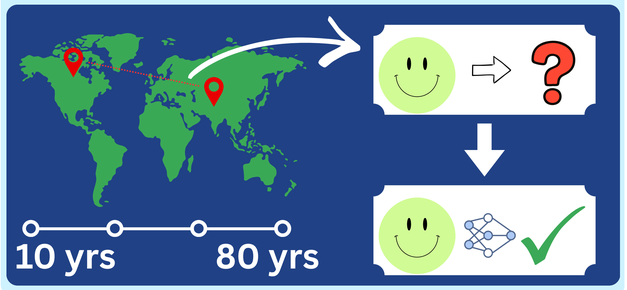
Using Deep Learning Model to Predict Individual Health Outcomes Due to Migration

Naaisha Agarwal

Advisor: Dr. Kevin Crowthers

Visual Abstract



Background

Problem

40 million migrants each year

- Migrants initially experience better resilience to diseases, but later on experience worsened health (Paez-Deggeller, 2025) ("Healthy migrant Effect")
- All migrant health studies so far are **population level** studies – does not help individual understand their specific situation

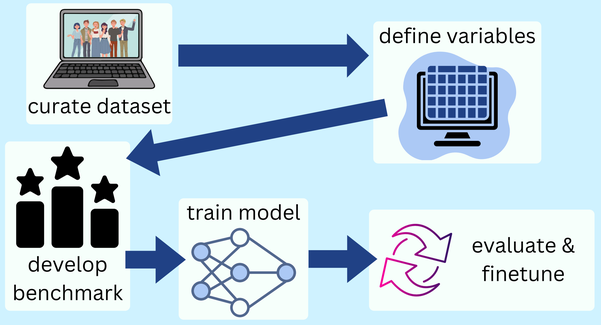
Opportunity

- Recent shift toward **personalized treatments** and risk diagnosis → provides increased accuracy for patients and doctors (Johnson et al., 2021)
- **Deep learning models** have been used to predict disease risk using **Electronic Health Record (EHR)** data through **temporal** patterns (Amirahmadi et al., 2023)

Key Challenges

- **Dataset** needs to be: large, long time period, migrant only, baseline and resulting health, individual demographics, residential history
- Modeling the **causal** and **temporal** nature of the data for the highest accuracy is tricky
- Model needs to be **interpretable** for the user

Methodology



Decision Matrix

Criterion	Weight	Linear Models	Deep Learning	Tree-Based Models
Temporal Modeling	25%	0.1	0.983	0.2
Feature Representation	20%	0.3	1	0.65
Overfitting Resistance	15%	0.7	0.645	0.85
Training Speed	10%	1	0.55	0.95
Interpretability	10%	0.9	0.325	0.9
Handling Missing / Sparse Data	10%	0.3	0.525	0.95
Scalability to Large Datasets	10%	0.6	0.917	0.8
Weighted Score	100%	0.47	0.77	0.67

Engineering Problem:

Migrants face several **health changes** after **relocating**, yet they don't have a way to **predict** their **individual** health changes so they can be **prepared**.

Engineering Goal

The goal is to engineer a **model** that **accurately** identifies **potential health risks individuals** will face based on their personal demographics, history, and location they are moving to.

Model Training Setup

Curated Dataset

Novel Continuous Health Risk Score

$$\text{HealthScore}_{j,t} = 0.5 \cdot \text{SRH}_{j,t} + 0.5 \cdot \left(1 - \frac{1}{m}\sum_{i=1}^{m} C_{i,j,t}\right)$$

Figure 1 Caption: This figure shows the formula to create the continuous health score where SRH represents the self rated health, m is the number of chronic health features, and C is whether the health feature is present. This score is higher when the person has a better overall health and lower when they have a worse overall health. It is used in the model training because the model predicts the change in health risk based on this score.

Relocation Event Definition

$$\text{Relocations}_j = [(t_1, S_{j,t_0}, S_{j,t_1}), (t_2, S_{j,t_1}, S_{j,t_2}), \ldots]$$

Figure 2 Caption: The relocation event was defined in the data by t, the year the person relocated, $S_{j,t1}$, the state they lived in before relocation, and $S_{j,t2}$, the state they moved to.

