicates Chi-squared values and accuracies while left Y-axis indicates *p*-values. Calculate feature importances of year, county, gender and age to influence

Fig. 1. Trends of Chi-squared and *p*-values.

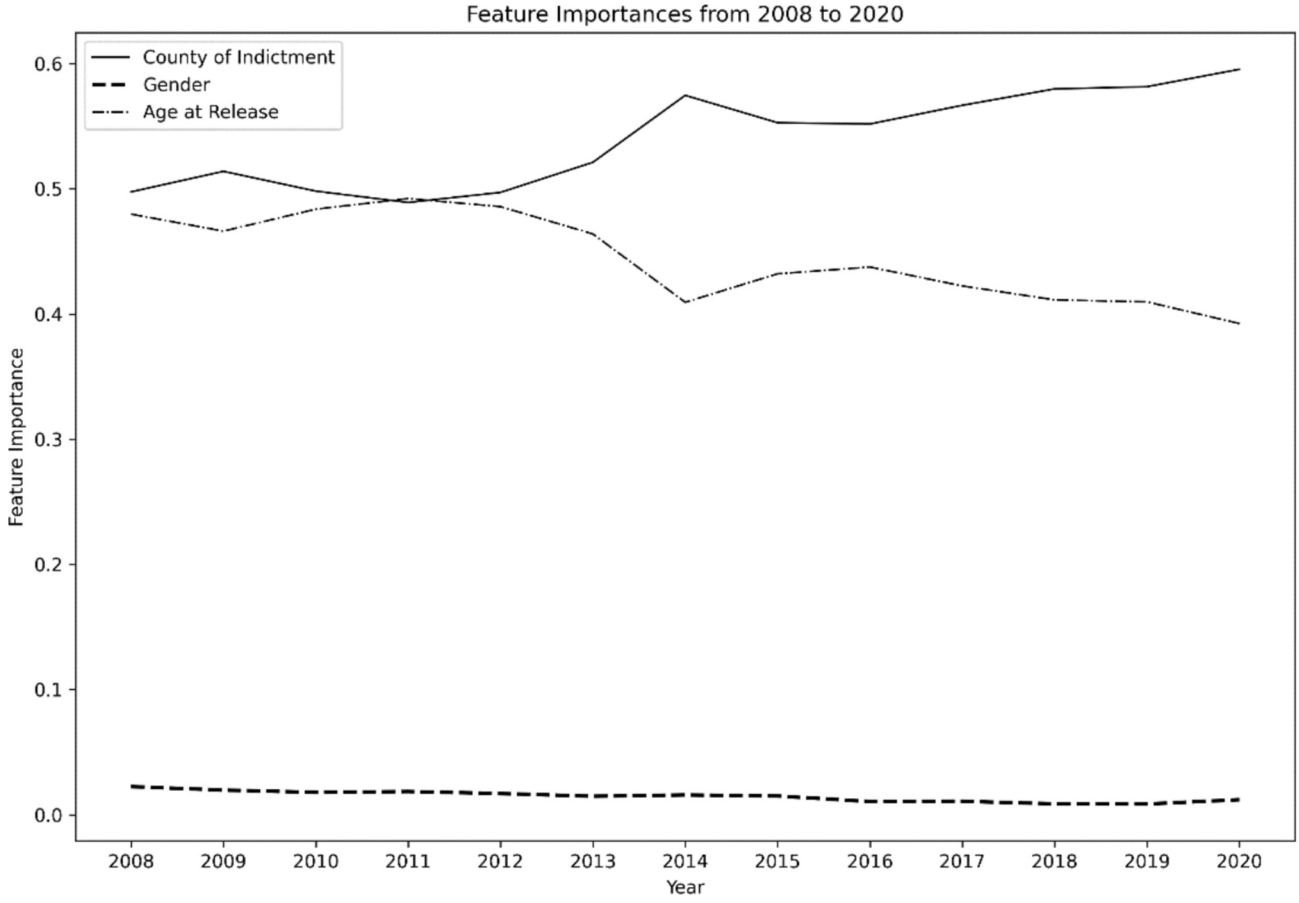


Fig. 2. Feature importances with random forest algorithms.

outcomes. Draw a graph of distributions of 4 feature importances from 2008 to 2020 with 4 linestyles. Show python full code.

