**Developing a Deep Learning Model to Predict the Health Risks for Individual Migrants**

**Grant Proposal**

Naaisha Agarwal

Massachusetts Academy of Math and Science

85 Prescott St, Worcester, MA 01605

**Author Note**

Thank you to Dr. Kevin Crowthers for his support with this project.

**Executive Summary**

Currently, around 40 million people relocate each year, whether within a country or internationally, yet none of them have a way of factoring health into their decision. This is because they do not know what their individual predicted health outcomes will be. There have been several studies showing general, population-level trends of migrant health. However, migrant health varies significantly depending on baseline health, initial and final residential locations, and demographics. So, migrants need a way to predict their individual health outcomes for their specific situation.

Machine Learning models offer a promising approach to solving this problem. They have been increasingly used in the health field to predict health outcomes. Furthermore, a specific type of Machine Learning model, Deep Learning models, has been shown to handle time-based and causal event-based predictions effectively. Deep Learning models have been used to make health predictions for individuals. So, this project aims to address the problem of migrants needing a way to predict their individual health outcomes by using a Deep Learning model to make those predictions.

First, a dataset of migrant health data containing each individual's demographics, residential location history, and health history will need to be collected. Then, a benchmark evaluating state-of-the-art models on the dataset will be created. Finally, a new model will need to be trained and evaluated on the dataset.

Overall, this project will create a model that accurately predicts health outcomes for people who grew up in one area and relocated to another.

*Keywords:* migrant health, health predictions, Deep Learning, Machine Learning, personalized health

**Developing a Deep Learning Model to Predict the Health Risks for Individual Migrants**

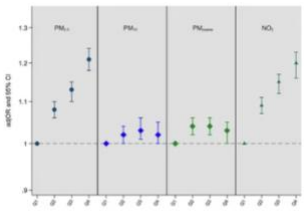


Figure 1: Shows how increasing pollution affects morbidity rates (Ronaldson et al., 2022)

Where each person lives has a massive impact on their health. Since each location has different air quality, allergens, lifestyle factors, access to fresh food, water quality, temperature, and other environmental factors, people have different health conditions based on where they live. For example, people living in more polluted areas, such as urban cities, tend to be more susceptible to respiratory diseases (Mohan et al., 2023). Air pollution also affects the risk of death, as shown in Figure 1. Furthermore, individual family history also affects respiratory disease risk (Hersh et al., 2011). Likewise, people living in tropical areas with high mosquito populations are more likely to encounter malaria or Lyme disease than those living in other areas (CDC, 2024). However, these conclusions only take into account people who live in that one area.

**Overview of Migrant Health**

Living in one area for an entire lifetime is not true for many people in the world. In fact, as of 2024, over 300 million people worldwide are international migrants, meaning they have moved to live in a new country long-term at least once in their lifetime (Paez-Deggeller, 2025). Researchers found that, in general, migrants experience better resilience to health risks when they first relocate, but this resilience can worsen over time as they continue to stay in their new location (Elshahat et al., 2022). Some examples of health changes that immigrants face include increased rates of cardiovascular disease, metabolic conditions, and respiratory diseases (Chen et al., 2024). Increased rates of disease impact quality of life, increase medical bills, and cause stress. Migrants cannot consider the health risks associated with their migration when deciding to relocate. When individuals can learn their predicted health risks before moving, they can choose a location that offers the best predicted health outcomes, or at least prepare themselves for specific conditions.

**Previous Migrant Health Studies**

Most studies on migrant health have been cohort-based, analyzing data on self-reported or clinically observed health. One study uses UK Biobank data to analyze how different categories of

relocation, related to changes in pollution levels, affect risk for major diseases and mortality (Chen et al., 2024). However, none of these examines specific trends in migrant health beyond general trends. The general results are not helpful to individuals who want to understand what their specific health risks might look like based on their personal history and data.

