Demographic Bias in Medical Datasets for Clinical AI

Bojana Velichkovska, Sandra Petrushevska, Bisera Runcheva and Marija Kalendar
Faculty of Electrical Engineering and Infomation Technologies, "Ss. Cyril and Methodius University" in Skopje, Rugjer Boshkovikj 18, Skopje, North Macedonia
{bojanav, kti1772019, kti1842019, marijaka}@feit.ukim.edu.mk

Keywords: Artificial Intelligence, Machine Learning, Gender Bias, Age Bias, Demographic Bias, Medical Datasets.

Abstract: Numerous studies have detailed instances of demographic bias in medical data and artificial intelligence (AI) systems used in medical setting. Moreover, these studies have also shown how these biases can significantly impact the access to and quality of care, as well as quality of life for patients belonging in certain under-represented groups. These groups are then being marginalised because of stigma based on demographic information such as race, gender, age, ability, and so on. Since the performance of AI models is highly dependent on the quality of data used to train the algorithms, it is a necessary precaution to analyse any potential bias inadvertently existent in the data, in order to mitigate the consequences of using biased data in creating medical AI systems. For that reason, we propose a machine learning (ML) analysis which receives patient biosignals as input information and analyses them for two types of demographic bias, namely gender and age bias. The analysis is performed using several ML algorithms (Logistic Regression, Decision Trees, Random Forest, and XGBoost). The trained models are evaluated with a holdout technique and by observing the confusion matrixes and the classification reports. The results show that the models are capable of detecting bias in data. This makes the proposed approach one way to identify bias in data, especially throughout the process of building AI-based medical systems. Consequently, the proposed pipeline can be used as a mitigation technique for bias analysis in data.

1 INTRODUCTION

There exit numerous factors which contribute to or exacerbate disparities in healthcare, as are implicit and explicit biases which imbibe discriminatory practices based on demographic information as race, ethnicity, gender, or age [1]. With biased practices preserving, patients can receive subpar care quality, which can range from delays in admission and poor treatment to inaccurate diagnosis and potential for worsened health conditions [2].

The impact of these issues largely affects underrepresented groups, and these (un)intended consequences even impede academic performance as medical professionals find themselves unable to treat certain populations. In example, dermatologists have spoken of their inability to accurately diagnose diseases in patients of colour due to under-representation of certain populations in medical textbooks [3]. Consequently, five-year melanoma survival estimations show the survival rate for Black patients is only 70% compared to the 94% for White patients [4]. Compared to the self-awareness of dermatologists, there is a different side to medical personnel, as shown by [5], where it is illustrated that physicians are significantly less likely to recommend bypass surgery for Black compared to White patients. The contributing factor in these decisions was physicians believing Black patients to be less educated and, therefore, less likely to adhere to necessary activity post-surgery.

Moreover, personnel biases extend to disability attitudes [6], with 83.6% of healthcare providers having a preference for able-bodied patients. Socioeconomic status is another aspect in which medicine is biased, and patients of lower status are likely to have worse self-reported health and at a risk of multimorbidity [7], in addition to having limited access to health care and being at a greater risk for substandard care [8].

The biases are not limited to preferences only, and extend to assumptions based on demographic information which personnel use when treating patients. The authors of [9] identify gender bias in patient-provider encounters and treatment decisions, with dichotomous depictions of "brave men" and "emotional women". The study also found that

physicians are likely to attribute woman's pain as a product of a mental health condition rather than as a physical condition. Medical personnel disregarding patients' conditions can lead at the very least to delays in diagnosis. One example is [10], which found that women wait longer on average for a diagnosis compared to men in 72% of cases. Worst case scenarios can result in increased risk of death, e.g., how lack of awareness of the impact of heart attacks on women contributes to higher rates of females dying from heart attacks [11]. Healthcare professionals are less likely to recommend older patients for invasive or aggressive procedures denoting the choice as a "compassionate" approach even if said decision impacts life quality and expectancy in these patients [12].

