Is Investment in Health Information Technology Influenced by Socioeconomic Factors?

By Hanyu Huang and Te Chuan (Vicky) Kuo

ITEC 6310

December 12, 2022

Introduction

The digital revolution is rapidly changing workflows in numerous industries. In most cases, the incorporation of digital infrastructure is associated with rising productivity (Sistrom, 2005, p. 435). Partly as a consequence, many industries have embraced this digital revolution. The healthcare industry is no exception to this trend to digitalize. In the United States, the *Health Information Technology for Economic and Clinical Health (HITECH) Act* established the roadmap for the American healthcare industry to integrate digital technology into its operations (U.S. Department of Health & Human Rights, 2009).

Despite its alleged benefits, the healthcare industry has been reluctant to embrace digital technology. A number of factors stymie the widespread adoption of health information technologies (HITs). Some of these reasons are the high cost of infrastructure, organizational resistance to digitalization, and technical shortfalls of the proposed systems (Cresswell & Sheikh, 2013, pp. 77–80). For these reasons, government intervention is crucial for the widespread adoption of HIT in the U.S. within a relatively short time span. Many healthcare providers would lack the incentive to purchase and deploy HIT infrastructure without some form of government intervention.

However, government intervention does not mean the results benefit everyone equally. In an era of badly planned infrastructure projects and rising economic inequality, it is possible that public investment decisions may cause more widening of the existing social fissures. Given the state of many disadvantaged populations, governments often need to pay especial attention to these groups so individuals from these groups can receive the same health outcomes as those from the more privileged groups. Otherwise, without this consideration for social equity and such affirmative actions, the disparities in health outcomes between different groups will likely widen. This widening of the existing disparities would place much-needed public services and goods even more out of the reach for the people who need them the most. This is especially important for a service as essential as healthcare.

This paper will explore one such relationship between government HIT infrastructure investment and social equity. In this paper, we look at the U.S. government's Electronic Healthcare Records (EHR) spending and whether it is equitable. We will compare the government subsidy spending for EHR implementation in the state of California, against the city incomes and expenditures. If the government spending is equitable, then cities with lower socioeconomic level populations should receive more financial assistance from the government to implement HIT.

Background

Electronic Health Records (EHR) Incentive Program

One of the most common of the new HITs being implemented is Electronic Healthcare Records (EHR). From both healthcare and administrative perspectives, EHR has several advantages over the previous dominant record-keeping method, which was to keep the patients' health records on paper. Digitalized EHR systems have lower transportation and storage costs compared to paper. With digitalization, patient records can be transferred quickly and efficiently to where it is needed, such as the hospital a patient is currently staying at (Chiu et al., 2015). In addition, EHR lengthens clinician-patient interactions. Longer interactions generally indicate improvement in the quality of care the clinicians provide (Poissant et al., 2005).

Despite these alleged advantages, the healthcare sector has been slow to embrace this technology. Researchers often point out the slow technology adoption rate in the healthcare industry (Cresswell & Sheikh, 2013; Sicotte et al., 1998). Cost continues to be one of the major barriers to the widespread adoption of HIT systems such as EHR. The high initial infrastructure cost and the low rate of financial returns is a significant deterrent for healthcare providers to invest in substantial HIT systems (Blumenthal, 2009). As a result, short of outside intervention or compulsion, the rate of EHR adoption would likely to have remained low.

Because of these reasons, the EHR Incentive Program for Medicaid and Medicare providers was established under the American Recovery and Reinvestment Act of 2009. As a part of the legislation, qualified Medicare and Medicaid practitioners and hospitals became eligible to apply for incentive payments starting in 2011 to help with purchasing, installing, and using electronic health information systems in their work processes. At least partially because of these incentives, the number of healthcare providers that have adopted some form of EHR system has greatly increased. Whereas in 2009 in the U.S. only around 17% of doctors and 10% of the hospitals had EHR systems; by 2015, around 84% of U.S. hospitals had adopted an EHR system.(Kuhn et al., 2010; *Office-Based Physician Electronic Health Record Adoption*, n.d.)

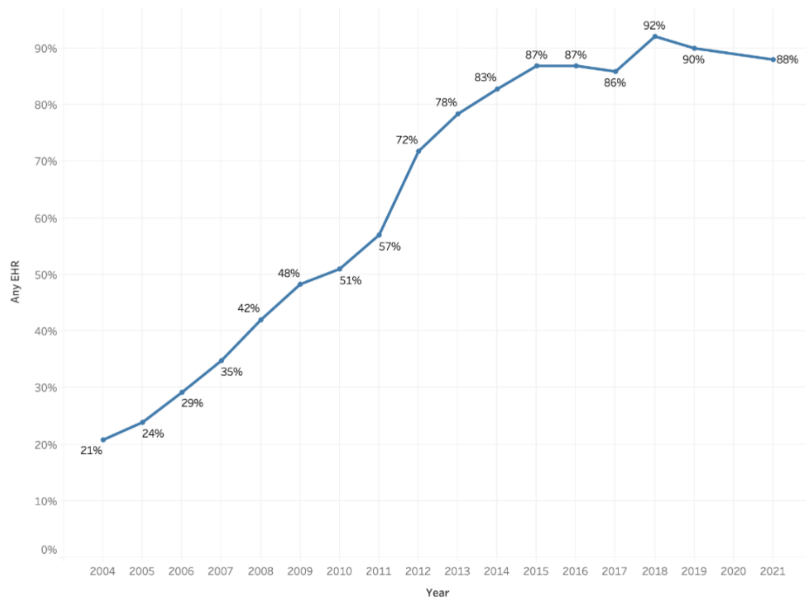


Fig.1: EHR uptake in the United States

The EHR Incentive Payments program introduced a graduate incentive payment structure. The general idea for the payment structure is to reward the healthcare providers who had started to implement EHR infrastructure early and penalize the latecomers. Under this structure, early adopters that have implemented EHR infrastructure and have demonstrated "meaningful use" can receive payments of up to $44,000 over five years. Latecomers that only become eligible from 2013 onwards would receive less amounts the later they start. Finally, healthcare providers who have not started to apply for EHR incentive funding by 2015 would not receive any subsidies at all (*CMS FINALIZES REQUIREMENTS FOR THE MEDICARE ELECTRONIC HEALTH RECORDS (EHR) INCENTIVE PROGRAM | CMS*, 2010).

		Starting Year				
Payment Year		2011	2012	2013	2014	2015
	2011	$18,000	---	---	---	---
	2012	$12,000	$18,000	---	---	---

	2013	$8,000	$12,000	$15,000	---	---
	2014	$4000	$8000	$12,000	$12,000	---
	2015	$2,000	$4,000	$8,000	$8,000	$0
	2016	---	$2,000	$4,000	$4,000	$0
	Total	$44,000	$44,000	$39,000	$24,000	$0

Fig. 2: the EHR payment structure

Despite this schedule, it is evident that CMS could not match this schedule. In 2015, the EHR Incentive program was extended to 2017 and beyond (CMS, 2022). Furthermore, EHR payments continued to be paid out at least up to 2020. The EHR Incentive program itself has since changed its name to the Promoting Interoperability program (CMS, 2022).