**Personalized vs. Population Health Predictions**

With improvements in technology, especially in artificial intelligence, there has been a shift toward personalized medical treatments and risk diagnosis, rather than a one-size-fits-all approach (Johnson et al., 2021). This personalized approach has been applied to create drugs, develop treatment plans, and even more accurately predict the risk or onset level of a disease for individual people (Parekh et al., 2023; Serrano et al., 2024). Previous clinical studies of migrant health have conducted population-level health assessments by examining general trends. However, this does not give people individual health predictions based on their personal history, which means it is not as accurate since it is not specific to them.

**Deep Learning with EHR Data to Predict Health**

| Prediction task | Outcome | # of papers | Papers |
|---|---|---|---|
| Diseases | Multiple diseases | 19 | [8–26] |
| | Cardiovascular disease | 17 | [15,18,19,27–40] |
| | Diabetes | 7 | [30,35,36,41–44] |
| | Kidney diseases | 7 | [27,29,36,45–48] |
| | Chronic obstructive pulmonary disease | 4 | [27,29,35,44] |
| | Alzheimer's diseases | 4 | [49–52] |
| Other | Mortality prediction | 6 | [10,19,26,49,53,54] |
| | Readmission prediction | 6 | [10,19,26,49,53,55] |

Figure 2: Shows different outcomes that deep learning models have been trained to predict (Amirahmadi et al., 2023)

One technique for predicting disease risk or onset uses Deep Learning models. Deep Learning is used to predict health changes and risks by analyzing temporal patterns in Electronic Health Record (EHR) data (Amirahmadi et al., 2023). They have been used to predict multiple different outcomes, shown in Figure 2.

**Datasets**

Several datasets exist with either EHR data or clinical data, with a specific focus on migrants, or can be filtered to focus on people who relocated. An example is the UK Biobank dataset, which includes patient demographics, baseline health, where they relocated, and their final health (Bycroft et al., 2018).

## Section II: Specific Aims

This proposal's objective is to develop a model that can predict health outcomes for individuals who relocate.

Our long-term goal is to help every single person be informed and prepared about their individual potential health outcomes if they relocate to a specific location.

**Specific Aim 1:** Create a dataset that contains information about a migrants residential location before and after their relocation, a baseline health before and after relocating, and their demographics.

**Specific Aim 2:** Create a benchmark evaluating state-of-the-art models on their ability to predict health outcomes of migrants.

**Specific Aim 3:** Train a Deep Learning model to predict health outcomes of migrants accurately.

The expected outcome of this work is a model that can effectively and accurately predict the health risks and outcomes of a potential relocation for each user.

## Section III: Project Goals and Methodology

### Relevance/Significance

This project is significant because it helps individuals prepare for their relocations and make informed decisions by accurately predicting health outcomes associated with a change in residential location. The proposed model is the first to accurately make these predictions for use, allowing the user to factor in their individual health when deciding where to relocate.

From a technical side, this model is significant because of its temporal and causal nature. It will be able to make a prediction years in the future, requiring a strong temporal understanding. Furthermore, it will factor in a causal event, the relocation, into the predictions.

### Innovation

This project is innovative due to its unique focus. There has been no model or solution to help individuals predict their health outcomes after migration. The dataset is innovative because it creates an entirely new dataset specifically for migrants. It also uses a new approach to link several datasets. The model itself will also be innovative because it will be the first Deep Learning model explicitly trained for predicting migrant health outcomes.

**Methodology**

The first step for this project is to create a dataset from longitudinal studies that track patients' health and residential locations. These datasets will have to be filtered to include only migrants. Then, the dataset features will have to be aligned to work across all combinations of datasets. Furthermore, the data will be normalized to account for differences in bias and collection methods across datasets. Finally, the data will be deidentified to remove any potential biases or risks of identification for the people who participated in the data collection.

Next, I will create a benchmark. The benchmark will consist of several state-of-the-art models' performances on the dataset. To do this, a pipeline will first need to be created to run the models on the dataset. Then I will run each model through the pipeline and collect the final results. For each model's results, I will need to run evaluation metrics to quantify performance and compare them. Since the model will predict several different aspects of health, I will not only have evaluation metrics for each aspect individually, but also need to create an evaluation metric that summarizes the model's overall performance across all aspects of health. This benchmark will help set a baseline for how well current state-of-the-art models predict health outcomes for migrants.