3. Results

Download the comma-separated values (CSV) file, referred to as the dataset, from the data.gov website (Data.Gov, n.d.). The Python scripts, chi.py and rf_fimportances.py, developed using generative AI, are available for public access on GitHub to ensure reproducibility, and additional details can be found in the Appendix (GitHub, n.d.). After downloading the dataset and programs, run these scripts to generate the desired results. ($) indicates the prompt from the system terminal.

$ python chi.py

$ python rf_fimportances.py

Fig. 1 illustrates the results of the chi.py analysis, while Fig. 2 presents the findings from the rf_fimportances.py analysis. In Fig. 1, the *p*-values for county, gender, and age are all zeros, indicating statistical significance. In Fig. 1, the associations for county and age hovered around the range of 400 to 600 until 2014 with twice intersected. Post-2014, the county association showed a marked increase, peaking in 2018 and then stabilizing. Conversely, the age association exhibited a decline from 2013 to 2019, with a slight uptick in 2020.

In Fig. 2, the county association consistently remained higher than the age association from 2008 to 2020, with the exception of 2011, where the age association briefly surpassed that of the county. These trends highlight significant differences in the predictive power of county and age over the years. Differences in Fig. 1 and Fig. 2 highlight biases introduced by machine learning.

4. Discussion

The results from Fig. 1 and Fig. 2 provide complementary insights into factors influencing recidivism globally. Fig. 1, based on chi-squared analysis, reveals gender has nearly zero association with recidivism rates—a finding consistent with international studies (Ko et al., 2024; McCarthy et al., 2008). The increasing chi-squared values for county post-2014 indicate that geographical location has become more influential, reflecting differences in urbanization patterns, socioeconomic conditions, and rehabilitation infrastructure that transcend national boundaries.

Fig. 2, based on random forest feature importances, confirms these patterns while highlighting methodological considerations relevant to researchers worldwide. The consistently higher feature importance of county compared to age reinforces that geographical factors are universally significant predictors. This methodological comparison demonstrates how different analytical approaches can yield complementary insights applicable across diverse contexts.

The implications extend beyond local policy to international practice. Our finding that geographical location significantly predicts recidivism aligns with global urbanization research, suggesting that rapid urban transitions create similar challenges for criminal justice systems worldwide. Policymakers internationally should implement place-based interventions that address specific urban-rural disparities in recidivism rates, regardless of national context.

Similarly, our age-related findings suggest universal life-course patterns in criminal behavior that require targeted interventions. We recommend that correctional systems globally develop age-appropriate rehabilitation programs while maintaining gender-neutral approaches, as our findings indicate gender has minimal impact on recidivism across contexts.

For researchers, our methodological comparison between statistical and machine learning approaches offers a transferable framework that can be adapted to different national datasets. We caution that machine learning models generate biased feature importances due to their model-specific nature (Lakens, 2022), highlighting the need for methodological triangulation in international recidivism research.

Future cross-national research should explore how urbanization processes specifically influence recidivism patterns across countries at different development stages. We recommend establishing international data-sharing protocols and standardized metrics to facilitate comparative analyses. Additionally, policymakers should create forums for knowledge exchange regarding successful place-based and age-targeted interventions, fostering global collaboration to address the universal challenge of reducing recidivism through evidence-based approaches.

CRediT authorship contribution statement

Yoshiyasu Takefuji: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Not applicable.

Declaration of Generative AI and AI-assisted technologies in the writing process

Not applicable.

Funding

This research has no fund.

Declaration of competing interest

The author has no conflict of interest.

Appendix A. Appendix A

APPENDIX: chi.py

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency

# Load the dataset
df = pd.read_csv('Recidivism__Beginning_2008.csv')

# Remove "UNKNOWN" from 'County of Indictment'
df = df[df['County of Indictment'] != 'UNKNOWN']

# Initialize lists to store chi-squared and p-values
chi2_county = []
p_county = []
chi2_gender = []
p_gender = []
chi2_age = []
p_age = []

# Loop through each year from 2008 to 2020
for year in range(2008, 2021):
    df_year = df[df['Release Year'] == year]

    # Chi-squared test for 'Return Status' and 'County of Indictment'
    contingency_table_county = pd.crosstab(df_year['Return Status'], df_year['County of
Indictment'])
    chi2, p, _, _ = chi2_contingency(contingency_table_county)
    chi2_county.append(chi2)
    p_county.append(p)

    # Chi-squared test for 'Return Status' and 'Gender'
    contingency_table_gender = pd.crosstab(df_year['Return Status'], df_year['Gender'])
    chi2, p, _, _ = chi2_contingency(contingency_table_gender)
    chi2_gender.append(chi2)
    p_gender.append(p)

