Addressing Healthcare Inequality through Predictive Analytics

Elizabeth Baker, Abby Hathaway, Halle Loveday, Mowaninuola Olorunsola, Shao Hua Weng

1 Project Description

The U.S. healthcare system suffers from deep-rooted disparities affecting accessibility, affordability, and quality of care. Many Americans (specifically marginalized and low-income communities) encounter barriers when seeking medical services, including inflated medical costs, variability in healthcare quality, limited hospital access due to geographical location, disparities in hospice care, and high incidences of hospital-acquired infections (HAIs). Healthcare inequality has detrimental effects on individuals and society, leading to disparities in health outcomes, financial burdens, and overall well-being.

For example, low-income individuals, especially in rural areas, often have limited access to quality care because of the lack of adequate healthcare facilities. This can result in adverse outcomes, such as significantly higher mortality rates than those of urban regions. High medical costs, inadequate insurance coverage, and other financial barriers force many to forgo necessary treatments or accumulate medical debt. Financial barriers like these compound the issue of disparity. The U.S. healthcare expenditure is the highest globally, yet outcomes rank in the bottom half among developed nations, as demonstrated by studies like those conducted by Zywiel et al. (2017). This indicates a disturbing paradox: exorbitant spending frequently results in poor patient outcomes, perpetuating a cycle of financial hardship and suboptimal care, particularly affecting marginalized groups.

Additionally, disparities in healthcare quality contribute to unequal patient experiences, where marginalized communities receive subpar care. These issues contribute to unequal health outcomes, further marginalizing already vulnerable populations, and addressing these inequalities is very important in advancing the goals of improving public health, reducing economic strain, and ensuring equitable access to medical services.

To address these critical disparities, this project aims to conduct an in-depth technical analysis to propose solutions for overall healthcare inequality in the United States. We specifically focus on five difficulties within the healthcare system: geographic proximity, financial costs, healthcare quality, disparities in hospice care, and hospital-acquired infections. Each issue will be substantiated with robust statistical evidence to highlight its prevalence and impact clearly.

Informed by this evidence, we will propose hypotheses regarding why these problems persist and explore a potential way to improve the healthcare system. Our hypothesis centers around the significant influence of geographic location, specifically zip codes, on healthcare accessibility and outcomes. We aim to provide actionable insights that policymakers, healthcare providers, and advocacy groups can leverage to promote healthcare equity.

2 Technical Solution

We propose a robust application that leverages predictive machine learning models trained on multiple healthcare datasets, including patient surveys, financial reports, and hospital information, to quantify and predict the level of healthcare access by ZIP code. Users can utilize an interactive dashboard of the U.S. that allows them to identify high-risk geographic areas with significant healthcare disparities.

By training our model on multiple datasets encompassing patient satisfaction, financial burdens, hospital effectiveness, infection rates, and hospice care quality, we aim to quantify and predict the level of healthcare access by ZIP code.

The user can easily explore various regions of the United States and receive informative statistics describing the state of healthcare within that specific region. The application also utilizes inferential statistics and machine learning principles to provide a prediction to the user characterizing the overall healthcare quality potentially received if care were sought within the selected region. The machine learning model was trained on various datasets gathered from the United States government healthcare database (data.cms.gov). Topics of these datasets include healthcare costs, patient satisfaction, hospital care quality, hospital infection rates, and patient equity metrics. More details describing model training can be found in the Methodology section of this report.

Our technical approach includes data aggregation, feature engineering, and model training for our machine learning model, as well as developing a comprehensive, interactive web dashboard featuring an interactive U.S. map. Users can explore geographic areas to identify regions with heightened healthcare risk factors. This transparent interface aims to promote informed decision-making and solutions in the U.S. healthcare system.

3 Related Work

3.1 Geographical Disparities

Population Health for Nurses (Ochs et al., 2024) discusses how a person’s zip code or general geographic location can significantly impact health outcomes by determining access to healthcare facilities, the quality of care within said facilities, and exposure to environmental risk factors. While this project is not directly examining the environmental risk factors associated with each region, it is an important note when discussing healthcare topics in the context of geography. This is especially true when evaluating inferential statistics and model predictions. In addition to the idea of quality of care, a study done by Brown and colleagues reveals that socioeconomically disadvantaged neighborhoods were more likely to lack or lose healthcare facilities between 2000 and 2014 (Brown et al., 2020). Not only do certain geographical areas have fewer facilities than others, but the facilities that exist often provide lower-quality care in comparison with national averages. These findings suggest that the uneven geographic distribution of healthcare resources can emphasize disparities in healthcare access, quality, and outcomes.