Technology and Inequality

All major technological changes almost inevitably create new winners and losers. There will always be groups better able to take advantage of changes. For instance, one study published in 2011 pertaining to the usage of HIT among the general American population found that whites, women, younger patients, English speakers, and the insured are more likely to use patient portals to access their health records than other groups (Ancker et al., 2011). Even when advanced methods such as machine learning (ML) are employed, biases still appear that affect population groups differently. A research paper from 2020 studying the use of ML to assist with booking patients with clinicians found that the booking algorithm often placed the poor and the uninsured patients into already overbooked slots to maximize “efficiency” (Samorani et al., 2021). As a result, the implementation of new technologies can lead to the deepening of existing disparities, and the creation of new ones, if such consequences are not taken into considerations in the design phase.

Literature Review

Unfortunately, there is little literature that directly address the relationship between EHR incentive payments and the socio-economic positions of the different demographic groups. As a result, we must use literature that is more tangential to our problem statement to provide context. The literature reviewed can be broken down into three main categories. These categories are:

1. The current state of HIT and lessons learned from previous implementations.
2. The effect of government investment on social inequality.
3. The impact of technology on equity and inequality.

Lessons Learned from Previous HIT Implementation

The literature regarding the state of HIT investment centers on drawing lessons from previous implementations. One of the main takeaways from the literature is the difficulty of gaining acceptance from healthcare providers for HIT. The difficulty of incorporating HIT into healthcare workflows is reported in *Cresswell & Sheikh (2013)*, *Frisse (2009)*, and *Gagnon et al. (2009)* (Cresswell & Sheikh, 2013; Frisse, 2009; Gagnon & Gravel, 2009). The difficulties lie with inserting HIT into a largely labor-intensive industry, and the conflicting incentives that often arise between the different management levels. The *Sicotte et al. (1998)* study found that often, the adoption of HIT also meant the imposition of new ways of working, and the divergence of interest between managers and front-line healthcare providers. This divergence often results in antagonistic relationship between the frontline healthcare providers, and their managers who try to impose the new technologies and performance metrics that were developed without input from those affected (Sicotte et al., 1998, pp. 441–443). As a result, the healthcare industry is in many ways intrinsically resistant to HIT. Many healthcare providers see HIT as the imposition of technocrats who lack understanding regarding how healthcare provision work in practice.

Another major theme of the HIT literature is the financial barriers of HIT adoption. Two of the biggest financial hurdles for HIT adoption are the high capital cost for the initial purchase of the HIT infrastructure, and the uncertain rates of return from these investments. *Blumenthal (2009)*, *Bower*, *(2005)*, and *Sistrom (2005)* all emphasized the difficulties of getting healthcare providers to invest in HIT due to the lack of financial incentives. *Bower (2005)* found the rate of IT investment in the healthcare industry is less than half of those in other industries (Bower, 2005, p. 2). *Blumenthal (2009)* claimed the high financial and technical barriers was a significant deterrent for healthcare providers (Blumenthal, 2009, p. 28). Finally, *Sistrom (2005)* explained that economically, labour intensive sectors all experience difficulties in "industrializing" and lower their costs. *Sistrom (2005)* supported this assertion with economic data that showed healthcare was one of four industries that had demonstrated little benefit from the growth of information technology (Sistrom, 2005, p. 435). Given these financial and economic barriers, external forces would be needed to push healthcare providers to adopt the not-so-profitable HIT offerings. Often, this external force comes in the form of the government and its regulators either offering incentives or requiring the actors to invest to meet compliance goals.

Does government investment increase inequality?

Some economists believe governments should take a lead role in incentivizing infrastructure investment when the private sector appear to be incapable of doing so (Boyle & Clarine, 2022). Compared to private businesses, public investments are in a unique position of not being driven by the need to maximize its rate of return. Additionally, governments also have the power of compulsion through legislation. Governments can leverage these resources to force private businesses to adopt unprofitable practices and technologies that society nevertheless deems to be beneficial, such as HIT.

However, in some circumstances, public investment may not be the remedy to market failures. Governments are political institutions. Governments may make investment decisions due to political considerations, such as rewarding their supporters, rather than measuring social benefits and costs accurately. Due to the large variation in governments across the world, there is no one-size fit all answer for whether public investment is beneficial for social equity. Furthermore, studies on government decision-making regarding infrastructure is hard to find. One of these case studies on how governments make infrastructure investment decisions, *Castells & Sole-Olle (2005)*, studied transportation infrastructure investment in Spain. The researchers found that most of the infrastructure investment decisions were made based on economic considerations (Castells & Solé-Ollé, 2005, p. 32). This finding suggests that it is economic, not political, considerations to be the biggest determinant in government infrastructure spending. While this is good news in that governments may be less susceptible to political pressure than one feared, economic efficiency may not necessarily be socially equitable. Public services may, intentionally or unintentionally, reduce social equity and increase inequality due to access and distribution issues.

The literature surrounding public infrastructure investment and its relationship to social equity and inequality suggests there may be social equity problems associated with public infrastructure provisioning by governments. *Chatterjee & Turnovsky (2012)* argued that public investment will always cause an increase in income inequality over time regardless of how the investment is financed. *Lustig (2015)'s* survey of 13 developing countries found more divisive results. Healthcare spending in some of the countries, such as Chile, Colombia, and Uruguay, benefit the poorest 20% of the population. However, healthcare spending in some of the other countries surveyed, including El Salvador, Ethiopia, and Guatemala; benefit the richest 20% of their population (Lustig, 2015, p. 313). Therefore, there is evidence that government spending, which may improve living standards overall, may create social inequity in the long run. Without policies mitigating inequity, essential services such as healthcare would become out of the reach of the poor and those who need it most.

What is the impact of technology on inequality?

The disruptions that occur during technological revolutions also creates new inequalities. The increasing adoption of HIT is no exception to this. Literature suggests many of the new HIT being implemented have unequal impact on their users. For instance, *Ahmed et al. (2020)* in their investigation of electronic health resources in Bangladesh found that despite the increasing electronic device coverage, not everyone benefitted equally from the resources' availability. The rural population lags behind their urban counterparts in using electronic health resources. Furthermore, young people, which is the demographic that is most likely to use electronic health resources, are more likely to have their concerns dismissed or unable to access the resources (Ahmed et al., 2020, pp. 9–10). As a result, HIT interventions are less effective than they can be due to the mismatch between those who have access to technology and those who can best make use of it.