Then, I will train a model. This benchmark will help demonstrate how the model improves overall performance compared to the current state of the art. To train the model, I will split the dataset and use 70% for training, 15% for validation, and 15% for testing. The train split of the dataset will be

used for the model to learn and adjust its weights and biases. The validation split will be used during training to fine-tune hyperparameters and prevent the model from overfitting to the training data. Finally, the test split will be used to evaluate the model's performance on the dataset.

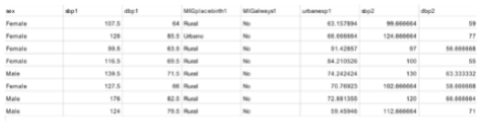
***Specific Aim #1:***

The objective is to create a dataset that contains information about a migrant's residential location before and after their relocation, a baseline health before and after relocating, and their demographics.

| sex | sbp1 | dbp1 | MIGplaceobrth1 | MIGallways1 | urbanesp1 | sbp2 | dbp2 |
|---|---|---|---|---|---|---|---|
| Female | 107.5 | 64 | Rural | No | 63.157894 | 99.666664 | 59 |
| Female | 126 | 85.5 | Urbano | No | 66.666664 | 124.666664 | 77 |
| Female | 99.5 | 63.5 | Rural | No | 91.42857 | 97 | 56.666668 |
| Female | 116.5 | 69.5 | Rural | No | 84.210526 | 100 | 55 |
| Male | 139.5 | 71.5 | Rural | No | 74.242424 | 130 | 63.333332 |
| Female | 127.5 | 66 | Rural | No | 70.78923 | 102.666664 | 58.666668 |
| Male | 176 | 82.5 | Rural | No | 72.881355 | 120 | 66.666664 |
| Male | 124 | 79.5 | Rural | No | 59.45946 | 112.666664 | 71 |

Figure 4: Screenshot of a sample from the preliminary dataset (Agarwal 2025)

**Justification and Feasibility.** Creating a dataset is important because the model needs data to train on. This specific aim is feasible because several datasets exist. For example, some migrant surveys include health and relocation data, such as the Peru Migrant Study (Carrillo-Larco et al., 2017). As shown in Figure 3, this dataset is helpful because it includes a significant number of migrants and tracks their health over several years. This shows that creating a dataset is feasible, as datasets like the Peru Migrant Dataset already exist that can be used. Other datasets track longitudinal health data for several participants, including migrants, such as the Panel Study of Income Dynamics from the University of Michigan (Gouskova et al., n.d.). These datasets can be incorporated into my final dataset by filtering for people with relocations.



| sex | sbp1 | dbp1 | MIGplaceobrth1 | MIGallways1 | urbanesp1 | sbp2 | dbp2 |
|---|---|---|---|---|---|---|---|
| Female | 107.5 | 64 | Rural | No | 63.157894 | 99.666664 | 59 |
| Female | 126 | 85.5 | Urbano | No | 66.666664 | 124.666664 | 77 |
| Female | 99.5 | 63.5 | Rural | No | 91.42857 | 97 | 56.666668 |
| Female | 116.5 | 69.5 | Rural | No | 84.210526 | 100 | 55 |
| Male | 139.5 | 71.5 | Rural | No | 74.242424 | 130 | 63.333332 |
| Female | 127.5 | 66 | Rural | No | 70.78923 | 102.666664 | 58.666668 |
| Male | 176 | 82.5 | Rural | No | 72.881355 | 120 | 66.666664 |
| Male | 124 | 79.5 | Rural | No | 59.45946 | 112.666664 | 71 |

Figure 4: Screenshot of a sample from the preliminary dataset

**Summary of Preliminary Data.** Using the Peru Migrant Study data, I have created a preliminary dataset. This dataset includes the participants' sex, the location before and after relocation, and their blood pressure before and after relocation, as shown in Figure 4. It contains data from over 300 migrant participants. Although it does not yet contain as many participants as needed and only contains a few features, it shows the feasibility of creating a migrant health dataset.