Graham (2016) also explores the divide that location creates when it comes to healthcare outcomes and quality. His research largely centers around policy changes that should be carried out in order to counteract the challenges associated with this phenomenon. His work provides a framework for what solutions can be presented following the examination of the model implementation.

3.2 Racial Disparities

The United States healthcare system is deeply intertwined with systemic inequalities that affect marginalized populations and communities, particularly people of color. An article published in PubMed Central (Butler et al., 2021) discusses the multitude of factors, both historical and modern, that contribute to racial and ethnic inequalities in healthcare. This paper also provides insight into solutions that can be implemented by healthcare providers to mitigate these disparities. Similar to Graham's work, we can leverage the solutions proposed here to further understand why some regions perform poorer than others in healthcare quality and what can be done to create an equilibrium across locations in the United States.

3.3 Financial Disparities

The United States has the highest per capita healthcare expenditure rate in the world. Despite this, the U.S. also ranks in the bottom fifty percent for healthcare outcomes among developed countries (Zywiel et al., 2017). The study by Zywiel and colleagues elucidates how high-value procedures often result in below-average outcomes, specifically in the orthopedic sector of healthcare. Poor outcomes coupled with high cost are often associated with the tendency within the American medical ideology to quickly resort to invasive procedures, such as surgery, despite not being completely necessary. This can disproportionately affect low-income patients due to the fact that poor medical outcomes result in further intervention from healthcare practitioners, creating a downward spiral of excessive spending and disappointing outcomes.

In accordance with the idea of unnecessary healthcare costs for low-income patients, a 2023 study examined healthcare affordability issues in high-income countries. The study found that Americans with lower incomes are more likely to skip necessary medical care and encounter problems paying medical bills compared to their counterparts in similar high-income countries (Doty et al., 2023).

4 Data

4.1 Dataset Descriptions

Our analysis relies on publicly available datasets related to healthcare performance, costs, and accessibility. These datasets are gathered from the public United States healthcare database (data.cms.gov) and include:

• **HCAHPS-Hospital.csv** – Contains patient star ratings (out of 5) for categories such as cleanliness, staff communication, staff responsiveness, the likelihood of recommending the hospital, and the overall rating.

• **Health_Equity_Hospital.csv** – Contains the number of domains (0–5) that the hospital can affirm were used to assess its commitment to health equity.

• **Family_Practice.csv and General_Practice.csv** – Provide the minimum and maximum copay amounts for both new and established patients in their respective types of medical practices.

• **Hospital_General_Information.csv** – Contains the zip code and address of each of the hospitals examined.

• **Healthcare_Associated_Infections-Hospital.csv** – Contains a score for each hospital indicating the severity and frequency of infections associated with receiving healthcare, taking into account bloodstream, urinary tract, and intestinal infections.

• **Provider_CAHPS_Hospice_Survey_Data_Feb2025.csv and Hospice_General-Information_Feb2025.csv** – Contain information about hospice providers. The former dataset includes scores for emotional support, respect, pain management, communication, and training; the latter indicates the type of ownership of the provider (non-profit, for-profit, or government-operated).

The purpose and function of each of the provided datasets are further discussed in section 5.1 (Data Aggregation).

4.2 Data Collection and Annotation

To ensure data quality and usability, we normalized data fields for consistency, filtered incomplete or incorrect records, assigned categorical labels where necessary (e.g., hospital type and ownership), and filtered inconclusive fields. Normalizing data fields includes ensuring that two different representations of the same variable (for example, treating “poor” and “bad” as equivalent in terms of healthcare quality measurements) are unified. It also ensures that all numerical values are represented in the same units. All categorical variables were encoded numerically so that rankings could be applied when necessary, ensuring that the analysis is mathematically sound and objective. Additionally, many fields within these datasets were incomplete and were removed from the analysis to prevent introducing outliers or skewed predictions.

5 Methodology

5.1 Data Aggregation

The datasets HCAHPS-Hospital.csv and Health_Equity_Hospital.csv were analyzed to evaluate patient satisfaction and overall hospital effectiveness based on survey metrics and equity measures. Healthcare costs were examined via the Family_Practice.csv and General_Practice.csv files to compare costs across regions and practices. Analyzing infection rates through Healthcare_Associated_Infections-Hospital.csv provides an objective measure of a provider's efficacy. Provider_CAHPS_Hospice_Survey_Data_Feb2025.csv supplies data regarding hospice provider ownership, allowing further insights into how profit may influence healthcare outcomes. Hospitals and other providers are grouped based on ZIP code data from Hospital_General_Information.csv to evaluate various metrics by regional accessibility.

