Fargo Health Group Cardiovascular Exam Projections – Abbeville Clinic:

A Review of Past Data and Predictions for the Future

Deanna Schneider

UW-Eau Claire

Fargo Health Group has seen significant growth in the quantity of disability exams they are performing throughout their health care network. Disability exams require a tight turnaround, with both monetary and reputation consequences if exams are not completed and reported on in time. It's imperative that Fargo Health Group be able to accurately predict the number of exams they will need to process, so that they can have the appropriate staff on hand to meet exam need. Predictive analytics can be employed to give Fargo Health Group the information they require to staff their health clinics and meet the growing demand.

For the pilot project, Fargo would like to focus on the Abbeville clinic, and has provided historical data of monthly exam demand. A spreadsheet was provided in which the majority of the data was already aggregated. However, there were several months for which data was spread across several worksheets. Time series projections require a single dataset that represents observations recorded at regular intervals (Kabacoff, 2015). Therefore, in order to use the provided data it needed to be cleaned and combined into a single complete dataset. I cleaned the data using the following steps.

Data Cleaning

The May, 2007 data and May-July, 2013 data was spread across several sheets and had inconsistent date formatting. The four sheets containing this data were combined into a single sheet (May-2007). When the data was combined, it became clear that duplicates were present in the data. (Request ID should be a unique field, but multiple rows shared the same Request ID.) I used conditional formatting to highlight the duplicate Request IDs and a flag column to ensure that only 1 of each Request ID was marked as non-duplicative. The dates were all adjusted by hand to follow standard date formatting. The month and year were extracted from the date. I created pivot tables for each of the months in question. The pivot tables summed the number of exams based on the month and year in question, the original hospital location (Abbeville), the non-duplicate identifier, and the exam type. Only heart-related conditions are in question, but I was not provided with a comprehensive list of exam types that related to cardiovascular conditions. I determined that the following exam types would be considered heart-related: angina, aortic valve stenosis, arrhythmia, CAD, cardiac, cardiovascular, chest pain, cor pulmonale, coronary artery disease (CAD), endocarditis, heart, heart palpitations, ischemic heart disease, myocardial infraction, myocardial ischemia, myocarditis, premature ventricular contraction, stress test, ventricular septal defect (VSD), and VSD. I combined the summed data with the existing information in the Abbeville, LA spreadsheet.

The December, 2013 data was provided as individual records in a separate spreadsheet. The documentation indicated that the Routing SYSID is a compound key, containing information about the routing hospital and the condition code. I broke the Routing SYSID into its component parts and created flags to identify rows of data that were relevant to this study. I made the assumption that the Routing SYSID is a unique identifier and flagged those rows that were duplicates. I summed the rows that met all the conditions (not a duplicate, came from Abbeville, related to heart conditions). The final sum was added to the Abbeville, LA spreadsheet.

Several other rows in the aggregate worksheet had data inconsistent with an exam count. March and June of 2006, May of 2009 and June of 2010 each had string characters in the exam count column. These values were converted to nulls. There were two outliers which clearly seemed to be data entry errors. I assumed as much and set December 2008 and January 2011 to null. There were two legitimate outliers in the data – October of 2008 and December of 2011. October of 2008 had a high number of exams due to a hurricane closing nearby facilities. Louisiana is prone to hurricanes, and therefore, this outlier was left in the dataset to represent the types of anomalies that might actually occur. The December 2008 data point was set to null, under the assumption that the clinic is not typically closed for the holidays. (It was not closed for any other year.) It is more logical to impute data for this outlier data point than to leave it at zero.

For December, 2009 through February, 2010 Fargo reported an aggregate number, but there's no valid mathematical way to break that number down across the three months. Simply dividing the data by 3 introduces a false damping effect (Kabacoff, 2015). Additionally, there are multiple missing data points for those three months in other years, making it difficult, if not impossible, to identify any ratio by which to divide the data. It is better to use a modern imputation method to fill those three months. All of the above logical decisions were based on the information in Chapter 18: Advanced Methods for Missing Data in *R in Action* (Kabacoff, 2015).

After completing the above methods, the data set was missing 10% of its values. Those values were imputed using the Amelia package (Honaker & King, 2016). Bounds were set on the exam counts column such that they should always be positive numbers. I attempted to use Amelia's priors to incorporate the aggregate amount for December, 2009-February, 2010 and the standard deviations seen in those three months in other years. But, the resulting dataset took longer to converge, varied more from the non-imputed distribution, and had more visual variance when over-imputed (Honaker & King, 2016). I proceeded with Amelia's imputations with bounds only. The Amelia function generated five distinct complete datasets. I combined the five back into one by averaging the Amelia data, resulting in a single complete dataset to use for predictions.