    # Chi-squared test for 'Return Status' and 'Age at Release'
    contingency_table_age = pd.crosstab(df_year['Return Status'], df_year['Age at Release'])
    chi2, p, _, _ = chi2_contingency(contingency_table_age)
    chi2_age.append(chi2)
    p_age.append(p)

# Plot the trends

years = list(range(2008, 2021))

fig, ax1 = plt.subplots(figsize=(14, 8))

# Chi-squared values on the left Y-axis
ax1.plot(years, chi2_county, label='Chi-squared (County)', linestyle='-', linewidth=2,
color='black')
ax1.plot(years, chi2_gender, label='Chi-squared (Gender)', linestyle=':', linewidth=2,
color='black')
ax1.plot(years, chi2_age, label='Chi-squared (Age)', linestyle='--', linewidth=2,
color='black')
ax1.set_xlabel('Year')
ax1.set_ylabel('Chi-squared Values')
ax1.tick_params(axis='y')

# P-values on the right Y-axis
ax2 = ax1.twinx()
ax2.plot(years, p_county, label='P-value (County)', linestyle='-', linewidth=1, color='black')
ax2.plot(years, p_gender, label='P-value (Gender)', linestyle=':', linewidth=1, color='black')
ax2.plot(years, p_age, label='P-value (Age)', linestyle='--', linewidth=1, color='black')
ax2.axhline(y=0.05, color='red', linestyle=':', linewidth=2, label='Reference Line (0.05)')
ax2.set_ylabel('P-values')
ax2.tick_params(axis='y')

# Combine legends
lines, labels = ax1.get_legend_handles_labels()
lines2, labels2 = ax2.get_legend_handles_labels()
ax2.legend(lines + lines2, labels + labels2, loc='upper center', bbox_to_anchor=(0.5, -0.15),
ncol=3)

plt.title('Trends of Chi-squared and P-values (2008-2020)')
plt.xticks(years, rotation=90)
plt.grid(True)
plt.tight_layout()
plt.savefig('chi.png',dpi=300)
plt.show()
```

APPENDIX: rf_fimportances.py

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import chi2
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score

# Load the data
df = pd.read_csv('Recidivism__Beginning_2008.csv')

# Data preprocessing
df = df[df['County of Indictment'] != 'UNKNOWN']
df['Gender'] = df['Gender'].map({'MALE': 1, 'FEMALE': 0})

# Encode categorical variables
le_county = LabelEncoder()
df['County of Indictment'] = le_county.fit_transform(df['County of Indictment'])

le_return_status = LabelEncoder()
df['Return Status'] = le_return_status.fit_transform(df['Return Status'])


# Train RandomForest model
X = df[['County of Indictment', 'Gender', 'Age at Release']]
y = df['Return Status']
model = RandomForestClassifier()
model.fit(X, y)

# Calculate feature importances for each year
feature_importances = {feature: [] for feature in ['County of Indictment', 'Gender', 'Age at
Release']}
for year in range(2008, 2021):
    df_year = df[df['Release Year'] == year]
    X_year = df_year[['County of Indictment', 'Gender', 'Age at Release']]
    y_year = df_year['Return Status']
    model.fit(X_year, y_year)
    importances = model.feature_importances_
    for i, feature in enumerate(feature_importances.keys()):
        feature_importances[feature].append(importances[i])

# Plot feature importances
plt.figure(figsize=(12, 8))
linestyles = ['-', '--', '-.', ':']
widths = [1, 2, 1, 2]
for (feature, linestyle, width) in zip(feature_importances.keys(), linestyles, widths):
    plt.plot(range(2008, 2021), feature_importances[feature], linestyle=linestyle,
linewidth=width, label=feature, color='black')

plt.xlabel('Year')
plt.ylabel('Feature Importance')
plt.title('Feature Importances from 2008 to 2020')
plt.legend()
plt.xticks(range(2008, 2021))
plt.savefig('feature-importances.png',dpi=300)
plt.show()
```

Data availability

The authors do not have permission to share data.

References[1]

Cava, W., Bauer, C., Moore, J. H., & Pendergrass, S. A. (2020). *American Medical Informatics Association Annual Symposium Proceedings, 2019*, 572–581 (Published 2020 Mar 4).

*CSG justice center. The cost of recidivism Accessed on Feb. 13, 2025 https://csgjusticecenter.org/publications/the-cost-of-recidivism/.

Data.Gov. recidivism: beginning 2008. https://catalog.data.gov/dataset/recidivism-beginning-2008.