To develop the predictive model, we first cleaned and aggregated the datasets (handling missing values, standardizing formats, removing duplicates, and correcting inconsistencies). After merging data from various CSV files using common identifiers (such as hospital facility codes and ZIP codes), a final, clean dataset was produced. The dataset contains healthcare providers identified by their identification number, along with the features selected for prediction. All providers are grouped based on zip code.

5.2 Feature Engineering

Features were created for each subfield of analysis, including:

- *Hospital Quality Metrics:* Average patient satisfaction, hospital performance benchmark, and equity of care.

- *Healthcare Costs:* Healthcare cost burden, doctor availability per capita, and low-income patient ratio.

- *Hospital Accessibility:* Hospital density by ZIP code.

- *Hospital-Acquired Infections (HAI):* Average infection rate per hospital and hospital infection risk category.

- *Hospice Care Quality:* Hospice satisfaction score and ownership type and quality correlation.

Statistical methods were used to determine the top 15 features contributing to the model. Once the Random Forest Classifier was trained, we extracted the model's built-in feature importance scores. Each importance score reflects how much its respective feature contributes to reducing impurity across all of the decision trees. Impurity refers to the uncertainty of classification. Therefore, the importance scores reflect how much influence each feature has with respect to final classification. A discussion surrounding the specific features used and their importance scores can be found in section 6, Experiments and Results.

5.3 Model Training

Model training includes implementing a "Multi-Class Classification" approach using random forest classification on structured healthcare data. Initially, the team opted to use a logistic regression model to analyze the data. However, evaluation metrics, such as precision and recall, were quite poor. This suggests that our data may not have a linear relationship with the features selected, and a more sophisticated model is required. Random forest classification has many advantages in comparison with logistic regression. These advantages include less sensitivity to outliers and multicollinearity, automatic interaction handling (no need to manually identify interactions between predictors), and simpler categorical variable encoding. Logistic regression is often more interpretable when it comes to feature-to-feature impact, but overall model performance is more important than single-feature analysis for this project.

The following features were selected to quantify risk scores for each zip code present in the datasets: average patient satisfaction, average infection score, equity score, median household income, hospital count, and average new patient copay. Thresholds were selected to determine whether or not a given quantity would increase the risk score associated with a location. Household income and patient satisfaction thresholds were set to be the value of the first quartile of that variable within the overall data. If either of these variables were above their threshold, the risk score was incremented by a value of 1. The threshold for infection score was set as the third quartile of the variable's data. If an infection score was above this value, risk score increases by 1. Since equity scores are given on a scale of 0-5, the threshold for equity was hard-coded as a value of 3. Equity scores below three result in incrementing the risk score by 1. The same is true for hospital count, except the hospital count threshold is 2 hospitals. The patient copay threshold was set to 40 dollars because this is a reasonable copay amount for a typical office visit. Copays above 40 dollars result in a 1 point increase in risk score.

The model identifies risk levels (e.g., low, mild, moderate, high) based on key socioeconomic and hospital quality indicators derived from the datasets. Risk scores less than or equal to 1 are classified as "low" risk, scores of 2 are classified as "mild" risk, scores of 3 are classified as "moderate" risk, and scores greater than 3 are classified as "high" risk areas.

The random forest classifier was trained using an 80% to 20% train to test split. 80% of the data was used for training and 20% was used for testing. The classifier contains 50 decision trees with each tree being limited to a depth of 5 (this mitigates the risk of overfitting). Results of the model are discussed in section 6, Experiments and Results.

5.4 Result Visualization and Interpretation

We developed an interactive map with the Leaflet library where users can enter a ZIP code and receive an overall risk score from 1 to 3 based on healthcare access, costs, and quality in that area. The map in Figure 1 displays the risk scores for all the zip codes the model currently supports. Individual scores can also be queried through the search box. We logged all the scores predicted by our model in a CSV file and imported the results onto the map for a static display.