Despite efforts to address and mitigate biased practices, health inequities persist, and infinitely worse get propagated in medical datasets and AI models which impact large populations. An algorithm was found to be racially biased since it used medical costs as a proxy for care needed, and consequently assigned the same level of risk to Black and White patients, even though the Black patients were in a worse medical condition [13]. A study of an algorithm for abnormalities in chest X-rays showed that highest rate of underdiagnosis exists in young females [14]. Another algorithm, which aimed to help with in-home care for patients, was found to recommend extreme cuts in cases of disabled patients, resulting in reduced quality of life and increased hospitalisation [15].

As integration of AI in medical systems is expected to increase in the upcoming years, it is necessary to address and resolve biased issues in order to limit negative impact, as well as understand where the bias originates in order to reduce the chances of propagating said bias into production stages, and thus, mitigation strategies will be necessary. Previous examples demonstrate that one potential source of bias for AI models can be the data used for the research and its distributions, as shown in [16] where the authors show the impact of gender imbalance in medical imaging datasets in computer-aided diagnostic tools. Additionally, the data used is the driving force for the algorithms, as they extrapolate information from said data in order to understand the problem and arrive at a decision.

Since the basic foundation for AI systems is the data, we wanted to investigate whether data bias is visible and easily discernable by the algorithms even when confounding variables are excluded from the training data. That is to say, we investigate whether potential biased issues can be detected with simple analysis of the data itself. However, as all data can be a subject to bias, medical datasets are not excluded from the influences of biased medical personnel or biased decisions in real practice. Moreover, even though it is necessary for developers to thoroughly investigate trained models before their active use, in many cases hidden (or implicit) biases are not observed before models are deployed. This results in biased real-world applications, which impact large populations [13, 14]. Normally, biases arise from using confounding variables, however bias can be present even when confounding variables are excluded from research.

For that reason, we wanted to investigate whether implicit biases can be found in data points where they should not exist, namely, measurements from bedside monitors. Therefore, we analyse bias from two demographic aspects, age and gender, using machine learning (ML) algorithms. The model is derived on 80% of the data, whilst the performance is evaluated from a holdout of 20% using a classification report [17] and confusion matrix as metrics [18].

Previous papers have shown both gender [19, 20, 21] and age [22] differences in biosignals. Moreover, ML algorithms have been used to predict age and gender from iris biometrics [23, 24]. ML has also been used for racial bias analysis in patient vital signs [25], but to the best of our knowledge researchers have not trained ML algorithms only on biosignals from bedside monitors to differentiate patient age and gender. This is a necessary analysis, and offers insights into whether differences in biosignals can unintendedly be learned by a model in a discriminatory way, and therefore make the model predict in favour of certain patient populations at the expense of others.

The paper is organised as follows. Section two describes the data used for the research as well as the applied methodology. Section three contains the results and discussion, whilst section four concludes the paper.

2 METHODOLOGY

This section outlines the data used for the research and the specifics of the preprocessing stage. Additionally, we give an overview of the algorithms used as well as the metrics which evaluate the trained models.

For the purposes of this research, we use the VitalDB dataset [26], which contains biosignals and clinical information from 6,388 non-cardiac surgical patients that underwent surgery in Seoul National

University Hospital in Seoul, Republic of Korea. The data has high-resolution with 2.8 million data points per case on average. The data of interest for us included: from demographic information, age and gender, and from vital information measured using Solar 8000M monitor, heart rate, respiratory rate, and (systolic, diastolic, and mean) blood pressure both invasively and non-invasively measured. As each of the biosignals was organised in a separate file, before proceeding with training the algorithms, it was necessary to merge the information while minding the time stamp of each measurement in order to maintain the continuity of the data. Additional information related to the surgical approach and the anaesthesia were not considered. The selected data was analysed in two different formats: first, the original data as recorded by the monitor without interference, and second, using features obtained with the tsfresh library [27]. In both cases, only patients with measurements for all biosignals of interest were considered, which reduced the population to 2905 patients.

The analysis of the demographic information, age and gender, is separate; namely, the gender analysis is a binary classification, whereas the age analysis is a multiclass classification problem. Each analysis was conducted both with the original data and with tsfresh statistics from the original data. All cases consider several ML algorithms: Logistic Regression (LR) [28], which estimates the probability of an event occurring, and so establishes baseline results, then Decision Trees (DT) [29] which represents a tree-like model showing series of decisions and possible consequences, Random Forest (RF) [30] which contains a collection of trees and uses a majority voting system to obtain the final prediction, and XGBoost [31], which compared to Random Forest operates on adjustable parameters through iterations, is proven as the most successful algorithm, even in cases of small and medium datasets, with limited feature count, as is the case here. However, as XGBoost is prone to overfitting when trained on small data, we performed parameter optimisation so to restrict the expansion of the model's structure.