Methods

Initial inspection of the historical data indicates a clear upward trend. (See Figure 1.) Monthly demand had a low point of 311 exams per month in February, 2007 and topped out at 6094 exams in October of 2013. Notice the hurricane outlier in 2008, and the fluctuations around the level. The fluctuations appear to be increasing with the trend, which could indicate the need for a multiplicative model (Kabacoff, 2015). The clear trend and the fluctuations indicate that simple exponential smoothing would not be a suitable approach for forecasting this data. Holt and ARIMA are both potential viable candidates for predicting data that has a trend, seasonality, or multiplicative effects (Kabacoff, 2015).
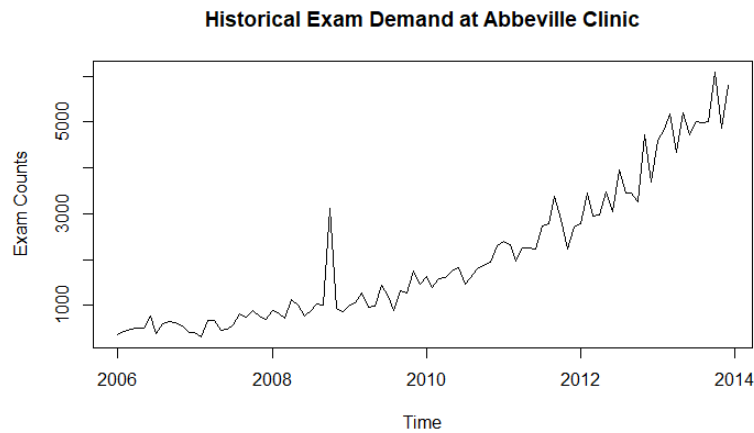
Figure 1

Holt Results

I used the ets package in R to generate a best-fit model using either Holt or Holt-Winters (Kabacoff, 2015). The model generated was a Holt model (A,A,N) with a level smoothing constant of .1166 and a trend smoothing constant of .0202. Both the alpha and beta smoothing constants were closer to zero than one, indicating that data further from the present was weighted more heavily that data close to the present (Foreman, 2014). I used the model to forecast out 12 additional months. Figure 2 contains a plot of the final year of historical data and the forecasted year.
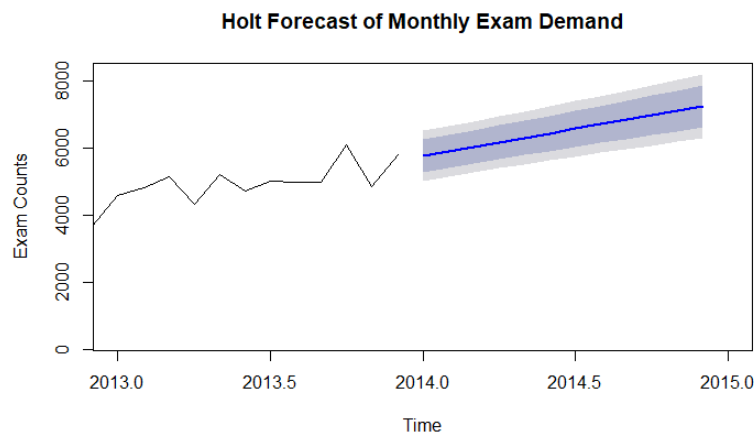

Figure 2

The plot shows that demand will continue to increase, but at a moderate rate. It may help to review the actual numbers. Table 1 shows the forecasted exam count for each of the next 12 months and the 80% and 95% confidence intervals.

		Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan	2014	5783.071	5297.149	6268.994	5039.917	6526.226
Feb	2014	5916.278	5425.832	6406.724	5166.205	6666.35
Mar	2014	6049.484	5553.143	6545.826	5290.395	6808.573
Apr	2014	6182.69	5678.939	6686.442	5412.268	6953.112
May	2014	6315.897	5803.097	6828.696	5531.638	7100.155
Jun	2014	6449.103	5925.52	6972.686	5648.352	7249.854
Jul	2014	6582.309	6046.132	7118.487	5762.297	7402.322
Aug	2014	6715.516	6164.882	7266.149	5873.394	7557.637
Sep	2014	6848.722	6281.743	7415.701	5981.602	7715.842
Oct	2014	6981.928	6396.708	7567.148	6086.911	7876.945
Nov	2014	7115.135	6509.79	7720.479	6189.34	8040.929
Dec	2014	7248.341	6621.017	7875.665	6288.931	8207.751

Table 1

ARIMA Results

ARIMA forecasting requires that the data meet the definition of stationary. Stationary time series have a constant mean and variance (Kabacoff, 2015). To ensure that the data from Abbeville is stationary, I performed an Augmented Dickey-Fuller (ADF) test. Kabacoff (2015) states that, "A significant result suggests stationarity" (pg. 361). The result from the ADF test on the Abbeville data was .9752, which is not significant. It is likely that the trend in the data means that differencing is required (Kabacoff, 2015). I then used the ndiffs function to determine if the data needed to be differenced. It required a single differencing. The plot of the differenced data shows that differencing the data effectively removed the upward trend (Figure 3).
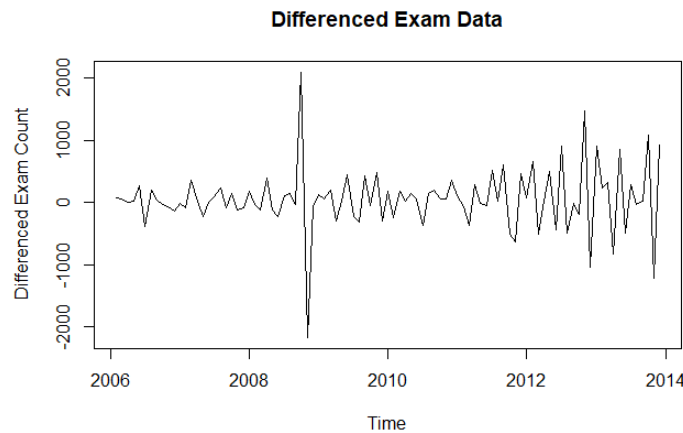

Figure 3

After differencing the data, I ran ACF and PACF tests to assess the best type of ARIMA forecast. Results of these tests are visible in Figure 4.
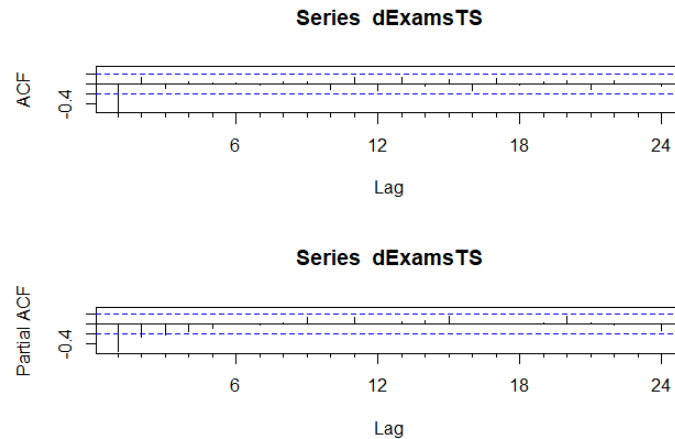
Figure 4

The ACF test shows one large spike at the beginning of the ACF plot and a gradual decay in the PACF plot. Additionally, the correlations for both the ACF and PACF are negative. This pattern suggests an MA(1) signature (Nau, n.d.).

The forecast package in R provides a simple automatic ARIMA modeling approach. The auto.arima function returned a model that used p of 0, d of 1, and q of 1. This matches the MA(1) signature that was expected based off of the ACF and PACF tests, and corresponds to simple exponential smoothing (once the trend has been accounted for) (Nau, n.d.). The ARIMA model was used to predict an additional 12 months. Figure 5 is a plot of the final year of actual demand and the forecast for the next 12 months.
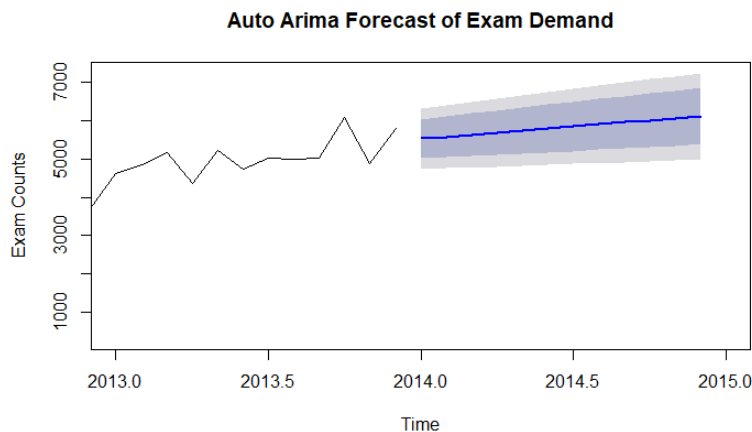