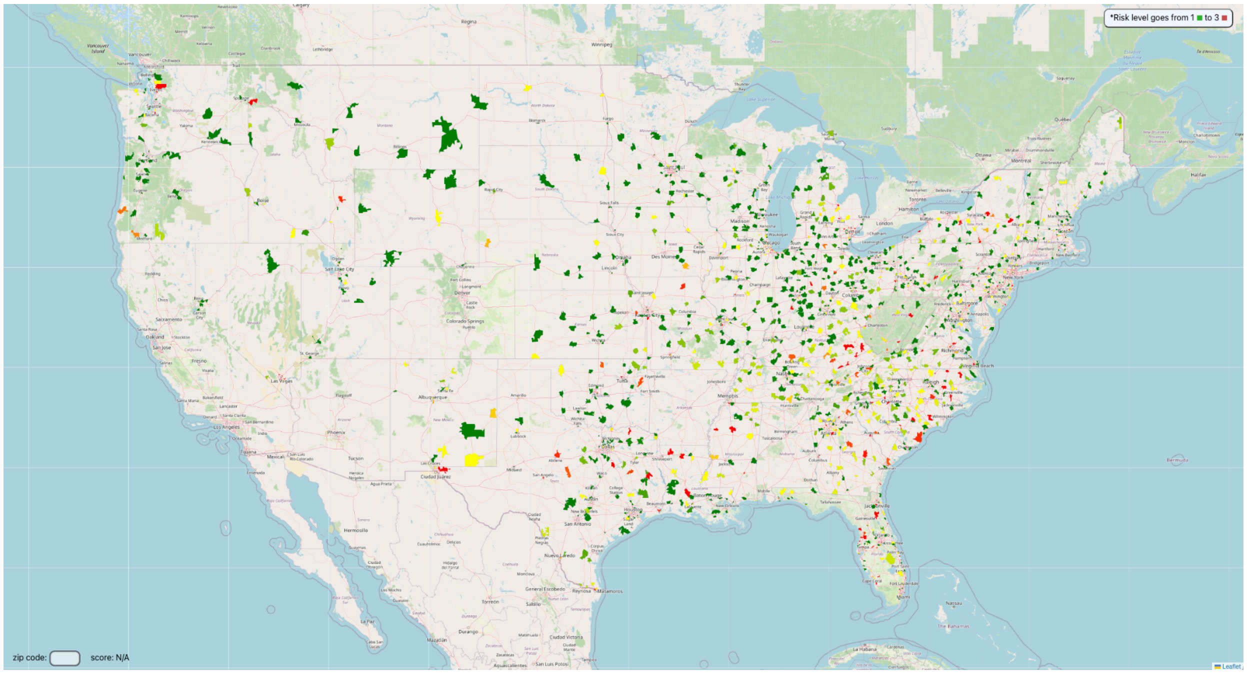


Figure 1. Interactive map color-coded risk score by zipcode.

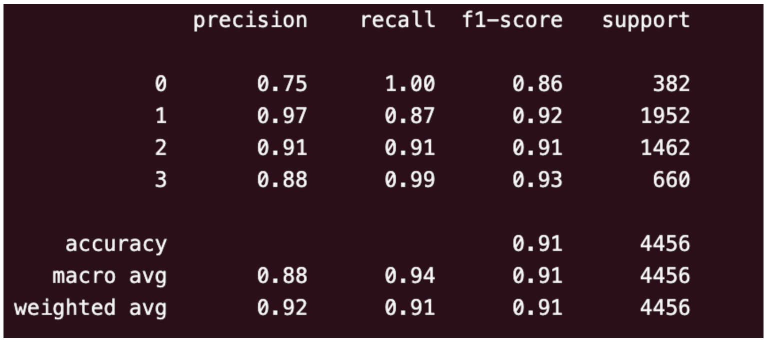
5.5 Evaluation Metrics

Our model and web application are evaluated based on model performance through precision, recall, F1 scores, and accuracy levels. These metrics are calculated for each risk classification to determine effectiveness across each level. Overall model accuracy is also measured in order to ensure that proper features have been selected. Evaluation metrics are extracted by comparing prediction labels from testing to the true labels defined during the training stage.

6 Experiments and Results

6.1 Model Results and Evaluations

The random forest classification model achieved strong performance across all risk score levels. Evaluation metrics across classes are depicted in Figure 2.



```
              precision    recall  f1-score   support

           0       0.75      1.00      0.86       382
           1       0.97      0.87      0.92      1952
           2       0.91      0.91      0.91      1462
           3       0.88      0.99      0.93       660

    accuracy                           0.91      4456
   macro avg       0.88      0.94      0.91      4456
weighted avg       0.92      0.91      0.91      4456
```

Figure 2. Evaluation metrics of the random forest classification model.

