Article

Higher Self-Reported Mental Health Predicts Better Perceived Physical Health in Aging Adults

Audrey Young , Tamara Qawasmeh , Serena McCalla *

Posted Date: 6 February 2025

doi: 10.20944/preprints202502.0438.v1

Keywords: aging populations; health outcome prediction; machine learning in healthcare; successful aging; mental health and aging; logistic regression analysis; predictive analytics; artificial intelligence; physical health integrated

Article

Higher Self-Reported Mental Health Predicts Better Perceived Physical Health in Aging Adults

Audrey Young [1,2,*,†], Tamara J. Qawasmeh [2,†] and Serena McCalla [2,†]

[1] Prospect High School, Saratoga, CA 95070

[2] iResearch Institute, Glen Cove, NY 11542, USA

* Correspondence: audreyoung08@gmail.com

† These authors contributed equally to this work.

Abstract: Aging is a global phenomenon that has driven interest in successful aging (SA), characterized by optimal physical, psychological, and social functioning without major disabilities. This study leveraged Machine Learning (ML) models to predict factors influencing SA using self-reported physical health data from the University of Michigan's 2022 Wave 10 National Poll on Healthy Aging (n = 2,277). ``Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) models were evaluated, with LR achieving the highest accuracy (77.7%) and F1 score (78.2%). LR identified significant predictors of physical health outcomes, demonstrating a moderate positive correlation (r = 0.29) between physical and mental health, especially in individuals with "Very good" and "Fair" mental health ratings. These findings underscore the critical role of mental well-being in SA and highlight the potential of ML models to enhance healthcare strategies by identifying key health interdependencies.

Keywords: aging populations; health outcome prediction; machine learning in healthcare; successful aging; mental health and aging; logistic regression analysis; predictive analytics; artificial intelligence; physical health integrated

1. Introduction

By 2050, it is estimated that the number of the aging population will double, reaching approximately 2 billion individuals worldwide [1,2]. Healthy aging is essential for improved quality of life (QOL) and reducing strain on healthcare systems [3,4]. Scientists focus on genetics in successful aging research, while gerontologists highlight lifestyle and environmental factors. Genetics establish the foundational health metrics, but lifestyle and environmental factors play a vital role in determining quality of life [5–7]. Highlighting lifestyle factors provides a comprehensive view that can be applied to a wider range of aging populations [8,9].

Identifying characteristics of physically healthy older individuals enables the implementation of preventative measures from a young age [10,11]. By analyzing survey data on healthy aging, traits contributing to longevity and well-being can be targeted and promoted. Machine Learning (ML) models in healthcare can identify patterns and predict outcomes, enabling early intervention and personalized treatment plans [12,13]. Previous studies have used machine learning to predict successful aging, **often relying on limited datasets** [14,15].

With the aging population increasing and people living longer, there are variations in how different nations experience aging. The range "old age" is inconsistent, with studies varying in their focus on ages sometimes starting from 50, while others start from 60, 65, or 80 and beyond [16,17]. This age group is most susceptible to leading health problems such as cancer, diabetes, heart disease, dementia, and Alzheimer's. With the growing population of older adults living longer lives, it increases demand on healthcare professionals, medical resources, and infrastructure which strains healthcare systems. Previous journals have attempted to address this issue by elucidating lifestyles and demographics impact on health.

The objective of this study is to identify and rank the most important features that contribute to "excellent" physical health, the highest possible self-rating, among aging populations using ML models (LR, DT, RF). The highest-performing model will determine the most influential variables associated with self-rated "excellent" physical health. These insights will guide healthcare providers and policymakers in promoting successful aging through targeted interventions. This study seeks to address these gaps by leveraging a more comprehensive dataset and exploring feature importance in predicting self-reported physical health, with a focus on investigating the use of ML models to enhance our understanding of successful aging.

2. Materials and Methods

This study utilizes data from the University of Michigan's 2022 National Poll on Healthy Aging (NPHA) Wave 10 survey, which includes responses from 2,277 U.S. adults aged 50-80 years old. The survey covers a wide range of topics including demographics, mental and physical, housing, employment status, region, access to technology and caregivers, personal mindset, residential mobility, chronic health conditions, and healthcare professionals' perspectives. This dataset provides a comprehensive understanding of successful aging by integrating physical and mental health variables with socioeconomic and lifestyle factors to understand the various dimensions of an individual's life that contribute to overall well-being. [18,19].