**Expected Outcomes.** The overall aim is to create a complete dataset that links each person's demographics, residential history, and medical records. This knowledge will be used to create a benchmark for state-of-the-art models to assess how well they predict health outcomes for migrants. This same dataset will also be used to train the model to predict health outcomes for people who lived in one place and then moved to another, as accurately as possible.

**Potential Pitfalls and Alternative Strategies.** Since there are no datasets for this problem, there might not be enough data to include all the necessary information. An alternative strategy will be to combine multiple datasets or collect data through linkage frameworks, such as the one used to link environmental and residential data to the UK Biobank (Vanoli et al., 2024). Another potential pitfall is that the data is not consistent across all the datasets I am trying to collect. For example, one dataset might have state-level locations, while another might have specific zip codes. An alternate strategy is to standardize on the most general location. In this example, I would have to convert the zip codes to their state locations.

***Specific Aim #2:***

The objective is to create a benchmark that assesses state-of-the-art models' capabilities to predict health outcomes for people who relocate.

**Justification and Feasibility.** Creating a benchmark is feasible because once I have the dataset, I need to define the evaluation metrics, and then I can run several state-of-the-art models to see how they perform. Some examples of state-of-the-art models I could use are Gemini 1.5 (Team et al., 2024), GPT 4o (OpenAI et al., 2024), and Foundation Health Models (Moor et al., 2023).

**Summary of Preliminary Data.** Using the preliminary dataset, Gemini 1.5 (Team et al., 2024) was evaluated on 300 participants and performed as shown in Figure 5 in Appendix A. The results show

that Gemini predicts clinically acceptable results less than 25% of the time. This is a very low score, showing that there is significant improvement to be made, justifying the project and the need for the benchmark.

**Expected Outcomes.** The expected outcome is that all state-of-the-art models will perform poorly on the benchmark, with less than 80% accuracy. This will demonstrate the need to train a model.

**Potential Pitfalls and Alternative Strategies.** Some potential problems are the fact that I might not be able to use some models because they are not open source. In that case, I will use other comparable, open-source models.

***Specific Aim #3:***

The objective is to create a benchmark that assesses state-of-the-art models' capabilities to predict health outcomes for people who relocate.

**Justification and Feasibility.** Creating a benchmark is feasible because once I have the dataset, I need to define the evaluation metrics, and then I can run several state-of-the-art models to see how they perform. Some examples of state-of-the-art models I could use are Gemini 1.5 (Team et al., 2024), GPT 4o (OpenAI et al., 2024), and Foundation Health Models (Moor et al., 2023).

**Summary of Preliminary Data.** Using the preliminary dataset, Gemini 1.5 (Team et al., 2024) was evaluated on 300 participants and performed as shown in Figure 5 in Appendix A. The results show that Gemini predicts clinically acceptable results less than 25% of the time. This is a very low score, showing that there is significant improvement to be made, justifying the project and the need for the benchmark.

**Expected Outcomes.** The expected outcome is that all state-of-the-art models will perform poorly on the benchmark, with less than 80% accuracy. This will demonstrate the need to train a model.

**Potential Pitfalls and Alternative Strategies.** Some potential problems are the fact that I might not be able to use some models because they are not open source. In that case, I will use other comparable, open-source models.

## Section III: Resources/Equipment

For this project, I will need several different resources. To start, I will need internet access to collect datasets to create my complete dataset. I will need access to a computer running Python and to Python libraries such as PyTorch and TensorFlow to train my model.

## Section V: Ethical Considerations

For this project, there are two primary ethical considerations to keep in mind. One ethical consideration is each person's privacy because the model is trained on people's individual data. However, since it is being obtained from public datasets, consent is already in place. One action I will take to mitigate any ethical concerns further is to ensure each data point is deidentified.

Another ethical concern is bias. The model will develop biases from whatever data it is trained on. So, if my final dataset contains biases, my model might inherit them. For this reason, I will track how features are distributed. To evaluate and reduce bias, I will create an evaluation set to assess bias and use specific metrics to quantify the bias in the dataset.

## Section VI: Timeline

The timeline is shown in Figure 7 in Appendix C.