The evaluation of the classification for each of the models was performed using a confusion matrix and a classification report (which observes metrics across each class), both for binary and multiclass classification. The confusion matrix visually represents the performance of the models, as it summarises the predicted and actual values obtained from the model and illustrates all misclassifications. The classification report shows the performance for each individual class and provides overall metrics for all classes. It observes the overall accuracy of the model and provides precision, recall, and F1-score values for each class. Precision measures how many of the positive predictions made are in fact correct, whilst recall measures how many of the positive cases from the overall positives were correctly predicted. The F1-score combines both metrics and shows intel into how many times the model made a correct prediction across the entire dataset.

3 RESULTS

The obtained results are divided into two separate groups: binary classification results for gender bias and multiclass classification results for age bias. The age bias results observe two age range divisions: one in three groups and another in four groups. The division of age ranges in three subgroups resulted in the first group of patients under 30 years, the second with patients between 30 and 49 years, and the third contained patients aged 50 and above. As majority of patients were aged 50 and over, and considering the age range considered for the third group was larger, we extended the analysis into a division of four groups, where the third range was split in two, with patients aged 50 to 69 years, and another with patients aged 70 and above.

3.1 Gender Bias

In order to perceive gender bias, the biosignals are used to classify patients as either male or female. The accuracy for all algorithms, both trained on the original data and the tsfresh features given in Table 1.

Table 1: Accuracy from gender bias analysis.

Models	Original Data	TSFRESH Features
LR	64%	61%
DT	99%	53%
RF	100%	63%
XGBoost	84%	58%

As can be observed from the Table 1, the prediction is better when trained on the original values of the data. As expected LR provides the baseline result, whereas the three remaining algorithms show improvement in performance. The accuracy of 84% for XGBoost shows that gender can be identified from biosignals in four from five patients, which is a significant number. The two remaining algorithms show an accuracy of 99% and 100% respectively, which essentially indicates that

biosignals can help AI algorithms to identify all patients' gender details.

The precision, recall, and F1-score are structured in Table 2. With XGBoost exists a drop in predictive power between the two classes, which is not the case with the results from DT and RF. The drop in the metrics for female patients can partially be due to a smaller pool of female patients. Nevertheless, these results are consistent with previous research data showing male patients have higher blood pressure compared to females [32].

Table 2: Classification report from gender bias analysis on the original data (M – male, F – female) (in %).

Models	Precision		Recall		F1-score	
	M	F	M	F	M	F
LR	64	61	87	29	74	39
DT	99	99	99	99	99	99
RF	100	100	100	99	100	100
XGBoost	83	85	91	73	87	78

This shows that models are able to detect subtle differences in data between patients of different genders, and while these subtle differences are necessary when analysing blood pressure information, they are not a beneficial feature when analysing biosignals in general, since models' performances need to be invariant to demographic information.

3.2 Age Bias

The results for age bias, obtained using the selected biosignals, are observed from two standpoints: first, where only three groups of patients are considered, and second, with four groups of patients considered (created by dividing one of the three groups from the first observation into two). As this approach uses multiclass classification, only three algorithms were considered; namely LR was not trained and tested for these data points. The results from the division of patients in three groups (under 30; between 30 and 49; 50 and over) are given in Table 3. The results from the division of patients in four groups (under 30; between 30 and 49; between 50 and 69; 70 and over) are given in Table 4.

Table 3: Accuracy from age bias analysis (3 groups).

Models	Original Data	TSFRESH Features
DT	99%	69%
RF	100%	76%
XGBoost	91%	75%

These results show that patients' age groups can be identified using biosignal information with an accuracy of 100% when using RF. The high accuracy results are obtained on the original data without value interference, whereas processing the data and using features extracted with tsfresh results in significant decrease of performance. When observing the behaviour of the models on the train and test data, the differences in metrics indicate that the models overfit when trained on the tsfresh features, which partially accounts for the worsened performance. Another reason is the difference in data points, meaning as there is lower data point count with tsfresh (since this approach aggregates the original data) the model is impacted by that reduction.