2.1. Data Collection

This study analyzes a 2022 dataset (n = 2,277) from the University of Michigan (U-M) National Poll on Healthy Aging (NPHA). The NPHA surveyed U.S. adults aged 50 and older via phone and web-based interviews, covering major topics such as health and household, aging in place, arthritis, integrative medicine, and women's health. Data collection occurred in January and February 2022 through internet and phone-based surveys, with 1,000 interviews from participants aged 50 to 64 and 1,000 from ages 65 to 80. Sampling was restricted to AmeriSpeak Panel members who had responded to at least one survey within the prior six months.

2.1.1. Data Preprocessing

The analysis was conducted using Python 3.12 in Visual Studio Code on macOS 10.15+, employing several packages, including pandas for data manipulation, numpy for numerical operations, sklearn for machine learning and statistical analysis, matplotlib and seaborn for data visualization, and imbalanced-learn for handling imbalanced datasets. Missing values for categorical variables were filled with the mode, while missing values for numerical variables were filled with the mean. One-hot encoding was applied to convert categorical variables into numerical values, ensuring that the machine learning algorithms could process these variables correctly. Columns deemed irrelevant to the analysis, such as duration and case ID, were removed. Columns related to question one on self-perception of physical health, except for "Excellent" physical health, were dropped to focus on the primary outcome variable.

Given the imbalanced nature of the dataset, with fewer instances of "Excellent" physical health, the Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the dataset. SMOTE generates synthetic samples for the minority class to ensure a balanced distribution, which is crucial for training robust ML models.

2.1.2. Machine Learning Classifications

Three machine learning models—Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT)—were used to predict health outcomes in aging populations. Logistic Regression identified key predictors by assigning coefficients to variables. Decision Tree captured non-linear relationships, while Random Forest reduced overfitting by averaging results from multiple decision trees.

The dataset was split into 80% for training and 20% for testing. The training set was used to train the models, while the testing set was used to evaluate their performance. Hyperparameter tuning was performed using GridSearchCV to find the optimal model parameters for each ML algorithm. Model performance was assessed using various metrics, including accuracy, precision, recall, F1 score, and ROC AUC score. These metrics provide a comprehensive view of model performance, given the imbalanced class distributions. Feature importance analysis was conducted for the most optimal model to identify key predictors of "Excellent" physical health.

3. Results

The performance metrics for the LR, DT, and RF models on both training and testing datasets are summarized in Tables 1 and 2. These metrics include accuracy, precision, recall, F1 score, and the Receiver Operating Characteristic Area Under the Curve (ROC AUC).

Table 1. Train Data Model Performance.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.909	0.912	0.909	0.91	0.996
Decision Tree	0.718	0.704	0.718	0.704	0.977
Random Forest	0.815	0.813	0.815	0.811	0.995

Table 2. Test Data Model Performance.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.777	0.788	0.777	0.782	0.972
Decision Tree	0.704	0.687	0.704	0.688	0.972
Random Forest	0.764	0.757	0.764	0.759	0.992

The LR model achieved the highest test accuracy at 97.2%, closely followed by the RF model at 96.5%, and the DT model at 94.4%. The F1 scores, which balance precision and recall, were also highest for the Logistic Regression model, indicating its strong performance in handling imbalanced data.

3.1. Feature Importance

Feature importance analysis was conducted for the Logistic Regression model, which was identified as the best-performing model. The top features contributing to "Excellent" physical health were identified and ranked based on their importance (Table 2):

Table 3. Top Feature Importances in Logistic Regression.