Such outcome disparities different population groups are not limited to the developing world. Studies such as *Craig et al (2021)*, and *Senteio et al. (2022)* found that HIT implemented in the United States also showed signs of inequity when dealing with marginalized populations. *Craig et al. (2021)* found general equity problems with HIT as implemented today. For example, marginalized populations are less likely to make use of health information portals (Craig et al., 2021, pp. 2–4), and scheduling software trained on machine learning were found to discriminate against patients from lower socioeconomic backgrounds by booking these patients into already overbooked slots (Craig et al., 2021, p. 5). The literature review by Senteio et al. (2022) looked at HIT interventions in black and Hispanic populations in the U.S. It found persistent disparities in chronic disease outcomes. This disparity stems from HIT design decisions and social influence (Samorani et al., 2021, p. 10; Senteio et al., 2022, p. 2825). Despite efforts to mitigate the effects of socioeconomic inequality, HIT and their implementers directly or indirectly benefit some population segments more than others. Whether it is in the developing or the developed economies, the equity effect of HIT need to be taken into consideration.

Data Collection

Our research is conducted using secondary data collected by other institutions. Our data regarding the EHR payments made to healthcare provided was published by the *California Department of Health Care Services*. The data comes in two sets, one set for the individual healthcare providers, and the other dataset for hospitals. The dataset covers the years 2012 to 2020.

Our data regarding the Californian cities' finances also comes in two datasets. One of the datasets covers revenue and the other one for expenditures. The data is collected by the *California State Controller's Office*. The datasets contain the total revenues and expenditures for cities in California from the years 2003 to 2021. Additionally, the datasets contain the estimated population of each city for each of the fiscal years, which allowed us to calculate the estimated EHR payments per capita.

We used the overlapping period from 2012 to 2020 as our period of study. One characteristic of the EHR datasets to keep in mind is that EHR subsidies are not paid out every year to every city. Therefore, it is possible for a city to have received its last EHR payment in a year prior to 2020. For example, a city may have received it its last EHR payment in 2016, and never received any subsequent payments. When such situations arise, we assume all eligible healthcare providers in that city have received all eligible EHR subsidies in 2016. All the payments the city received up to 2016 constitute the total payments received. When we compared the correlation between EHR per Capita and the independent variables, we compared total EHR per capita with the city's revenue and expenditures of the last year of payment. Therefore, if a city received its last EHR subsidy payment in 2016, we would compare the total EHR subsidies received by 2016 with the city's 2016 revenue and expenditures. If a city received its last HER payment in 2017, we would compare the EHR per capita of the city in 2017 with its revenue and expense in that year, and so forth.

Definition of Variables

Our dependent variable is the EHR spending per capita (*EHR per Capita*), measured in United States dollars (USD). This variable was intended to measure the government subsidies given to each community when broken down by the population size of that community. The decision to use *EHR per Capita* rather than the gross EHR for each city was because we believe using *per Capita* better reflects the investment's equity in each community. For example, if two cities both receive one million USD in EHR subsidies, but one city has a population of one million and the other city has a population of 500,000, then the latter city's residents effectively received double the EHR subsidies payment from the government compared to the former city's residents. For this reason, we believe using the per capita measurement of the variable is a better measurement of the social equity of EHR investment.

Revenue per Capita and *Expenditure per Capita* are our variables for measuring socioeconomic status of each city in California. *Revenue per Capita* and *Expenditure per Capita* are the best measurements of how affluent a community that is are publicly available. We believe the *Revenues* and *Expenditures* per capita reflect their community's affluence because, in most cases, the more prosperous a community is, the more tax revenue a municipal government is able to draw on. Therefore, the higher these two variables are when measured, the higher the affluence, and the higher up the socioeconomic ladder the average resident of the community should be.

However, our choice of *Revenue per Capita* and *Expenditure per Capita* is not ideal. Our choice in using these variables were constrained by the available of the secondary datasets. Ideally, other measures such as income per capita, education levels, or other more direct measurements of human development levels would be more preferable. However, we were unable to locate publicly available versions of these demographic data. Given the absence of these data, we believed the *Revenues* and *Expenditures per Capita* measures are the best approximate measures we have access to.

When it comes to equity, ideally the lower down the socioeconomic ladder a community is, the more money the government should be spending. This is because generally the communities lower down the socioeconomic ladder are already disadvantaged and may require additional resources from external sources to reach the same level of achievement as their counterparts from more affluent communities. If a poorer community receives less funding per capita from the government compared to wealthier communities, then we can term the distribution *regressive*. If the policy is regressive, then higher *revenue* and *expenditure* per capita would be positively correlated with higher *EHR per Capita* spending. On the other hand, if the government spending per capita is higher in the poorer communities compared to the wealthier ones, then the spending can be termed *progressive.* In this case, higher *revenue* and *expenditure* per capita would be correlated with lower *EHR per Capita*.

Proposed Hypothesis

For our research project, we have two null hypotheses. These hypotheses are:

- $H0_1$: EHR incentive payments to *healthcare providers* is not affected by a community government's revenues and expenditures.
- $H0_2$: the EHR incentive payments to *hospitals* is not affected by a community government's revenues and expenditures.

Research Design

Data Transformations

We used the log-transformed dataset for our analysis. The primary reason for transforming this data is due to reduce the skewness of the distribution. Our initial analysis indicates the raw data's distribution is heavily positively skewed for our independent and dependent variables. By transforming our data, we hoped the result would be more amiable to analysis and increase our models' predictive power. Please refer to Appendix A for the distribution histograms of before and after the transformation.

Machine Learning Algorithms

We aim to determine the correlation of *EHR per Capita* with the per capita *Revenues* and *Expenditures* of cities in California by using popular machine learning algorithms (MLAs). We hope, through the MLAs, that we can develop a model that can predict *EHR per Capita* using the available socioeconomic indicators of a community. The MLAs used in this study are linear regression (LR), decision tree regression (DTR), random forest regression (RFR), support vector machine (SVM), random sample consensus (RANSAC), and AdaBoost algorithm.

Validation of the MLAs was performed with statistical indices such as coefficient of determination (R^2), mean absolute error (MAE), and mean square error (MSE). R-squared is the most common measure of the correlation between variables. MAE and MSE are different measures of the error of a model. These measures together can inform us how accurate a model is once we have trained the algorithm using our training data.

Compared to purely statistical linear regression method, the machine learning models are theoretically more resistant to collinearity, outliers, or data overfitting. We chose the decision tree algorithm because it mimics human thinking. This makes the decision tree algorithm a "white box" algorithm, which means there is no need for significant data preprocessing such as feature selection, regularization, and multi-collinearity. White box algorithms typically require less time to train the model than other algorithms such as neural networks. Therefore, this algorithm and its derivative, random forest algorithm, are both proposed in this study.

Regarding the outliers, SVM handles those better than LR as it will only use the most relevant points (support vectors) to find a linear separation. RANSAC is another type of robust estimator which randomly selects several data points from the data set to make a model robust to the outliers. On the other hand, Adaboost is less susceptible to overfitting than other algorithms and also has the benefit of being easy to implement and there are no parameters need to be adjusted. Therefore, we hope this barrage of algorithms can offset each other's weaknesses arrive at a more accurate answer for our research question.