Figure 5

This plot shows less upward trend than the Holt forecast. Again, the actual forecasted demand numbers and confidence intervals can be helpful in planning physician availability. That data can be seen in Table 2.

		Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan	2014	5523.678	5007.663	6039.693	4734.501	6312.855
Feb	2014	5577.739	5038.593	6116.886	4753.186	6402.293
Mar	2014	5631.8	5070.475	6193.126	4773.327	6490.274
Apr	2014	5685.861	5103.2	6268.523	4794.758	6576.965
May	2014	5739.923	5136.68	6343.165	4817.343	6662.503
Jun	2014	5793.984	5170.839	6417.128	4840.966	6747.001
Jul	2014	5848.045	5205.614	6490.475	4865.532	6830.557
Aug	2014	5902.106	5240.952	6563.26	4890.958	6913.254
Sep	2014	5956.167	5276.806	6635.529	4917.174	6995.161
Oct	2014	6010.228	5313.135	6707.322	4944.116	7076.341
Nov	2014	6064.289	5349.904	6778.675	4971.731	7156.848
Dec	2014	6118.351	5387.081	6849.62	4999.971	7236.73

Table 2

Holt versus ARIMA – Model Fit

While each model shows similar visual results, we can also evaluate model fit based on a number of factors. One way to evaluate model fit is to review the residuals and autocorrelations. Kabacoff (2015) states that, "the residuals should be normally distributed with mean zero, and the autocorrelations should be zero for every possible lag" (p. 365). I tested this assumption using the qqnorm and qqline functions (which shows distributions of residuals) and the Box-Ljung test (which tests that autocorrelations are all zero). For both models, the Box-Ljung test returned insignificant results, supporting the assertion that the autocorrelations don't differ from zero (Kabacoff, 2015). Plots for the Holt and ARIMA residuals are shown below. As you can see, the Holt residuals fall more closely along the line. This indicates that they are closer to being normally and independently distributed than the ARIMA residuals (Kabacoff, 2015). This may be our first indication that the Holt model is a more accurate fit than the ARIMA model.
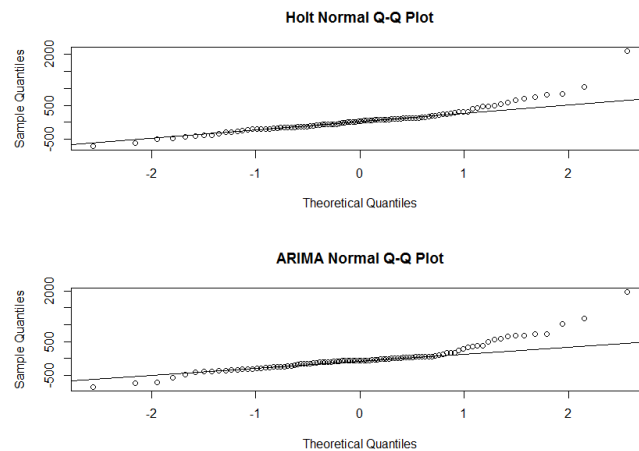

Figure 6

In addition to examining the residuals, there are a number of accuracy statistics that can be used to compare models. The following table summarizes the accuracy measures of the two models:

Model	Mean Absolute Scaled Error (MASE)	Root Mean Error Squared (RMSE)	Mean Absolute Percentage Error (MAPE)	AIC	BIC
Holt	.35	379.17	15.80%	1588.27	1601.09
ARIMA	.36	396.30	18.32%	1413.88	1414.14

An explanation of these accuracy measures is in order. Kabacoff describes the Mean Absolute Scaled Error as "the most recent accuracy measure and is used to compare the forecast accuracy across time series on different scales" (p 355). The Root Mean Error Squared describes the magnitude of the errors, and the Mean Absolute Percentage Error represents error as a percentage (Bukralia, n.d.). In all cases, the lower the number, the less error the model has. AIC and BIC are measures of the goodness of fit and the simplicity of the model (Keshvani, 2013). However, because the ets model and the arima model treat initial values differently, AIC is not a valuable tool for comparing the two models (Hyndman, 2013). Therefore, in comparing these two models, the Holt model appears to be the more accurate of the two.