Rank	Feature Variables	Feature Importance
1	Excellent Mental Health Rating	0.603581
2	Have Post Grad Study or Professional Degree	0.307683
3	No High Blood Pressure or Hypertension	0.269748
4	Very Satisfied with Current Social Life	0.264357
5	Definitely Yes Able to Live in Current Home Through Aging	0.233204
6	Very Confident in Ability to Afford Services for Help	0.224191
7	Very Comfortable Talking to Healthcare Providers About Integrative Health Approaches	0.218751
8	Not Currently Receiving Integrative Medicine	0.218018
9	Very Satisfied with Sexual Activity in the Past Year	0.210833
10	Did Not Experience Weight Gain or Slowed Metabolism	0.205516

Mental health, educational attainment, absence of hypertension, and social satisfaction are among the most significant predictors of excellent physical health [20,21].

3.1.1. Confusion Matrix Analysis

The confusion matrix for the Logistic Regression model (Figure 1 and Figure 2) measured the model's performance in classifying self-reported physical health across six categories: "Excellent," "Fair," "Good," "Poor," "SKIPPED ON WEB," and "Very Good." The majority of predictions fall along the diagonal of the matrix, indicating correct classifications. The Logistic Regression correctly identified 664 instances of "Excellent," 627 instances of "Fair," 564 instances of "Good," 700 instances of "Poor," 706 instances of "Skipped on Web," and 608 instances of "Very Good" in the training data (Figure 1). The model correctly classified 160 instances of "Excellent," 116 instances of "Fair," 98 instances of "Good," 170 instances of "Poor," 180 instances of "Skipped on Web," and 103 instances of "Very Good" in the testing data (Figure 2).

The highest number of correct classifications for both the training and testing datasets occurred in the "Poor" category, which reflects the higher frequency of respondents selecting poor self-reported physical health. This suggests that the model was more successful at classifying this common category due to its frequency in the dataset.

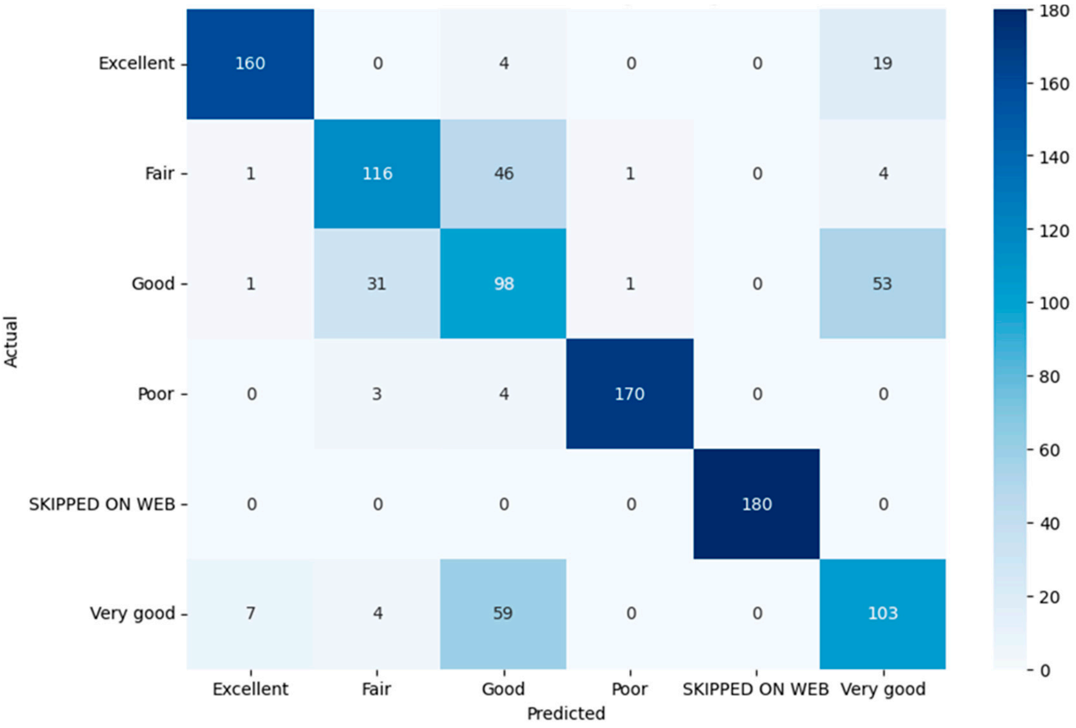
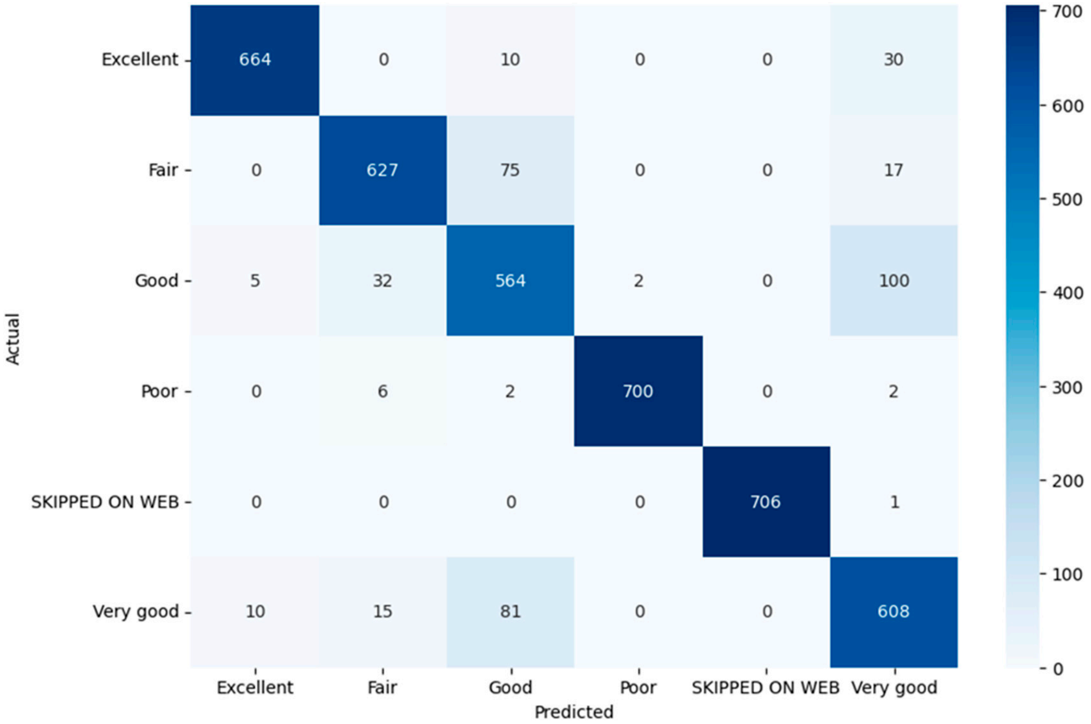