## Section VII: Appendix

### Appendix A: Figure of Preliminary Benchmark Results

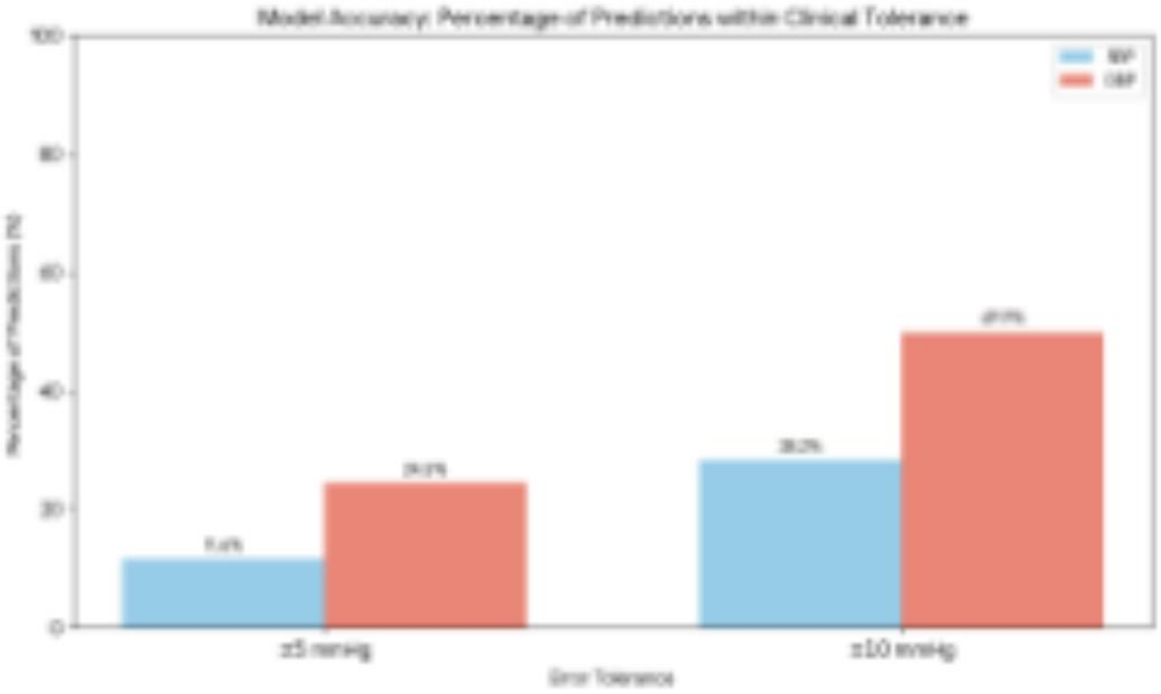



Figure 5: Comparison of Gemini 1.5's percentage of SBP predictions and DBP predictions within 5mmHg and 10mmHg (Agarwal 2025)