Erion, G., Janizek, J. D., Sturmfels, P., et al. (2021). *Nature Machine Intelligence, 3*, 620–631. https://doi.org/10.1038/s42256-021-00343-w

GitHub. chi.py and rf_fimportances.py. https://github.com/y-takefuji/recidivism.

Ko, B. S., Lee, S. B., & Kim, T. K. (2024). A brief guide to analyzing expression quantitative trait loci. *Molecules and Cells, 100*, 139. https://doi.org/10.1016/j.mocell.2024.100139

Lakens, D. (2022). *Collabra: Psychology, 8*, 1. https://doi.org/10.1525/collabra.33267

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., et al. (2008). *Nature Reviews. Genetics, 9* (5), 356–369. https://doi.org/10.1038/nrg2344

Michelucci, U. Springer, Cham. 2024. https://doi.org/10.1007/978-3-031-56431-4_10.

Nichols, E. S., Nelson, G., Wild, C. J., & Owen, A. M. (2024). *PLoS One, 19*(4), e0298899. Published 2024 Apr 16 https://doi.org/10.1371/journal.pone.0298899.

Saarela, M., & Jauhiainen, S. (2021). *SN Applied Sciences, 3*, 272. https://doi.org/10.1007/s42452-021-04148-9

Slack, D., Krishna, S., Lakkaraju, H., et al. (2023). *Nature Machine Intelligence, 5*, 873–883. https://doi.org/10.1038/s42256-023-00692-8

*Takefuji, Y. (2024a). Mitigating biases in feature selection and importance assessments in predictive models using LASSO regression. *Oral Oncology, 159*, Article 107090. https://doi.org/10.1016/j.oraloncology.2024.107090

*Takefuji, Y. (2024b). Unveiling feature importance biases in linear regression: Implications for protein-centric cardiovascular research. *Atherosclerosis.* , Article 119049. https://doi.org/10.1016/j.atherosclerosis.2024.119049

*Takefuji, Y. (2024c). Reassessing feature importance biases in machine learning models for infection analysis. *The Journal of Infection, 89*(6), Article 106357. https://doi.org/10.1016/j.jinf.2024.106357

*Takefuji, Y. (2024d). Reply to the editor. *Clinical Nutrition*. https://doi.org/10.1016/j.clnu.2024.11.031

*Takefuji, Y. (2024e). Evaluating feature importance biases in logistic regression: Recommendations for robust statistical methods. *European Journal of Internal Medicine*. https://doi.org/10.1016/j.ejim.2024.11.022

*Takefuji, Y. (2024f). Addressing feature importance biases in machine learning models for early diagnosis of type 1 Gaucher disease. *Journal of Clinical Epidemiology*. , Article 111619. https://doi.org/10.1016/j.jclinepi.2024.111619

*Takefuji, Y. (2025a). Model-specific feature importances: Distinguishing true associations from target-feature relationships. *Journal of Affective Disorders, 369*, 390–391. https://doi.org/10.1016/j.jad.2024.10.019

*Takefuji, Y. (2025b). Unveiling hidden biases in machine learning feature importance. *Journal of Energy Chemistry, 102*, 49–51. https://doi.org/10.1016/j.jechem.2024.10.032

*Takefuji, Y. (2025c). Reevaluating feature importances in machine learning models for schizophrenia and bipolar disorder: The need for true associations. *Brain, Behavior, and Immunity, 124*, 123–124. https://doi.org/10.1016/j.bbi.2024.11.036

*Takefuji Y. Chi-squared and P-values vs. machine learning feature selection. Annals of Oncology doi:https://doi.org/10.1016/j.annonc.2024.10.013.

Tang, A. S., Rankin, K. P., Cerono, G., et al. (2024). *Nature Aging, 4*(3), 379–395. https://doi.org/10.1038/s43587-024-00573-8

Theng, D., & Bhoyar, K. K. (2024). *Knowledge and Information Systems, 66*, 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Wan, Y. R., Koşaloğlu-Yalçın, Z., Peters, B., & Nielsen, M. (2024). *NAR Cancer, 6*(1), zcae002. Published 2024 Jan 29 https://doi.org/10.1093/narcan/zcae002.

Ziegenfeuter, J., Delbridge, C., Bernhardt, D., et al. (2024). *Eur J Nucl Med Mol Imaging*. https://doi.org/10.1007/s00259-024-06782-y. Published online June 5.

[1] References with * are the citations converted to AMA format.