Figure 1. Train Confusion Matrix of Logistic Regression.



	Excellent	Fair	Good	Poor	SKIPPED ON WEB	Very good
Excellent	664	0	10	0	0	30
Fair	0	627	75	0	0	17
Good	5	32	564	2	0	100
Poor	0	6	2	700	0	2
SKIPPED ON WEB	0	0	0	0	706	1
Very good	10	15	81	0	0	608

Figure 2. Test Confusion Matrix of Logistic Regression.

3.1.2. ROC Curves

The ROC curves for the Logistic Regression model were plotted for both the training and testing datasets (Figure 3 and Figure 4). The area under the curve (AUC) for the training data was 0.996, demonstrating the model's ability to distinguish between "Excellent" and "Not Excellent" physical health in the training set. The AUC for the testing data was 0.992, indicating that the model maintained high performance when applied to unseen data. The high AUC values for both training and testing datasets confirm its effectiveness in predicting physical health outcomes.

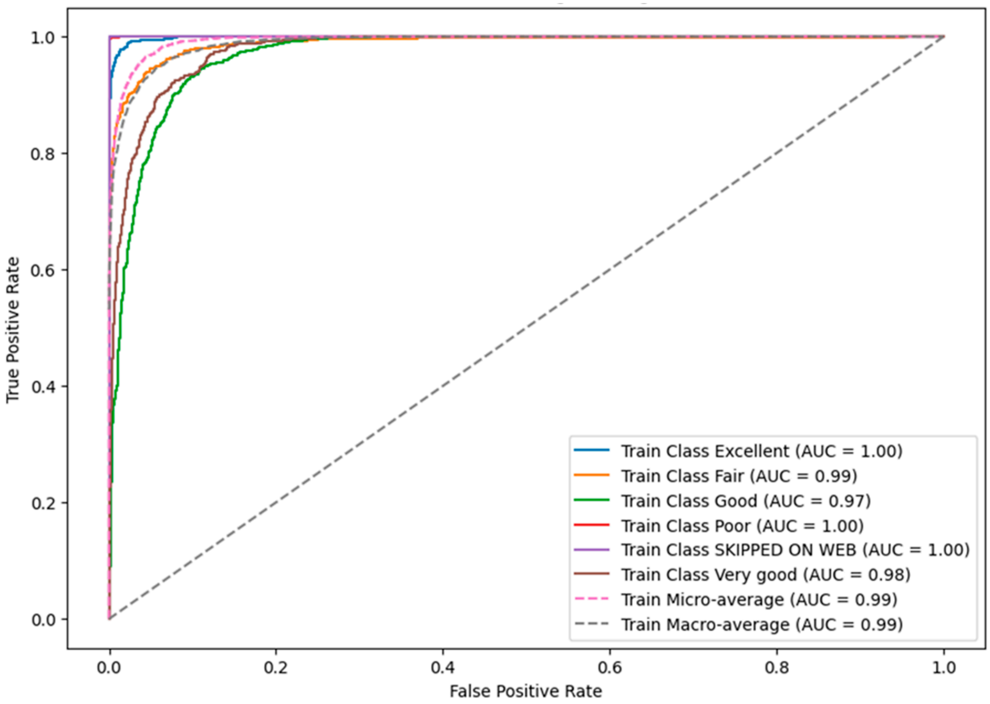


Figure 3. Train ROC Curve for Logistic Regression.

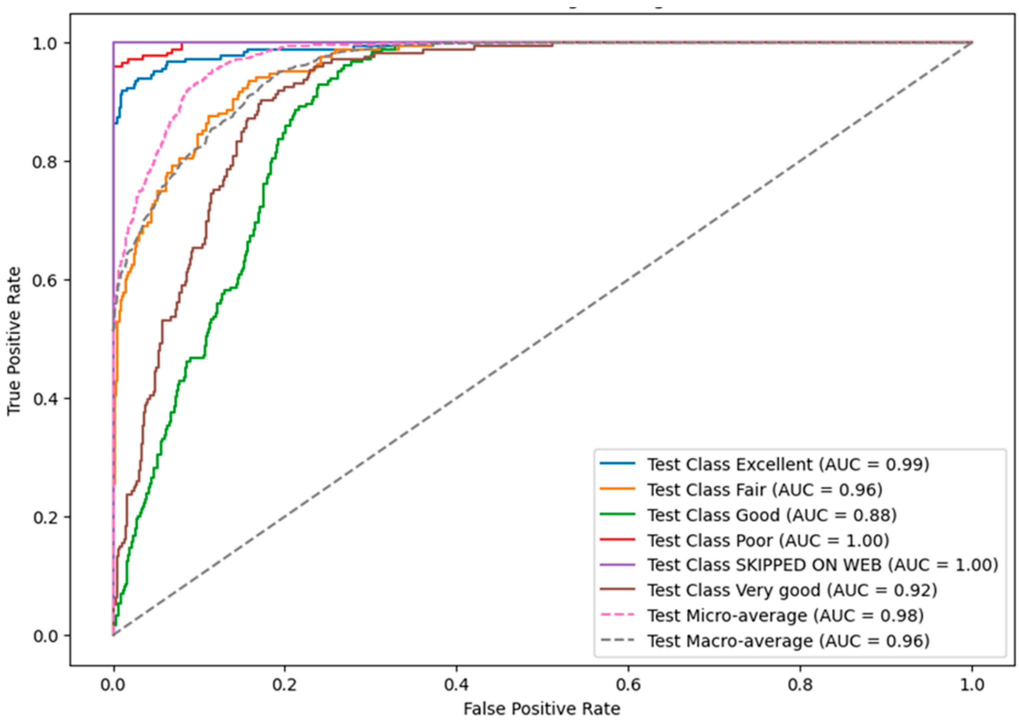


Figure 4. Test ROC Curve for Logistic Regression.