Overall model accuracy was 91% across 4,456 samples. The slightly large discrepancy between precision and recall for the low-risk class likely indicates that there is some overprediction of this class. The mild risk class had few false positives based on the precision score. The moderate risk class gives a balanced performance with equal precision and recall. The high-risk class has a very high recall score, capturing almost all high-risk cases, and the precision value indicates few false positives. The macro average and weighted average F1 scores prove that the model performs well across classes of varying sample sizes and imbalanced classes. The model as a whole is both reliable and accurate. It is especially accurate for high-risk classifications, which is especially important for individuals living in these areas. An inaccurate classification for a high-risk area could lead potential patients to poor outcomes when seeking out healthcare.

Feature importance scores were calculated to evaluate the effectiveness of each feature. Importance scores are depicted in Figure 3. Hospital count is the best predictor by far. Zip code seemed to be a relatively poor predictor, but this may mean that zip code may be too tight of a constraint for our purposes. The high importance score of hospital count suggests that areas with low hospital counts are more likely to be at risk. This supports the hypothesis that geographical location is heavily predictive of healthcare quality because rural areas typically have fewer healthcare providers in comparison with urban areas.

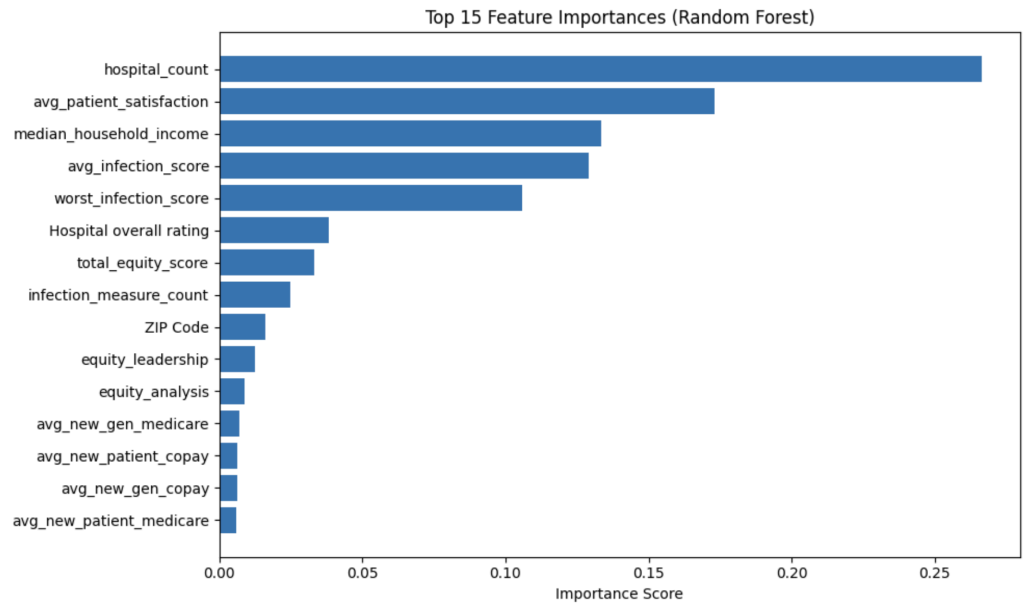


Figure 3. Feature importance scores to evaluate the effectiveness of each feature on healthcare quality.

Preliminary evaluations indicate that the application is scalable—the large datasets used could potentially be extended to multiple countries if required. One limitation noted is that the current data are not real-time, necessitating manual updates to refresh model predictions. Model training efficiency is acceptable, with each run taking between 2–3 minutes.

6.2 User Interface Progress

Our frontend dashboard successfully reads the zipcode risk score data from our predictive model through CSV file outputs, and displays the corresponding risk scores on the US map for the relevant zip code areas. Users can also query zip codes for their risk scores.

6.3 System Integration

Future cloud integrations with the learning model would be useful when we expand our predictive model to support producing risk scores for all zip codes without front-loading all the work in the web application, but for our current range of zip codes, this approach to reading data will suffice.

7 Conclusions

Our project provides meaningful insights into healthcare inequality in the U.S. and delivers a practical tool for policymakers, healthcare providers, and advocates aimed at mitigating these disparities. The analysis underscores the role of geographic and socioeconomic factors in determining healthcare outcomes.