### Appendix B: Figure of Preliminary Data from Training

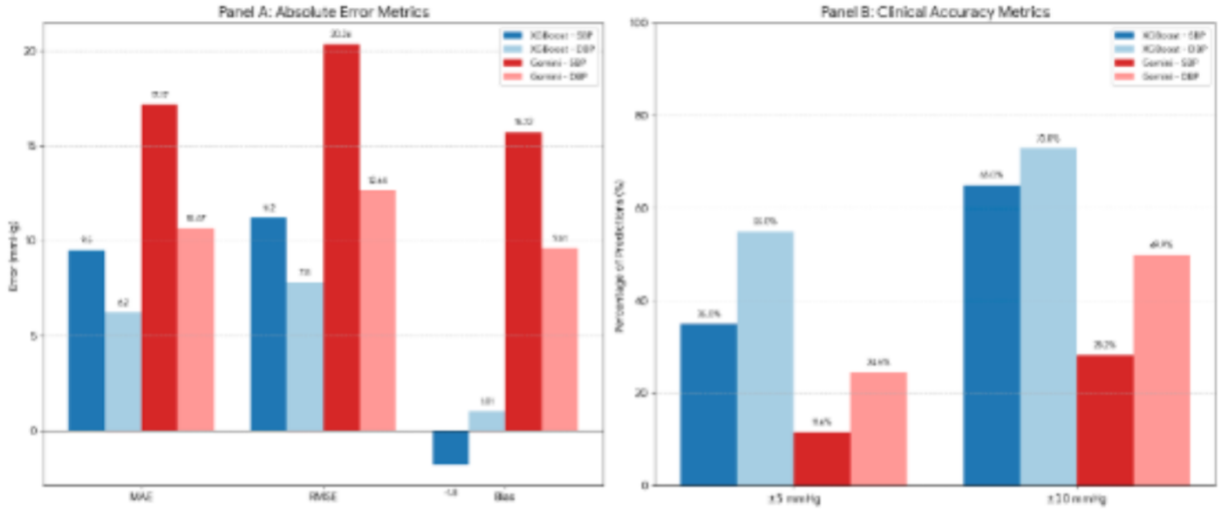



Figure 6: Comparing the trained XGBoost model against the Gemini Benchmark on a variety of metrics. (Agarwal 2025)

**Appendix C: Timeline Image**

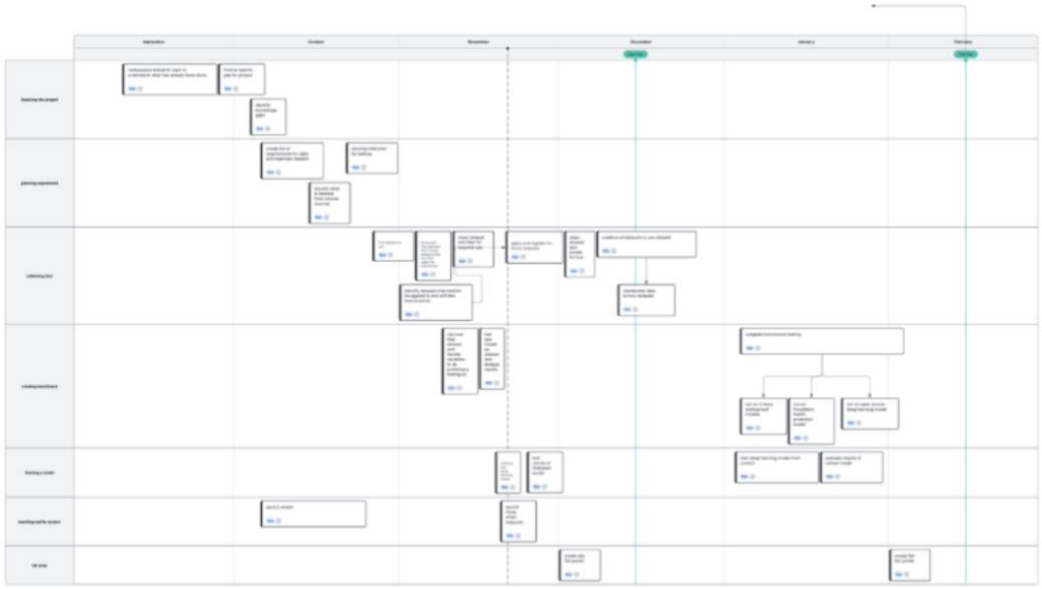


Figure 7: Image of a Gantt chart made in Lucid Chart of a timeline plan for the project (Agarwal 2025)

**Section VIII: References**

Amirahmadi, A., Ohlsson, M., & Etminani, K. (2023). Deep learning prediction models based on EHR trajectories: A systematic review. *Journal of Biomedical Informatics*, *144*, 104430. https://doi.org/10.1016/j.jbi.2023.104430

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Carrillo-Larco, R. M., Ruiz-Alejos, A., Bernabé-Ortiz, A., Gilman, R. H., Smeeth, L., & Miranda, J. J. (2017). Cohort Profile: The PERU MIGRANT Study–A prospective cohort study of rural dwellers, urban dwellers and rural-to-urban migrants in Peru. *International Journal of Epidemiology*, *46*(6), 1752–1752f. https://doi.org/10.1093/ije/dyx116