4. Discussion

This study evaluated the performance of three machine learning models, Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), to predict significant predictors of successful aging using data from the 2022 National Poll on Healthy Aging (NPHA). The results indicate that LR outperformed the other models with the highest accuracy (97.2%) and F1 score (97.2%) on the test data. The high performance of the LR model, as evidenced by its ROC AUC scores of 99.6% for training and 99.2% for testing datasets, demonstrates its ability to correctly classify instances of “Excellent” and “Not Excellent” physical health, with minimal false positives and false negatives. This model’s performance is likely due to its ability to handle linear relationships and assign coefficients to each predictor, facilitating the identification of significant variables.

The feature importance analysis of the LR model revealed that mental health, educational attainment, absence of hypertension, and social satisfaction are the most influential predictors of excellent physical health. Specifically, “Excellent” mental health emerged as the strongest predictor, determining the critical role of mental well-being in overall health outcomes. Educational attainment, specifically having a postgraduate or professional degree, also significantly contributed to excellent physical health, likely reflecting better access to resources and healthier lifestyles.

4.1. Limitations

Despite the effective results, this study has limitations. The self-reported nature of the data collected through phone and web-based interviews may introduce bias, as it may not fully represent the broader aging population. Additionally, the cross-sectional design of the survey limits the ability to draw causal inferences suggesting that incorporating longitudinal data may help to better understand the dynamic nature of aging and validate the findings across diverse populations.

5. Conclusion

LR emerged as the most accurate model for predicting key variables associated with "Excellent" physical health ratings, achieving an F1 score of 97.2%. The strongest predictor identified was "Excellent" mental health, highlighting the role of mental well-being in physical health outcomes.

These insights support the need for integrated health strategies focusing on mental health to improve physical health among older adults. Recognizing the traits of physically healthy older individuals can guide the implementation of early preventative measures, ultimately reducing the strain on healthcare systems and the economy. Future research should incorporate secondary datasets with diverse populations to further explore treatment options for those with poor health.

1. Future studies should investigate the role of mental health interventions, such as music therapy, in improving physical health outcomes. Integrating music-related factors into predictive models can help identify the role of simple and commonly-seen remedies like music in maintaining excellent mental health.
2. Future research should examine the impact of access to affordable, nutritionally dense foods on physical health outcomes. Integrating dietary factors into predictive models can help identify the role of nutrition in maintaining excellent physical health.

Author Contributions: Conceptualization, A.Y.; Methodology, A.Y. and T.Q.; Software, A.Y.; Formal Analysis, A.Y; Investigation, A.Y.; Resources, A.Y.; Data Curation, A.Y. and T.Q.; Writing-original draft preparation, A.Y.; Writing—review and editing, A.Y. and T.Q.; Visualization, A.Y.; Supervision, T.Q. and S.M.; Project Administration, S.M.; Funding, Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used for this study can be found at https://doi.org/10.3886/E204882V2.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rudnicka, E., Napierała, P., Podfigurna, A., Męczekalski, B., Smolarczyk, R., & Grymowicz, M. (2020). The World Health Organization (WHO) approach to healthy ageing. Maturitas, 139, 6–11. https://doi.org/10.1016/j.maturitas.2020.05.018
2. Marzo, R. R., Khanal, P., Shrestha, S., Mohan, D., Myint, P. K., & Su, T. T. (2023). Determinants of active aging and quality of life among older adults: Systematic review. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1193789
3. Asghari Varzaneh, Z., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2022). Prediction of successful aging using Ensemble Machine Learning Algorithms. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-022-02001-6
4. Caramenti, M., & Castiglioni, I. (2022). Determinants of self-perceived health: The importance of physical well-being but also of mental health and cognitive functioning. Behavioral Sciences, 12 (12), 498. https://doi.org/10.3390/bs12120498
5. Aiello, A., & Accardi, G. (2019). Aging successfully: The role of Genetics and environment in the era of the aging-boom. potential therapeutic implications. Current Pharmaceutical Design, 25 (39), 4131–4132. https://doi.org/10.2174/138161282539191226114927
6. Rosoff, D. B., Mavromatis, L. A., Bell, A. S., Wagner, J., Jung, J., Marioni, R. E., Davey Smith, G., Horvath, S., & Lohoff, F. W. (2023). Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for Healthy Aging. Nature Aging, 3(8), 1020–1035. https://doi.org/10.1038/s43587-023-00455-5
7. Zhang, J., Wang, S., & Liu, B. (2023). New insights into the genetics and epigenetics of aging plasticity. Genes, 14 (2), 329. https://doi.org/10.3390/genes140203