Data Analysis

$H0_1$: Individual Healthcare Providers

Descriptive Statistics

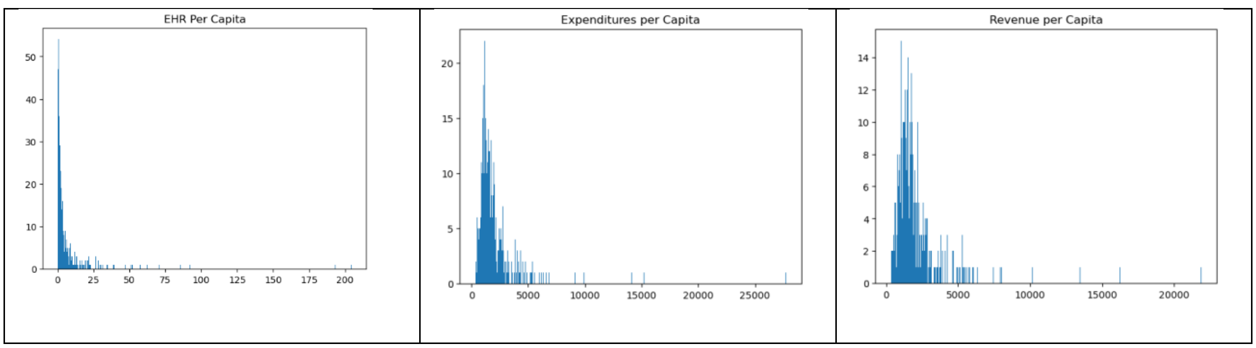


Fig.3: The Preliminary histograms

The dataset for *EHR* spending for individual healthcare providers contains 404 datapoints for cities across the state of California. When we plotted out the frequency and distribution of the data, the result is all three variables' distributions are positively skewed. This means most of the observations are towards the lower end of the income spectrum.

Data Analysis

At the first glance, the three variables seem to be well correlated with each other. The majority of the observation points seemed to appear approximately in the same places in their distribution. This distribution seems to have indicated some form of correlation.

	Statistic	P-Value
EHR per Capita vs. Revenue per Capita	0.078	0.116
EHR per Capita vs. Expenditures per Capita	0.039	0.439

Fig.4: Spearman Rank Test

However, our Spearman Rank test result soon confirms that there is little correlation between our independent and dependent variables. Individually, both *Revenue per Capita and Expenditure per Capita* have p-values that are significantly higher than the alpha value of 0.001 required for the difference to be statistically significant.

Ordinary Least Squared Regression

OLS Regression Results			
Dep. Variable:	y	R-squared:	0.018
Model:	OLS	Adj. R-squared:	0.014
Method:	Least Squares	F-statistic:	3.778
Date:	Sat, 19 Nov 2022	Prob (F-statistic):	0.0237
Time:	20:52:31	Log-Likelihood:	-700.31
No. Observations:	404	AIC:	1407.
Df Residuals:	401	BIC:	1419.
Df Model:	2		
Covariance Type:	nonrobust		

	Coef	Std err	t	P>|t|	[0.025	0.975]
Inter-cept	4.9251	1.406	3.502	0.001	2.161	7.690
x	0.0046	0.0046	0.0046	0.051	-2.29e-05	0.009
z	-0.0033	0.002	-1.519	0.130	-0.008	-0.008

Omnibus:	538.162	Durbin-Watson:	1.942
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59066.181
Skew:	6.578	Prob(JB):	0.00

Kurtosis:	60.756	Cond. No.	6.17e+03

Fig.5: OLS Regression Report for individual healthcare providers

The result of our Ordinary Squared Regression results also suggests a lack of correlation between our variables. The R-squared value of the model is exceptionally low, at only 0.018. This means that even with our data transformation, the model cannot explain most of our observations.

The model's lack of predictive power is also reflected in the analysis as well. Both *Revenues per Capita* (x) and *Expenditures per Capita* (z) have p-values that are significantly higher than the 0.05 alpha value required to meet the minimum requirement for statistical significance. Furthermore, even if the finding is statistically significant, the Pearson correlation coefficient suggests there is no substantive significance. At best, a one percent increase in either of the variables would cause a less than one percent change in *EHR per Capita*. If anything, the correlation between *Expenditures per Capita* and *EHR per Capita* is negative. If this correlation had been significant, then it would have suggested a substitution effect where higher expenditures results in less money being spent on healthcare.

Outlier treatment

To make the distribution look more symmetric and more normalised, the log transformation technique is used to improve the positively skewed variables.

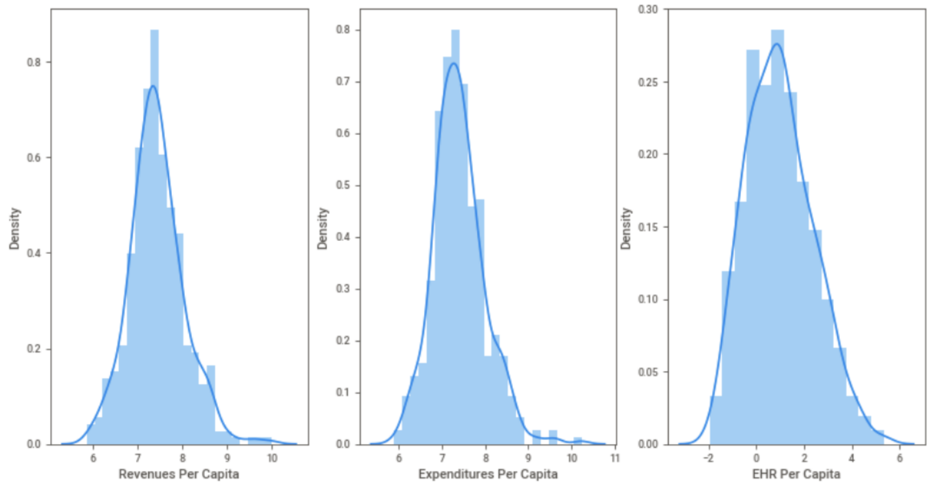


Fig.6: Features distribution plot after the log transformation

Machine Learning Models

Our ML models confirmed our initial observations regarding the lack of correlation between the variables. None of our ML models produced a model with enough predictive power to be considered reliable. Even when using 70% of our available data as training data, none of the ML algorithms produced a model that can account for any substantial portion of the observations. Notably, the Decision Tree Regression and Random Forest Regression algorithms were overfitted. Despite performing very well on the training data, their models too did not do well when applied to the testing data.

Due to the inability of the models to demonstrate any form of correlation, we cannot reject the null hypothesis that *EHR per Capita* is not affected by *Revenues* and *Expenditures* per capita.