Table 4: Accuracy from age bias analysis (4 groups).

Models	Original Data	TSFRESH Features
DT	98%	43%
RF	99%	54%
XGBoost	80%	51%

With DT and RF obtaining near perfect results, it is interesting to analyse the performance of XGBoost and potential reasons for its performance. The confusion matrix from the analysis of three groups using the original data, given in Figure 1, shows the model mistakes patients aged between 30 and 49 with patients aged 50 and over, which might indicate that the model struggles with differentiating blood pressure values per age [33]. Potential conflicts in age-related medical problems can stem from differences in biological and chronological age [34], however with the other two algorithms performing with an accuracy approaching 100%, this is unlikely the case here.

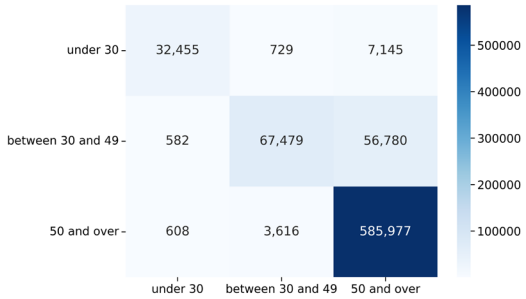

Figure 1: Confusion matrix for original data with three age groups analysed.

Another thing which can be noted is that the performance of the models decreases when patients aged 50 and over are divided in two groups, with DT dropping from 99% to 98%, RF dropping

from 100% to 99%, and XGBoost significantly dropping from 91% to 80%. The change in performance can be observed in the confusion matrix for the original data for four groups, as seen in Figure 2. Namely, once the patients are divided, the model is impacted and unable to successfully learn the difference between patients aged 50 to 69 and patients aged over 70. As the confusion matrix shows, a third of patients aged over 70 are misclassified into the group containing patients aged 50 to 69.

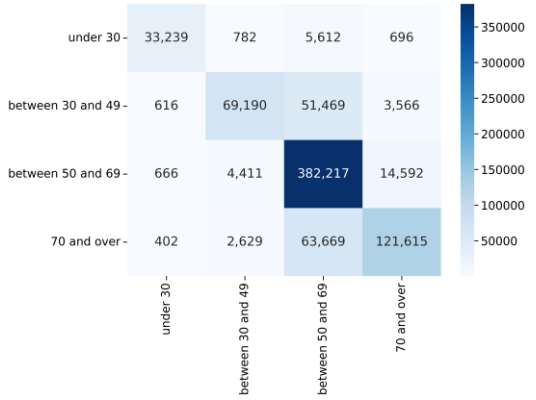

Figure 2: Confusion matrix for original data with four age groups analysed.

3.3 Discussion

The results for both gender and age bias show that ML algorithms are able to differentiate genders and age ranges based on biosignals, which in turn shows that identifying potential biases in data can be accomplished by observing whether specific input information can be used to predict classes belonging to a variable carrying said potential bias. In cases where the algorithms accomplish near perfect score, as is the situation here with DT and RF, it is safe to say that using the data in the same format might confuse the algorithms and lead them to predict based on information which should be disregarded.

With results showing high accuracy in predicting demographic information, it is necessary to discuss potential reasons behind the successful performance of the models. Namely, differences in biosignals based on gender and age have been shown, and it is likely the ML models observe these differences and make predictions on them. With models being able to differ between patient groups based on biosignals, it is possible that ML models trained on these biosignals for various other medical purposes also make their decisions based on these differences, and adjust predictions based on demographic information.

Therefore, another interesting discussion to touch up on are the implications of these results and the challenges which they pose for real-world use of ML algorithms in medical setting. Namely, implicit bias can easily be propagated along the pipeline, and create biased application, which in turn can lead to skewed outcomes and inequity among different patient populations. This can lead to favouritism of certain patient groups as well as reduced or inaccurate performance of models based on demographics. Depending on the application and the purpose of the algorithms, serious illnesses can be disregarded or overlooked, patients can be silenced on important health problems, patients might receive substandard preventive care, and many others. All of the above can lead to higher chances of worsening medical conditions, health complications, disruption of patients' lives, and in extreme cases, deaths which could have been avoided.