8. Noto, S. (2023). Perspectives on aging and quality of life. Healthcare, 11 (15), 2131. https://doi.org/10.3390/healthcare11152131
9. McManus, D. T. (2024). The intersection of spirituality, religiosity, and lifestyle practices in religious communities to successful aging: A review article. Religions, 15 (4), 478. https://doi.org/10.3390/rel15040478
10. Kankaanpää, A., Tolvanen, A., Heikkinen, A., Kaprio, J., Ollikainen, M., & Sillanpää, E. (2022). The role of adolescent lifestyle habits in biological aging: A prospective twin study. eLife, 11. https://doi.org/10.7554/elife.80729
11. Jarosz, E. (2023). Lifestyle differentiation among older adults: Exploring the links between individuals' behaviours, socio-demographic characteristics, health and wellbeing in later life. Ageing and Society, 43 (9), 2157-2172. https://doi.org/10.1017/S0144686X21001586
12. Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing Healthcare: The role of Artificial Intelligence in Clinical Practice. BMC Medical Education, 23 (1). https://doi.org/10.1186/s12909-023-04698-z
13. Chioma, N. S., Ayodeji, A. E., & Mayokun, A. D. (2024). Transforming healthcare with data analytics: Predictive models for patient outcomes. GSC Biological and Pharmaceutical Sciences, 27 (3), 025–035. https://doi.org/10.30574/gscbps.2024.27.3.0190
14. Odden, M. C., & Melzer, D. (2019). Machine learning in aging research. The Journals of Gerontology: Series A, 74 (12), 1901–1902. https://doi.org/10.1093/gerona/glz074
15. Ahmadi, M., Nopour, R., & Nasiri, S. (2023). Developing a prediction model for successful aging among the elderly using machine learning algorithms. DIGITAL HEALTH, 9, 205520762311784.https://doi.org/10.1177/20552076231178425
16. Prasad, G. L., Anmol, N., & Menon, G. R. (2018). Outcome of traumatic brain injury in the elderly population: A tertiary center experience in a developing country. World Neurosurgery, 111. https://doi.org/10.1016/j.wneu.2017.12.034
17. Yu, Z., Kong, D., Peng, J., Wang, Z., & Chen, Y. (2021). Association of malnutrition with all-cause mortality in the elderly population: A 6-year cohort study. Nutrition, Metabolism and Cardiovascular Diseases, 31 (1), 52–59. https://doi.org/10.1016/j.numecd.2020.08.004
18. Grønning, K., Espnes, G. A., Nguyen, C., Rodrigues, A. M., Gregorio, M. J., Sousa, R., Canhão, H., & André, B. (2018). Psychological distress in elderly people is associated with diet, wellbeing, health status, social support and physical functioning- A hunt3 study. BMC Geriatrics, 18 (1). https://doi.org/10.1186/s12877-018-0891-3
19. Bosnes, I., Nordahl, H. M., Stordal, E., Bosnes, O., Myklebust, T. Å., & Almkvist, O. (2019). Lifestyle predictors of successful aging: A 20-year prospective Hunt Study. PLOS ONE, 14 (7). https://doi.org/10.1371/journal.pone.0219200
20. Szychowska, A., & Drygas, W. (2021). Physical activity as a determinant of successful aging: A narrative review article. Aging Clinical and Experimental Research, 34 (6), 1209–1214. https://doi.org/10.1007/s40520-021-02037-0
21. Piccardi, L., Pecchinenda, A., Palmiero, M., Giancola, M., Boccia, M., Giannini, A. M., & Guariglia, C. (2023). The contribution of being physically active to successful aging. Frontiers in Human Neuroscience, 17. https://doi.org/10.3389/fnhum.2023.1274151

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.