Training Data

	Linear Regression	Decision Tree Regression	Random Forest Regression	Support Vector Machine	RANSAC Regression	Stochastic Gradient Descent	Adaboost Algorithm
R2	0.027	1.00	0.82	0.04	-0.02	0.02	0.11
MAE	1.78	0.00	0.32	1.75	1.85	1.80	1.62
MSE	1.07	0.00	0.46	1.04	1.11	1.08	1.05
RMSE	1.33	0.00	0.57	1.32	1.36	1.34	1.27

Testing Data

	Linear	Decision Tree	Random Forest	Support Vector	RANSAC	Stochastic	Adaboost

	Regression	Regression	Regression	Machine	Regression	Gradient Descent	Algorithm
R2	-0.00	-1.22	-0.34	-0.00	-0.06	0.01	-0.02
MAE	2.11	4.67	2.82	2.11	2.24	2.09	2.14
MSE	1.19	1.70	1.33	1.16	1.23	1.17	1.17
RMSE	1.45	2.16	1.68	1.45	1.5	1.45	1.46

Fig 7: ML outcomes for individual healthcare providers

H02: Hospitals

Descriptive Statistics

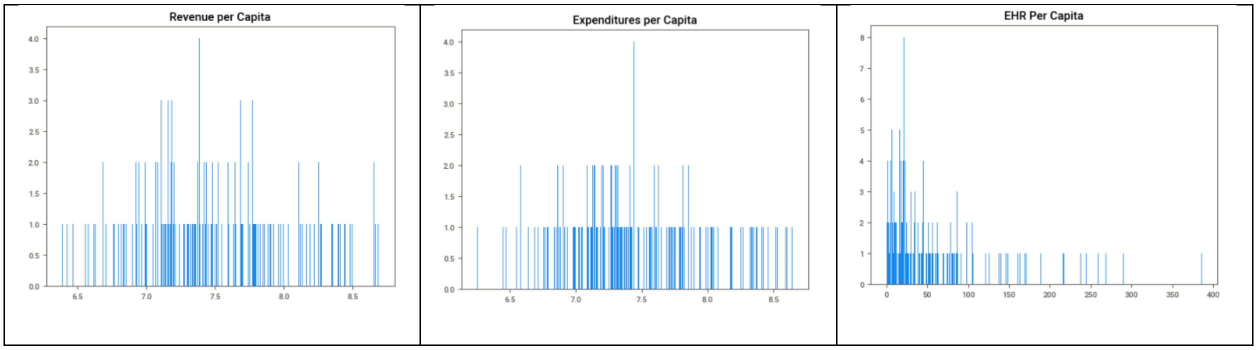


Fig.8: The Preliminary histograms

The dataset of hospitals provides significantly fewer observation points. Between the years 2013 and 2020, only 169 communities have received EHR payments. The significantly lower number of observation points makes sense for hospitals. As larger institutions, hospitals are more likely to be present only in large population centers. However, because they are large institutions, hospitals are also more likely to benefit from the economics of scale that information technology systems provide. Therefore, hospitals are possibly more willing to adopt HIT that can manage large amount of data more quickly than traditional systems. We can see more financial resources involved when it comes to hospitals.

Like our analysis with individual healthcare providers, we plotted out the frequency observations of our variables in histograms. However, in the case of hospitals, the independent variables, *Expenditures per Capita* and *Revenues per Capita*, appear to have a more normal distribution. *EHR per Capita* however continue to show a heavily positively skewed distribution.

Data Analysis

	Rho	P-Value
EHR Per Capita vs. Revenue Per Capita	-0.078	0.309
EHR Per Capita vs. Expenditures Per Capita	-0.084	0.277

Fig.9: Spearman Rank Test

Similar to what we had done with the individual healthcare providers, we also ran a Spearman Rank test with each independent variable against the dependent variable. The result is also similar to what we found with the Spearman Rank test for individual healthcare providers. Namely, the p-value is significantly higher than the distribution's alpha level. This indicates the independent variables do not appear to have significant effects on the independent variable.

Ordinary Least Squared Regression

OLS Regression Results			
Dep. Variable:	y	R-squared:	0.010
Model:	OLS	Adj. R-squared:	-0.006
Method:	Least Squares	F-statistic:	0.6457
Date:	Mon, 05 Dec 2022	Date:	Mon, 05 Dec 2022
Time:	22:18:11	Log-Likelihood:	-701.76
No. Observations:	126	AIC:	1410.
Df Residuals:	123	BIC:	1418.
Df Model:	2		
Covariance Type:	nonrobust		

	Coef	Std err	T	P>|t|	[0.025	0.975]
Intercept	37.8604	19.026	1.990	0.049	0.199	75.522
x	-0.0242	0.035	-0.693	-0.693	-0.693	0.045
z	0.0363	0.038	0.948	0.948	-0.040	0.112

Omnibus:	77.629	Durbin-Watson:	1.818
Prob(Omnibus):	0.000	Jarque-Bera (JB):	326.708
Skew:	2.292	Prob(JB):	1.14e-71
Kurtosis:	9.421	Cond. No.	7.36e+03

Fig.10: OLS Regression Report

The results of our Ordinary Least Squared regression for hospitals are, again, similar to the results for individual healthcare providers. Once again, the R-squared value is very low. This means the model has low explanatory power. For both the *Revenues per Capita* and *Expenditures per Capita* variables, their p-value is significantly higher than the alpha level and their Pearson correlation coefficients are also substantively insignificant.

Outlier treatment

Same as the outlier treatment for the provider dataset, the log transformation technique is also used to reduce the amount of right-skewness so as to get grater symmetry.

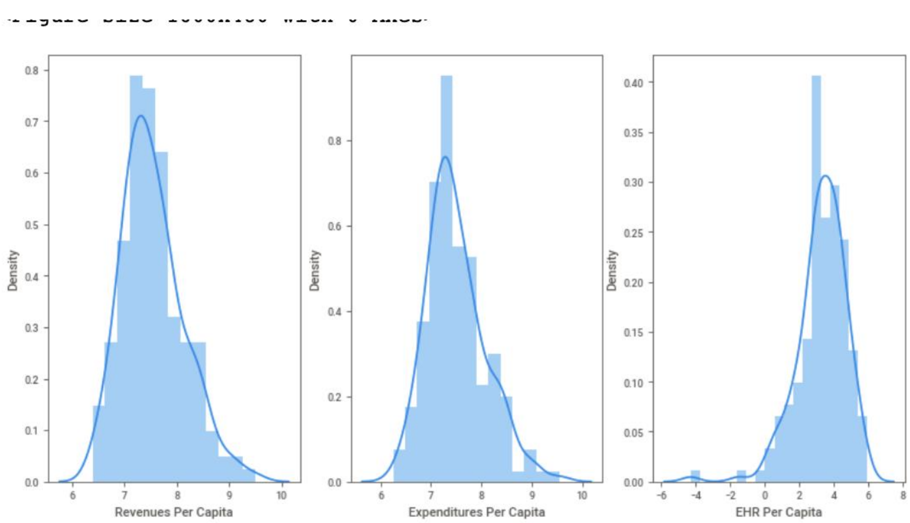


Fig.11: Features distribution plot after the log transformation

Machine Learning Models

Similar to the provider's data result, our first observations about the lack of association between the variables were supported by our ML models. None of our ML models have the ability to generate testable predictions when using 70% of the data for training, and the remaining 30% of the data for testing. Also, the Decision Tree Regression and Random Forest Regression algorithm still witnessed the phenomenon of overfitting.