Recommendations for Fargo Health Group

Based on the predictive models, Fargo Health Group should be prepared for a gradually increasing number of exams over the next 12 months. The historical max workload that the Abbeville clinic has experienced is 6094 exams in October of 2013. The Holt model indicates that there's an 80% certainty Abbeville will exceed its 6094 record by August of 2014 (with an estimated low threshold of exams at 6164). However, the actual number of exams that month could be as high as 7266. You can see that there is a wide range of possible exam demand within the 80% confidence interval.

This project only encompassed a single health center, and a single type of examination. To fully understand physician scheduling needs, I recommend that Fargo Health Group expand their data analysis to include all health centers and all exam types. A forecast that predicts demand based on center and exam type may give a more accurate representation of true need.

Perhaps the most important recommendation for Fargo Health Group is an increased emphasis on accurate data collection and repeated model-fitting as newer, more accurate data is collected. In the initial dataset, for the months that individual data was given, there was a 1.5% duplicate rate in one dataset and a 3.7% duplicate rate in another. Additionally, there was a 6.2% data entry error rate for the aggregate data (not including those months for which data was not available). These error rates aren't terrible. But, the more errors can be reduced, the more accurate the predictions will be. The following recommendations can improve data accuracy:

- Develop a data governance policy. Data governance will improve the quality of the data, as well as defining roles for data access and data lifecycle and setting your organizations data principles. (Khatri & Brown, 2010)
- Use a single, consolidated system of record for tracking patient visits. Most modern health clinics have implemented patient-tracking software that functions across all clinics in the system. Software such as this will greatly improve data accuracy.
- Have a data custodian assigned, who has primary responsibility for the accuracy of data.

I would recommend continuing to collect data on a monthly basis and running 3 month predictions, because time series predictions are most accurate in the short term.

Ethics

I was asked to comment on the ethical implications of this project. Ethics in the medical records community is governed by HIPAA, the Health Insurance Portability and Accountability Act of 1996. HIPAA ensures the privacy of all information "that identifies the individual or for which there is a reasonable basis to believe it can be used to identify the individual" (U.S. Department of Health & Human Services, n.d., What Information is Protected section, para. 1). While, on the surface, the data I was provided may not appear to be personally identifiable, with the advent of machine learning and the proliferation of social media, the data may very well be personally identifiable. I received individualized data that contained date of exam, condition codes, and location of the exam. If an individual in our dataset posted, "Saw my doc today – the ticker's all right!" on Facebook or Twitter, it would be possible to use the date of the post, their location, and the information in our dataset to potentially link them to a specific exam type or specific diagnosis (Daries, et al., 2014). This becomes a significant issue for HIPAA compliance. However, there are "no restrictions on the use or disclosure of de-identified health information" (U.S. Department of Health & Human Services, n.d., De-Identified Health Information section, para. 1). Therefore, the aggregate data provided in the Abbeville, LA spreadsheet, which consists merely of total counts of heart-related exams per month, would not violate the privacy laws for HIPAA. I recommend that Fargo Health employ an internal employee to clean and aggregate the data prior to engaging with an outside consultant to generate models and forecasts.

Privacy is not the only concern. Another key component of ethical data science in the health services field is informed consent. The U.S. Department of Health and Human Services requires that investigators "obtain the legally effective informed consent of the subject" (HHS, n.d.b, para. 1) in most circumstances. I received no information as to whether or not informed consent was obtained. I would strongly encourage Fargo Health Department to ensure that all patients are signing informed consent forms that allow for predictive analysis.