CDC. (2024, April 1). *Where Malaria Occurs*. Malaria. https://www.cdc.gov/malaria/data-research/index.html

Chen, G., Qian, Z. (Min), Zhang, J., Wang, X., Zhang, Z., Cai, M., Arnold, L. D., Abresch, C., Wang, C., Liu, Y., Fan, Q., & Lin, H. (2024). Associations between Changes in Exposure to Air Pollutants due to Relocation and the Incidence of 14 Major Disease Categories and All-Cause Mortality: A Natural Experiment Study. *Environmental Health Perspectives*, *132*(9), 097012. https://doi.org/10.1289/EHP14367

Elshahat, S., Moffat, T., & Newbold, K. B. (2022). Understanding the Healthy Immigrant Effect in the Context of Mental Health Challenges: A Systematic Critical Review. *Journal of Immigrant and Minority Health*, *24*(6), 1564–1579. https://doi.org/10.1007/s10903-021-01313-5

Gouskova, E., Andreski, P., & Schoeni, R. F. (n.d.). *Panel Study of Income Dynamics*.

Hersh, C. P., Hokanson, J. E., Lynch, D. A., Washko, G. R., Make, B. J., Crapo, J. D., & Silverman, E. K. (2011). Family History Is a Risk Factor for COPD. *CHEST*, *140*(2), 343–350. https://doi.org/10.1378/chest.10-2761

Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., Miceli, M., Kim, N. C., Orillac, C., Schnurman, Z., Livia, C., Weiss, H., Kurland, D., Neifert, S., Dastagirzada, Y., … Oermann, E. K. (2023). Health system-scale language models are all-purpose prediction engines. *Nature*, *619*(7969), 357–362. https://doi.org/10.1038/s41586-023-06160-y

Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., & Snowdon, J. L. (2021). Precision Medicine, AI, and the Future of Personalized Health Care. *Clinical and Translational Science*, *14*(1), 86–93. https://doi.org/10.1111/cts.12884

Mohan, A., Alupo, P., Martinez, F. J., Mendes, R. G., Zhang, J., & Hurst, J. R. (2023). Respiratory Health and Cities. *American Journal of Respiratory and Critical Care Medicine*, *208*(4), 371–373. https://doi.org/10.1164/rccm.202304-0759VP

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, *616*(7956), 259–265. https://doi.org/10.1038/s41586-023-05881-4

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., … Malkov, Y. (2024). *GPT-4o System Card* (No. arXiv:2410.21276). arXiv. https://doi.org/10.48550/arXiv.2410.21276

Paez-Deggeller, V. (2025, August 25). *Top Statistics on Global Migration and Migrants*. Migrationpolicy.Org. https://www.migrationpolicy.org/article/top-statistics-global-migration-migrants

Parekh, A.-D. E., Shaikh, O. A., Simran, Manan, S., & Hasibuzzaman, M. A. (2023). Artificial intelligence (AI) in personalized medicine: AI-generated personalized therapy regimens based on genetic and medical history: short communication. *Annals of Medicine and Surgery*, *85*(11), 5831. https://doi.org/10.1097/MS9.0000000000001320

Ronaldson, A., Arias de la Torre, J., Ashworth, M., Hansell, A. L., Hotopf, M., Mudway, I., Stewart, R., Dregan, A., & Bakolis, I. (2022). Associations between air pollution and multimorbidity in the UK Biobank: A cross-sectional study. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.1035415

Serrano, D. R., Luciano, F. C., Anaya, B. J., Ongoren, B., Kara, A., Molina, G., Ramirez, B. I., Sánchez-Guirales, S. A., Simon, J. A., Tomietto, G., Rapti, C., Ruiz, H. K., Rawat, S., Kumar, D., & Lalatsa, A. (2024). Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine. *Pharmaceutics*, *16*(10), 1328. https://doi.org/10.3390/pharmaceutics16101328

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., … Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* (No. arXiv:2403.05530). arXiv. https://doi.org/10.48550/arXiv.2403.05530