To conclude, the null hypothesis that EHR per capita is unaffected by Revenues and Expenditures per capita is not rejected. The models showed the negative relationships between the built models and the dependent variables. Therefore, we cannot conclusively demonstrate there is any causal relationship between *EHR per Capita* and *Revenues per Capita* or *Expenditures per Capita*.

Training Data

	Linear Regression	Decision Tree Regression	Random Forest Regression	Support Vector Machine	RANSAC Regression	Stochastic Gradient Descent	Adaboost Algorithm
R2	0.00	0.99	0.84	0.02	0.00	-0.04	0.47
MAE	1.08	0.04	0.44	1.05	1.08	1.12	0.91
MSE	2.17	0.02	0.34	2.13	2.17	2.26	1.16
RMSE	1.47	0.12	0.59	1.46	1.47	1.50	1.07

Testing Data

	Linear Regression	Decision Tree Regression	Random Forest Regression	Support Vector Machine	RANSAC Regression	Stochastic Gradient Descent	Adaboost Algorithm
R2	-0.43	-1.51	-0.82	-0.30	-0.44	-0.60	-0.59
MAE	1.05	1.48	1.19	1.00	1.06	1.15	1.06
MSE	1.72	3.00	2.18	1.55	1.72	1.92	1.90

RMSE	1.31	1.73	1.48	1.25	1.31	1.38	1.38

Fig 12: ML outcomes for individual healthcare hospitals

Validity Results

Internal Validity

There are several factors that affect our analysis' internal validity. Our first concern is the span of time the dataset covers and the changes in government policy in between the start and end dates. Another concern is how accurately our measurements reflect what we intend to measure, which is the socioeconomic status of the population, and how it addresses the problem we have identified in our research question.

Our dataset covers the *EHR* payments to Californian healthcare providers between the years 2011 to 2020. During that time, the *EHR* incentive program had underwent several changes. Examples of the changes include adding a third stage that extended the program, changes to reporting requirements, and even the program's name to "Promoting Interoperability" to reflect the new program goal (CMS, 2022). Outside of these publicized changes from CMS, there would have been changes in government and subsidy policies from the U.S. federal government legislations all the way down to the municipal bylaws. The data was unable to take any of these changes into account. Consequently, effects caused by these extraneous factors can cause skews the results and hide the individual effects that socioeconomic factors have on *EHR* payments.

Another factor that may compromise our analysis' internal validity is the degree to which our independent and dependent variables measure social equity and the population's socioeconomic status. We intended the *EHR per Capita* variable to measure the equity of government subsidies. The *Revenues* and *Expenditures* per capita should be the measures for the socioeconomic status of the affected populations. While important, EHR infrastructure is only one small part of the much larger HIT infrastructure project the U.S. is building. EHR is also not a front-end system, meaning is intended for use by healthcare providers and not directly accessible by the general population. Therefore, the social equity impact in relation to the EHR system is limited. EHR systems does not directly interface with the general population. It is also fixed infrastructure where flexibility and scalability matter less.

Our concern with the variables *Revenues* and *Expenditures* per capita lies in whether they can accurately represent socioeconomic status of the studied population. Both *Revenue* and *Expenditures* are impacted by tax policies and the different sources of revenue and commitments a regional government may have. In most cases, revenues and expenditures should be reflective of a community's prosperity. The largest source of income for local governments is property taxes (Miller et al., 2018). It is logical that more prosperous areas would have more expensive property, and so higher tax revenue and higher expenditures. On the other hand, it is entirely plausible that different cities would have different tax policies and dissimilar sources of revenue that is not reflective of the socio-economic levels of their residents. Hypothetically, the revenue stream of a mining town can be very different from that of an agricultural settlement, which in turn would be very different from that of a large metropolis such as San Francisco. Compounding this difficulty would be other considerations such as the cost of living in the different municipalities. These problems mean we cannot draw equivalence between municipal revenues, expenditures, and the needs of their residents, as we would like. Therefore, the average resident of a high-revenue, high-expenditure city may not necessarily be better off than the average resident of another city that appear have fewer financial resources.

External Validity

Our concern regarding our test's external validity mainly lies with whether the results are applicable to other American states outside of California, and what our findings mean for populations living there. Different American states would have different laws. While the number of subsidies paid to healthcare providers was set by CMS as far as we can tell, it is also clear that each state's healthcare administration also has roles to play in recording and disbursing the funds. Therefore, while our analysis looked at the effects of socioeconomic indicators and EHR payment in California, we are unsure how applicable our results are to the other American states, and even other countries.

Conclusion & Recommendations

The results of our analysis suggest that *revenue per capita* and *expenditure per capita* have no impact on the *EHR per capita* per city. The lack of correlation between our dependent and independent variables means that *EHR* subsidies spending is not influenced by a community's wealth and affluence. Both statistical and ML methods have not found any correlation between the dependent and independent variables. The models themselves all have low predictive power with high error rate. Therefore, with the data available, we can conclude that EHR subsidies paid out to healthcare providers or hospitals are not affected by the level of affluence of the community.

While our investigation has not identified any causal link, conjecture regarding any correlation between HIT and socioeconomic factors should not be entirely dismissed. There is sufficient literature to suggest that there is inequality when it comes to healthcare spending and HIT infrastructures. The disruptive effects of HIT on healthcare providers and users is well documented (Sicotte et al., 1998). Therefore, we would recommend further investigations on the equity effects of HIT and how to mitigate them, so marginalized populations will not be further disadvantaged in accessing this essential service and exercising their rights.

References

Ahmed, T., Rizvi, S. J. R., Rasheed, S., Iqbal, M., Bhuiya, A., Standing, H., Bloom, G., Waldman, L., & others. (2020). Digital health and inequalities in access to health services in Bangladesh: Mixed methods study. *JMIR MHealth and UHealth*, *8*(7), e16473.

Ancker, J. S., Barrón, Y., Rockoff, M. L., Hauser, D., Pichardo, M., Szerencsy, A., & Calman, N. (2011). Use of an electronic patient portal among disadvantaged populations. *Journal of General Internal Medicine*, *26*(10), 1117–1123.

Blumenthal, D. (2009). Stimulating the adoption of health information technology. *West Virginia Medical Journal*, *105*(3), 28–30.