Our contributions include:

• Identifying key areas of healthcare disparities.

• Building an accurate and reliable predictive model.

• Creating an interactive, web-based tool that provides actionable insights.

Looking ahead, future work and further enhancements could include incorporating real-time healthcare data, testing more sophisticated predictive models, and expanding our data sources. These improvements would yield more precise predictions and support nationwide efforts toward achieving healthcare equity.

8 Addressing Reviewers' Comments

8.1 Implemented Updates

In direct response to reviewer feedback—particularly comments noting that our front-end was underdeveloped and that visual evidence was lacking—we have added several concrete enhancements to the report while preserving our checkpoint-3 structure.

Comprehensive Metrics Table
Inspired by reviewers' requests to provide a full table of performance metrics for each class, Section 6.1 now embeds a detailed classification report showing precision, recall, F_1-score, and support for all four risk levels, alongside overall accuracy, macro average, and weighted average.

Feature Importance Visuals
Following requests to include visualizations of feature contributions, we inserted a bar chart of the top 15 feature importances and their numeric values. Each is captioned and explicitly referenced (e.g., "classes are depicted in Figure 2…") to strengthen our narrative flow.

Working UI Map Screenshot
To address feedback that the UI section is underdeveloped, we now showcase a working user-interface map in Section 6.2. This is a static, color-coded ZIP-code risk map (green through red) that clearly illustrates our dashboard's output. Although it is not yet interactive, this static map demonstrates that the underlying front-end logic is functional and visually aligned with our objectives. Reviewers' emphasis on "visual and demonstrative elements" directly inspired this inclusion.

Methodological Rationale
Responding to comments to "justify model choice," we expanded the "Model Training" subsection to articulate why random forest was selected over logistic regression, highlighting

RF's robustness to multicollinearity, ability to capture non-linear interactions, and superior handling of imbalanced classes.

Prose and Formatting Improvements
We conducted a thorough proofreading pass to standardize terminology (*risk_score*, *risk_level*, *risk_category*), corrected typos, and ensured uniform heading and caption styles.

8.2 Potential Future Enhancements

While our final report is necessarily constrained to completed coursework, several reviewer-sourced suggestions point to meaningful extensions beyond the deadline:

Interactive Dashboard Prototype: Building on static visuals, we could implement filters, hover-activated tooltips, and drill-down charts to bring the map to life, directly addressing calls for a "demonstrative UI walkthrough."

Baseline Model Benchmarking: Inspired by requests for "comparisons against simpler models," we might train and report logistic regression or decision-tree metrics side by side to quantify the random forest's performance gains.

Mathematical Formalization: To satisfy suggestions for "explicit formulas," we could add a concise indicator-function summation of our threshold-based risk scoring, improving methodological clarity.

Automated Update Pipeline: Reviewers' interest in "data update processes" motivates sketching an ETL workflow: scheduled CMS data pulls, preprocessing scripts, CI/CD triggers for retraining, and AWS Lambda redeployment to ensure model freshness in production.

These potential enhancements, driven directly by reviewer insights, represent valuable next steps for evolving this proof-of-concept into a fully interactive, benchmarked, and maintainable system.

References

Elizabeth J. Brown, Daniel Polsky, Cristina M. Barbu, Jeffrey W. Seymour, and David Grande. 2020. Patterns in geographic access to health care facilities across neighborhoods in the United States based on data from 2000 to 2014. *JAMA Network Open*, 3(5):e205105.

Anisa Butler, Kyla Covington, and Bridget Parsh. 2021. Identifying and tackling racial disparities in healthcare. *Nursing (Jenkintown, Pa.)*, 51(9):40–43.

Michelle M. Doty, Roosa Tikkanen, and Melinda K. Abrams. 2023. The cost of not getting care: Income disparities in the affordability of health care. Accessed: 2025-04-02.

Garth N. Graham. 2016. Why your zip code matters more than your genetic code: Promoting healthy outcomes from mother to child. *Breastfeeding Medicine*, 11:396–397. Epub 2016 Aug 11.

Jessica Ochs and others. 2024. 9.4 Geographical Disparities. OpenStax, Houston, Texas.

Michael G. Zywiel, Tiffany C. Liu, and Kevin J. Bozic. 2017. Value-based healthcare: The challenge of identifying and addressing low-value interventions. *Clinical Orthopaedics and Related Research*, 475(5):1305–1308.