Assuming both privacy and consent are covered, it's still important to understand what Fargo Health Group intends to do with this analysis. In a best-case scenario, the analysis will be used to ensure that enough physicians are on staff to cover the increasing needs of disability examination patients. In the worst case scenario, this analysis is used to determine that disability examinations are too time-consuming and costly and availability of disability examinations is limited at one or more of their clinics. The latter scenario could be seen as discriminatory. The

Americans with Disabilities act is a comprehensive piece of disability-related legislation. Most relevant to our study, the act requires that "health care providers provide individuals with disabilities full and equal access to their health care services and facilities" (ADA National Network, n.d., Introduction section, para 1). If the data in this analysis was used to funnel disability examinations to specific locations, thereby limiting access at other locations, it is an ethical problem. Though it may be going too far to say that Fargo Health group has a moral obligation to meet the growing demand for disability examinations, it is not unreasonable to expect them to use this information to better prepare themselves for meeting this growing need. (There is never any one clear and simple decision in data analysis for business needs. This is just one component of their overall business plan, and meeting this need should not be done at the expense of meeting the needs of other patients.)

In any data analytics project, accountability is a key component of a successful data project. In this project, the client was not available to answer questions or clarify assumptions made by the data scientist. That is a recipe for failure. The best data science projects are those in which all members are actively engaged and there is ongoing dialogue as questions arise. Ultimately, accountability falls on the shoulders of the project sponsor, who should be working closely with the data scientist to ensure that accurate, on-going prediction can be made.

Final Thoughts

The takeaways from this analysis are as follows:

1. There is a continuing upward trend in the need for disability examinations at Fargo Health Group's Abbeville clinic.
2. Continue to collect data, with an emphasis on accuracy and completeness and re-run the analysis on a regular basis.
3. Make sure to meet all HIPAA compliance points by de-identifying data given to consultants and ensuring informed consent.

Data analysis and forecasting is a powerful tool that could help Fargo Health Group better meet the needs of its patients, funders, and providers. I urge you to continue to use predictive analytics to position your business for success.

References

ADA National Network. (n.d.) *Accessible Health Care*. Retrieved from
https://adata.org/factsheet/accessible-health-care.

Bukralia, R. (n.d.). *DS700 – Foundations of Data Science: Forecasting*. Retrieved from
https://uwli.courses.wisconsin.edu/d2l/le/content/3713709/viewContent/22681703/View.

Daries, J., Reich, J., Waldo, J., Young, E., Whittinghill, D., Seaton, T.D…., Chuang, I. (2014).
Privacy, Anonymity, and Big Data in the Social Sciences. Retrieved from
http://queue.acm.org/detail.cfm?id=2661641

Foreman, J. (2014). *Data Smart: Using Data Science to Transform Information into Insight*.
Indianapolis, IN. John Wiley & Sons, Inc.

Honaker, J., King, G., Blackwell, M. (2016). *Package 'Amelia'*. Retrieved from https://cran.r-
project.org/web/packages/Amelia/Amelia.pdf.

Hyndman, R. (2013). *Facts and Fallacies of the AIC*. Retrieved from
https://robjhyndman.com/hyndsight/aic/.

Kabacoff, Robert. (2015). *R In Action: Data Analysis and Graphics with R*. Shelter Island, NY:
Manning Publications Co.

Khatri, Vijay & Brown, Carol V. (2010). Designing Data Governance. Retrieved from
http://delivery.acm.org/10.1145/1630000/1629210/p148-
khatri.pdf?ip=144.92.114.82&id=1629210&acc=ACTIVE%20SERVICE&key=066E7B0
AFE2DCD37%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E3
8B35&CFID=793212033&CFTOKEN=75343802&__acm__=1501781941_1973f25fdda
97226a7835b89e5113742

Keshvani, A. (2013). *Using AIC to Test ARIMA Models*. Retrieved from
https://coolstatsblog.com/2013/08/14/using-aic-to-test-arima-models-2/.

Nau, R. (n.d.) *Notes on nonseasonal ARIMA models*. Retrieved from
https://people.duke.edu/~rnau/Notes_on_nonseasonal_ARIMA_models--
Robert_Nau.pdf.

Paiz, J. M., Angeli, E., Wagner, J., Lawrick, E., Moore, K, Anderson, M.…, Keck, R. (2016)
Reference List: Basic Rules. Retrieved from
https://owl.english.purdue.edu/owl/resource/560/05/.

Paiz, J. M., Angeli, E., Wagner, J., Lawrick, E., Moore, K, Anderson, M.…, Keck, R. (2016)
General Format. Retrieved from https://owl.english.purdue.edu/owl/resource/560/01/.

U.S. Department of Health and Human Services. (n.d.). *Summary of the HIPAA Privacy Rule*.
Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/laws-
regulations/index.html.

U.S. Department of Health and Human Services. (n.d.). *Informed Consent FAQs*. Retrieved from
https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/informed-consent/index.html/