Bower, A. G. (2005). Federal investment in health information technology: How to motivate it? *Health Affairs*, *24*(5), 1263–1265.

Boyle, M. J., & Clarine, S. (2022, December 7). *Market Failure: What It Is in Economics, Common Types, and Causes*. Investopedia. https://www.investopedia.com/terms/m/marketfailure.asp

Castells, A., & Solé-Ollé, A. (2005). The regional allocation of infrastructure investment: The role of equity, efficiency and political factors. *European Economic Review*, *49*(5), 1165–1205.

Chiu, W.-T., Hsu, M.-H., & Teng, S.-W. (2015). Using health information technology to reduce regional health inequality in Taiwan. *J. Formos. Med. Assoc*, *114*, 1–2.

CMS. (2022). *Promoting Interoperability Programs*. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms

CMS FINALIZES REQUIREMENTS FOR THE MEDICARE ELECTRONIC HEALTH RECORDS (EHR) INCENTIVE PROGRAM | CMS. (2010, July 16). Centers for Medicare & Medicaid Services. https://www.cms.gov/newsroom/fact-sheets/cms-finalizes-requirements-medicare-electronic-health-records-ehr-incentive-program

Craig, S., McPeak, K., Madu, N., Dalembert, G., & Wasserman, R. C. (2021). Health information technology and equity: Applying history's lessons to tomorrow's innovations. *Current Problems in Pediatric and Adolescent Health Care*, 101110.

Cresswell, K., & Sheikh, A. (2013). Organizational issues in the implementation and adoption of health information technology innovations: An interpretative review. *International Journal of Medical Informatics*, *82*(5), e73–e86.

Frisse, M. E. (2009). Health Information Technology: One Step At A Time. *Health Affairs*, *28*(2), w379-.

Gagnon, MP, Légaré, F, Labrecque, M, Frémont, P, Pluye, P, Gagnon, J, Car, J, Pagliari, C, Desmartis, M, Turcot, L., & Gravel, K. (2009). Interventions for promoting information and communication technologies adoption in healthcare professionals. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org/10.1002/14651858.CD006093.pub2

Kuhn, A., Gaiser, T., & Gandonou, E. (2010). Simulating the effects of tax exemptions on fertiliser use in Benin by linking biophysical and economic models. *Agricultural Systems*, *103*(8), 509–520. https://doi.org/10.1016/j.agsy.2010.05.003

Lustig, N. (2015). Chapter 16. The Redistributive Impact of Government Spending on Education and Health: Evidence from Thirteen Developing Countries in the Commitment to Equity Project. In *Inequality and Fiscal Policy* (p. ch016). International Monetary Fund. https://doi.org/10.5089/9781513567754.071.ch016

Miller, R., Chu, V., Chu, C., Garosi, J., Kerstein, S., Uhler, B., & Weatherford, B. (2018). *California's Tax System: A Visual Guide* [LAO Report]. Legislative Analyst's Office. https://lao.ca.gov/reports/2018/3805/ca-tax-system-041218.pdf

Office-based Physician Electronic Health Record Adoption. (n.d.). [Government Website].

HealthIT.Gov. Retrieved November 12, 2022, from https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption

Poissant, L., Pereira, J., Tamblyn, R., & Kawasumi, Y. (2005). The Impact of Electronic Health Records on Time Efficiency of Physicians and Nurses: A Systematic Review. *Journal of the American Medical Informatics Association : JAMIA*, *12*(5), 505–516. https://doi.org/10.1197/jamia.M1700

Samorani, M., Harris, S. L., Blount, L. G., Lu, H., & Santoro, M. A. (2021). Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management*.

Senteio, C., Murdock, P. J., & others. (2022). The Efficacy of Health Information Technology in Supporting Health Equity for Black and Hispanic Patients With Chronic Diseases: Systematic Review. *Journal of Medical Internet Research*, *24*(4), e22124.

Sicotte, C., Denis, J. L., & Lehoux, P. (1998). The Computer Based Patient Record: A Strategic Issue in Process Innovation. *Journal of Medical Systems*, *22*(6), 431–443. https://doi.org/10.1023/A:1020674527409

Sistrom, C. (2005). The socioeconomic aspects of information technology for health care with emphasis on radiology1. *Academic Radiology*, *12*(4), 431–442.

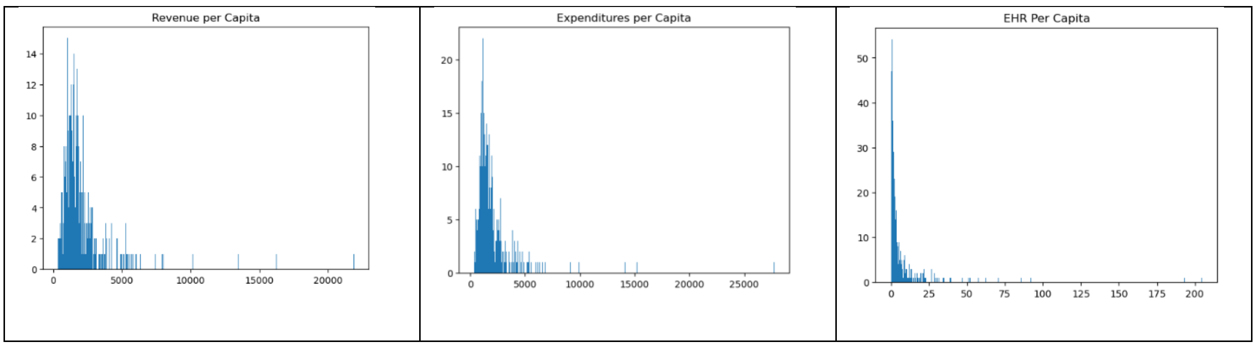
U.S. Department of Health & Human Rights. (2009, October 28). *HITECH Act Enforcement Interim Final Rule* [Text]. HHS.Gov. https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.html