Temporal Sequences

$$\mathbf{x}_{j,t_{\text{move}}+\tau} = [\tau,\ F_j,\ \text{Embed}(S_{j,t_{\text{move}}}),\ \text{Embed}(S_{j,t_{\text{move}}+\Delta t})]$$

$$y_{j,t_{\text{move}}+\tau} = H_{j,t_{\text{move}}+\tau} - H_{j,t_{\text{move}}}$$

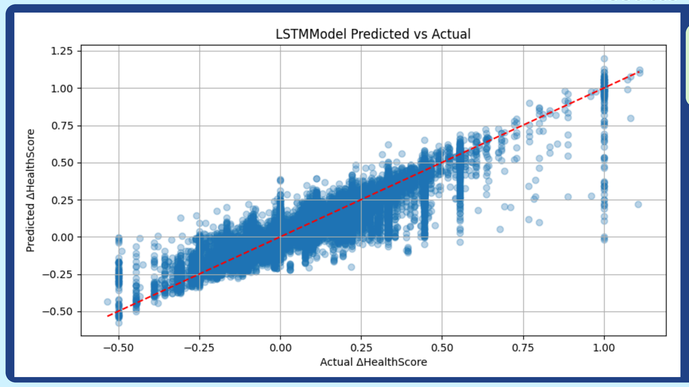
Figure 3 Caption: The data was transformed into temporal sequences so the model could learn the pattern as a change in time with t being the year, F_j representing baseline covariates such as sex, race, income, and each S is the state before and after. S is embedded because it is a categorical value.

Model Training

$$\hat{y}_j^{(k)} = f_\theta(\mathbf{X}_j^{(k)})$$

Figure 4 Caption: The model was trained where the model is f the function and it was given x (the temporal sequence) as an input and needed to predict y, the change in health outcome.

Results



Model accurately predicts health risk

Figure 5 Caption: This figure shows the comparison of the actual change in health score to the model's predictions. Overall, the model shows the capability to understand the change in health score as the points generally match the trend of the actual health score. However, the model struggles with the extreme change in health scores on either side, positive or negative. It also shows variance from the actual health score. But, since the direction is generally the same, it is evident that the model is able to predict close to the actual change in health scores demonstrating its ability to accurately predict health risk due to migration.

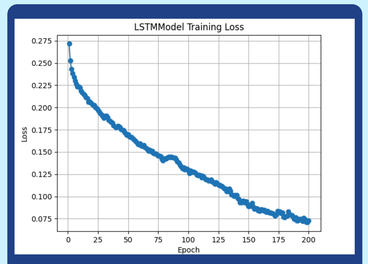


Figure 6 Caption: This figure shows the model's loss across epochs. The model shows consistently decreasing loss over epochs showing the need for all 200 epochs as it never plateaus. The model's loss decreases by a scale factor of almost 4, which shows a very strong decrease of loss, highlighting the model's capability to learn the data.

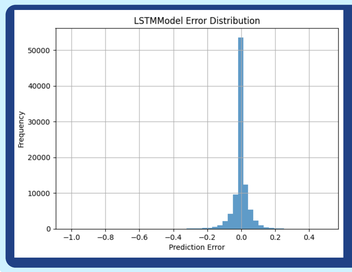


Figure 7 Caption: This figure shows the model's predicted error distribution. The highest frequency of error is very close to 0, with only a slight bias toward the negative. Overall the model shows balanced error distribution on both sides from there. This shows that the model does not tend to under predict or over predict, which highlights its strength.

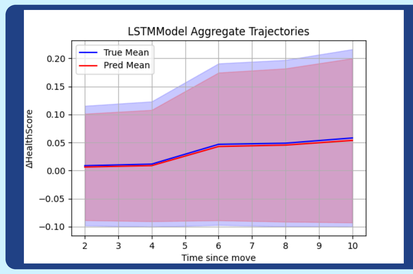


Figure 8 Caption: This figure shows a comparison of the aggregate trajectories of the ground truth and the model predictions. As shown in the figure, the aggregate trajectories follow almost the same path, which highlights the model's ability to accurately predict the change in health risk over time due to migration. However, the model's aggregate trajectory is slightly under the true mean, suggesting that the model might slightly under predict. Additionally, the model tends to not predict extreme values which is clear since the light blue shaded area extends longer in the positive and negative directions compared to the light red shaded area.