4 CONCLUSIONS

This paper proposes a demographic bias analysis approach from patients' biosignals, using ML algorithms to perform binary and multiclass classification in order to identify patient gender and age. The approach focused on analysing two types of results, firstly, the original data was used, and secondly, the data was processed and extracted tsfresh features were used. In both cases, bias could be seen, however bias was more prominent with the original data. This indicates that extracting features using tsfresh can be seen as a marginal mitigation technique in partially handling bias in this dataset. However, further research is required in order to understand whether the same holds for other data. Moreover, with results showing that biosignal information can be used to classify patients according to gender and age (with two separate analyses into three and four age groups), the approach can allow researchers to understand whether algorithms might detect hidden bias in data which cannot be easily observed by the developer. Therefore, the approach itself can be used to mitigate potential biases in creating and selecting datasets, as well as throughout the processing stages when developing AI-based medical systems. This would reduce the propagation of biased data and practices in real-world applications before they are deployed into production, which would greatly benefit patients discriminated upon by biased applications.

REFERENCES

[1] J.A. Sabin, "Tackling implicit bias in health care," N. Engl. J. Med. 2022, vol. 387, pp. 105-107, July 2022.

[2] J.R. Marcelin, D.S. Siraj, R. Victor, S. Kotadia, and Y.A. Maldonado, "The impact of unconscious bias in healthcare: how to recognize and mitigate it," The Journal of Infectious Diseases, vol. 220, issue Supplement_2, pp. S62-S73, September 2019.

[3] J.C. Lester, S.C. Taylor, and M.M. Chren, "Under-representation of skin of colour in dermatology images: not just an educational issue," British Journal of Dermatology, vol. 180, no. 6, pp. 1521-1522, June 2019.

[4] American Cancer Society, "Cancer Facts and Figures 2023,". [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf, [Accessed on Sep. 24, 2023].

[5] J.F. Dovidio, S. Eggly, T.L. Albrecht, N. Hagiwara, and L.A. Penner, "Racial biases in medicine and healthcare disparties," TPM, vol. 23, no. 4, pp. 489-510, December 2016.

[6] L. VanPuymbrouck, C. Friedman, and H.A. Feldner, "Explicit and implicit disability attitudes of healthcare providers," Rehabilitation Psychology, vol. 65, no. 2, February 2020.

[7] A. Dugravot, A. Fayosse, J. Dumurgier, K. Bouillon, T.B. Rayana, A. Schnitzler, Prof. M. Kivimaki, S. Sabia, and Prof. A. Singh-Manoux, "Social inequalities in multimorbidity, frailty, disability, and transitions to mortality: a 24-year follow-up of the Whitehall II cohort study," The Lancet Public Health, vol. 5, issue 1, pp. E42-E50, January 2020.

[8] H.R. Burstin, S.R. Lipsitz, and T.A. Brennan, "Socioeconomic status and risk for substandard medical care," JAMA., vol. 268, issue 17, pp. 2383-7, November 1992.

[9] A. Samulowitz, I. Gremyr, E. Eriksson, and G. Hensing, ""Brave men" and "emotional women": A theory-guided literature review on gender bias in health care and gendered norms towards patients with chronic pain," Pain Res. Manag., vol. 2018, pp. 6358624, February 2018.

[10] D. Westergaard, P. Moseley, F.K. Hemmingsen Sørup , P. Baldi, and S. Brunak, "Population-wide analysis of differences in disease progression patterns in men and women," Nat. Commun., vol. 10, no. 1, pp. 666, February 2019.

[11] B.N. Greenwood, S. Carnahan, and L. Huang, "Patient–physician gender concordance and increased mortality among female heart attack patients," Proc. Natl. Acad. Sci. U S A, vol. 115, no. 34, pp. 8569-8574, August 2018.

[12] A. Ben-Harush, S. Shiovitz-Ezra, I. Doron, S. Alon, A. Leibovitz, H. Golander, Y. Haron, and L. Ayalon, "Ageism among physicians, nurses, and social workers: findings from a qualitative study," Eur. J. Ageing, vol. 14, no. 1, pp. 39-48, March 2017.