Appendix A – Descriptive Statistics of Variables

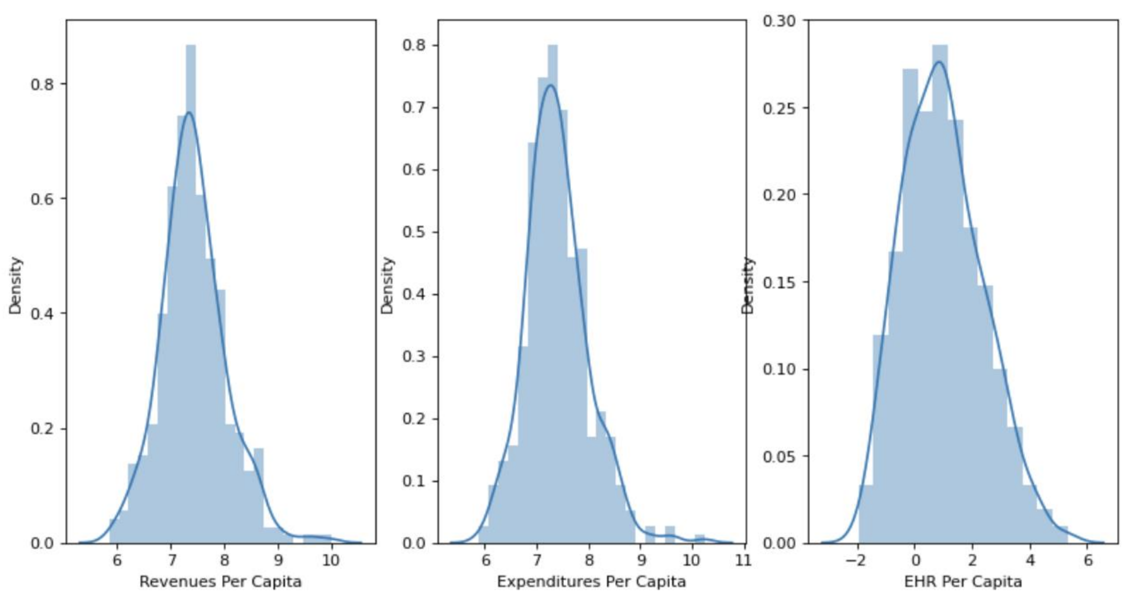
Data Transformation Histograms

Individual Providers

Before Transformation

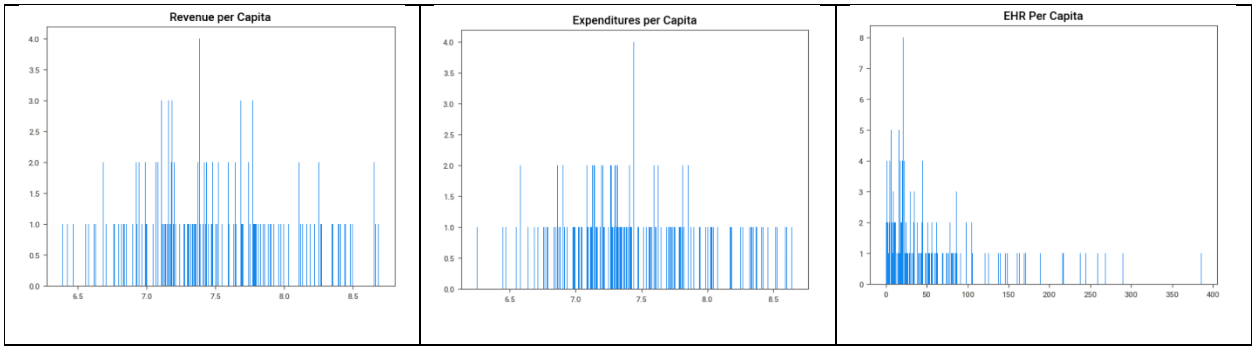


After Transformation

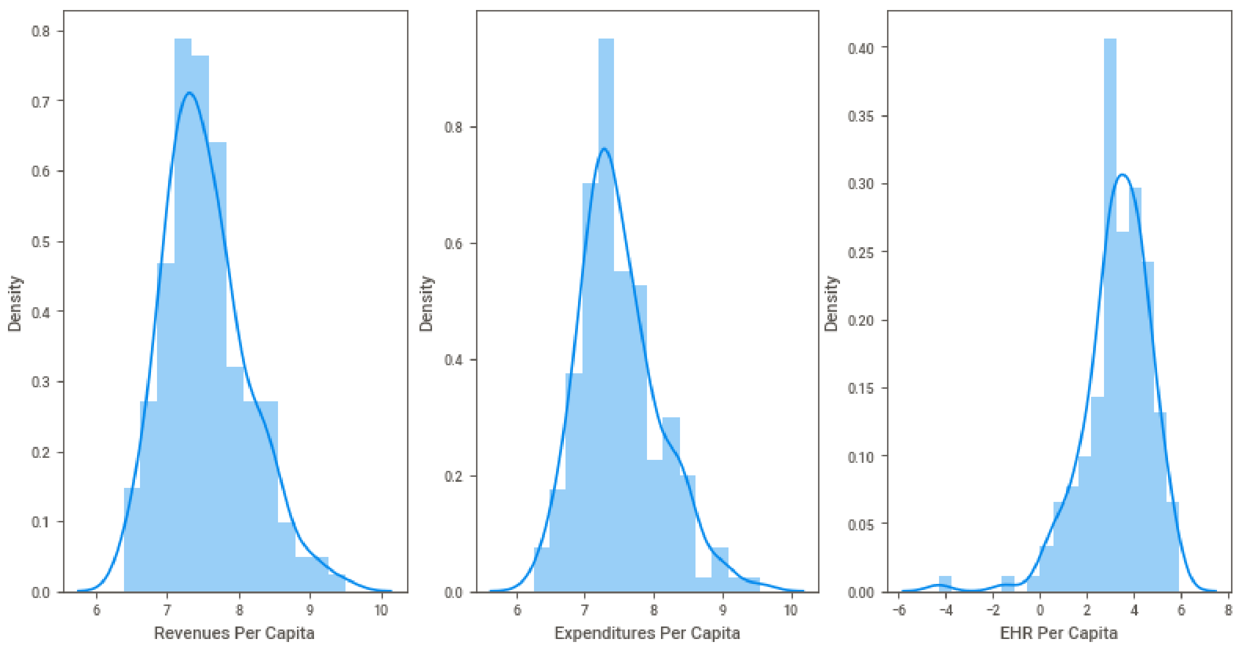


Hospitals

Before Transformation



After Log Transformation



Descriptive Statistics

Individual Healthcare Providers

	EHR per Capita	Revenue per Capita	Expenditure per Capita
Count	404	404	404
Mean	7.79	2082.49	2019.60

Standard Deviation	18.37	1875.80	2012.09
Minimum	0.14	350.00	361.00
25%	0.96	1171.25	1100.50
50% (Median)	2.49	1619.00	1544.00
75%	6.82	2341.25	2260.25
Max	204.62	21898.00	27753.00

Hospitals

	EHR per Capita	**Revenue per Capita**	**Expenditure per Capita**
Count	169		
Mean	52.19	2281.66	2194.14
Standard Deviation	64.07	1766.45	1723.80
Minimum	0.01	594.00	518.00
25%	11.27	1264.00	1242.00
50% (Median)	27.28	1697.00	1618.00
75%	63.12	2569.00	2460.00
Max	385.62	13461.00	14160.00

Appendix B – The Code Implementation

Analysis on the provider’s dataset:

- https://github.com/vicky-playground/EHR-GDP/blob/main/city_comparison_code/EHR%20City%20Corrolation.ipynb
- https://github.com/vicky-playground/EHR-GDP/blob/main/city_comparison_code/EHR%20City.ipynb

Analysis on the hospital's dataset:

- https://github.com/vicky-playground/EHR-GDP/blob/main/city_comparison_code/hospitals.ipynb