Dataset

About:

Sourced from Panel Study of Income Dynamics (Gouskova et al., 2026)

Key Facts:

- Total participants: 18890
- 879 features tracked
- 6 health features

~17 million data points

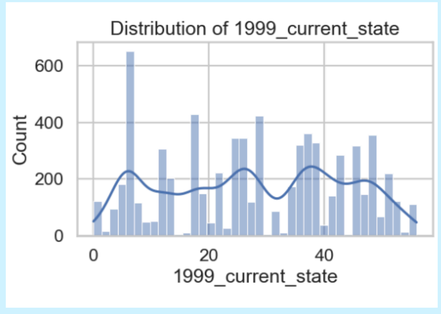


Figure 9 Caption: This shows the distribution of the families based on their starting state in 1999. Although certain states have many more participants than other states, overall the dataset shows a balanced mix of participants from many states, showing how it is more representative of the entire United States and not just one state or group of states.

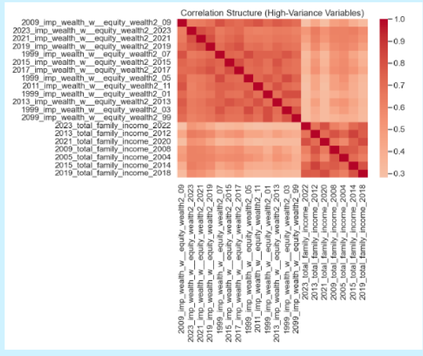


Figure 10 Caption: This figure shows a correlation heatmap of the highest correlated features in the dataset. Almost all the features are wealth and income across many years. This makes sense because one feature across years should show correlation. This shows that the dataset serves its function as it should show a correlation between the same variable across years.

Benchmark

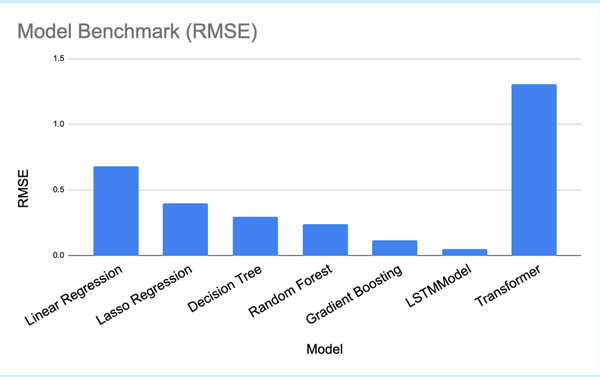


Figure 11 Caption: This figure shows a bar graph comparing the results of several different model structures after training on the curated dataset, representing a benchmark for this problem. RMSE was used as the error metric for comparison. LSTMModel, the model trained for this project, has the lowest RMSE, demonstrating that it has the least loss compared to all other model structures, thereby performing the strongest.

Analysis

- LSTM performs with the least error compared to several other model structures justifying its use as the model structure for this project
- Model has a consistent decreasing loss across epochs showing its need for 200 epochs of training
- Model's error distribution is balanced, showing that it does not bias toward either direction
- Model's predictions generally match trajectory of health changes, although it struggles with predicting extreme risk accurately and still shows variance

Conclusion

- **First** individualized migrant health prediction tool
- Empowers **informed decision-making** for migrants
- Establishes a **scalable framework** for personalized migrant health predictions and preparedness
- Can be further used by doctors and healthcare professionals to improve suggestions for individuals based on relocation history
- Shows model's ability to learn temporal and causal data

Key Contributions

- A new **dataset** containing demographics, baseline health, health after relocation, and residential histories specifically for migrants
- A **benchmark** various model structures on their ability to train to this dataset type
- A novel continuous health risk score helping define health risk over years
- A new **model** that is trained on the dataset and accurately predicts health outcomes for people who live in one area and migrate to another

Future Work

- Map each state to environmental factors
- Create counterfactual experiments for comparisons
- Test model on other datasets to evaluate performance
- Apply model to specific disease predictions to improve predictions based on relocations and suggest potential relocations for improved disease trajectories

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

Chi, G., Abel, G. J., Johnston, D., Giraudy, E., & Bailey, M. (n.d.). Measuring global migration flows using online data. Proceedings of the National Academy of Sciences of the United States of America, 122(18), e2409418122. https://doi.org/10.1073/pnas.2409418122

Gouskova, E., Andreski, P., & Schoeni, R. F. (2026). Panel Study of Income Dynamics.

Johnson, K. B., Wei, W., 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

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

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