[13] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447-453, October 2019.

[14] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," Pac. Symp. Biocomput., vol. 26, pp. 232-243, 2021.

[15] R. Lutz, Incident Number 110. In McGregor, S. (ed.) Artificial Intelligence Incident Database, Responsible AI Collaborative, 2023.

[16] A.J. Larrazabal, N. Nieto, V. Peterson, D.H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," Proc. Natl. Acad. Sci. U S A, vol. 117, no. 23, pp. 12592-12594, May 2020.

[17] Sklearn, "Classification report,". [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html, [Accessed on Sep. 24, 2023].

[18] Sklearn, "Confusion matrix,". [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html?fbclid=IwAR2TABvX6ymzxMD1Db7cV-wzNOFI6WzifcZHQez4FrVj0f6iCWv9RZON8es, [Accessed on Sep. 24, 2023].

[19] K. Prabhavathi, K.T. Selvi, K.N. Poornima, and A. Sarvanan, "Role of Biological Sex in Normal Cardiac Function and in its Disease Outcome – A Review," Journal of Clinical and Diagnostic Research, vol. 8, no. 8, pp. BE01–BE04, August 2014.

[20] A. LoMauro, and A. Aliverti, "Sex differences in respiratory function," Breathe (Sheff), vol. 14, no. 2, pp. 131-140, June 2018.

[21] R. Maranon, and J.F. Reckelhoff, "Sex and Gender Differences in Control of Blood Pressure," Clinical Science, vol. 125, no.7, pp. 311-318, October 2013.

[22] C.E. Boeke, M.E. Pauly, H.H. Stock, S. Pavlis, and J. B. Jackson, "The association of gender, age, body mass index, and vital signs in healthy plateletapheresis donors," Transfusion and Apheresis Science, vol. 41, no.3, pp. 175-178, December 2009.

[23] M. Erbilek, M. Fairhurst, and M.C.D.C. Abreu, "Age Prediction from Iris Biometrics," IET, December 2013.

[24] M. Rajput, and G. Sable, "Deep Learning Based Gender and Age Estimation from Human Iris,", SSRN Electronic Journal, June 2019.

[25] B. Velichkovska, H. Gjoreski, D. Denkovski, M. Kalendar, B. Mamandipoor, L.A. Celi, and V. Osmani, "Vital signs as a source of racial bias," medRxiv, February 2022.

[26] H.C. Lee, Y. Park, S.B. Yoon, S.M. Yang, D. Park, and C.W. Jung, "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients," Sci. Data, vol. 9, no. 1, pp. 279, June 2022.

[27] Tsfresh, "tsfresh,". [Online]. Available: https://tsfresh.readthedocs.io/en/latest/, [Accessed on Sep. 24, 2023].

[28] Sklearn, "Logistic regression,". [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, [Accessed on Sep. 24, 2023].

[29] Sklearn, "Decision Trees,". [Online]. Available: https://scikit-learn.org/stable/modules/tree.html#decision-trees, [Accessed on Sep. 24, 2023].

[30] Sklearn, "Random Forest,". [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, [Accessed on Sep. 24, 2023].

[31] XGBoost, "XGBoost Documentation,". [Online]. Available: https://xgboost.readthedocs.io/en/stable/?fbclid=IwAR0s6F9c8B3SFqZ6hvn81OhByL0e-a00h06g8rnRSUEzrUkYdGqXPd-1Yq4, [Accessed on Sep. 24, 2023].

[32] J.F. Reckelhoff, "Gender differences in the regulation of blood pressure," Hypertension, vol. 37, no. 5, pp. 1199-1208, May 2001.

[33] J.L. Lapum, M. Verkuyl, W. Garcia, O. St-Amant, and An. Tan, Vital sign measurement across the lifespan, 1st Canadian ed., Ryerson University, 2018, pp. 123-124.

[34] M.R. Hamczyk, R.M. Nevado, A. Barettino, V. Fuster, and V. Andrés, "Biological versus chronological aging," J. Am. Coll. Cardiol., vol. 75, no. 8, pp. 919-930, March 2020.