Big data analytics for deriving predictive healthcare insights

Ankit Agrawal and Alok Choudhary
Department of Electrical Engineering and Computer Science
Northwestern University
2145 Sheridan Rd
Evanston, IL 60201
USA
{ankitag,choudhar}@eecs.northwestern.edu

Abstract

This chapter describes the application of big data analytics in healthcare, particularly on electronic healthcare records so as to make predictive models for healthcare outcomes and discover interesting insights. A typical workflow for such predictive analytics involves data collection, data transformation, predictive modeling, evaluation, and deployment, with each step tailored to the end goals of the project. To illustrate each of these steps, we shall take the example of recent advances in such predictive analytics on lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) program. This includes the construction of accurate predictive models for lung cancer survival, development of a lung cancer outcome calculator deploying the predictive models, and association rule mining on that data for bottom-up discovery of interesting insights. The lung cancer outcome calculator illustrated here is available at http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator.

Introduction

The term "big data" has become a ubiquitous buzzword today in practically all areas of science, technology, and commerce. It primarily denotes datasets that are too large, complex, or both, to be adequately analyzed by traditional processing techniques. Scientific and technological advances in measurement and sensor devices, databases, and storage systems have made it possible to efficiently *collect*, *store* and *retrieve* huge amounts of and different kinds of data. However, when it comes to the *analysis* of such data, we have to admit that our ability to generate big data has far outstripped our analytical ability to make sense of it. This is true in practically all fields, and the field of medicine and healthcare is no exception to it, where the Fourth paradigm of science (data-driven analytics) is increasingly becoming popular and has led to the emergence of the new field of healthcare informatics. The Fourth paradigm of science [1] unifies the first three paradigms of science – namely theory, experiment, and simulation/computation. The need for such data-driven analytics in healthcare has also been emphasized by large-scale initiatives all around the world, such as Big Data to Knowledge (BD2K) and Precision Medicine Initiative of National Institutes of Health in USA, Big Data for Better Outcomes Initiative in Europe, and so on.

The bigness (amount) of data is certainly the central feature and challenge of dealing with the so-called big data, but it is many times accompanied by one or more of other features that can make the collection and analysis of such data even more challenging. For example, the data could be from several heterogeneous sources, may be of different types, may have unknown dependencies and inconsistencies within it, parts of it could be missing or not reliable, the rate of data generation could be much more than what traditional systems could handle, and so on. All this can be summarized by the famous Vs associated with big data, as presented in Figure 1 and briefly described below:

- *Volume*: It refers to the amount of data. Datasets of sizes exceeding terabytes and even petabytes are not uncommon today in many domains. This presents one of the biggest challenge in big data analytics.

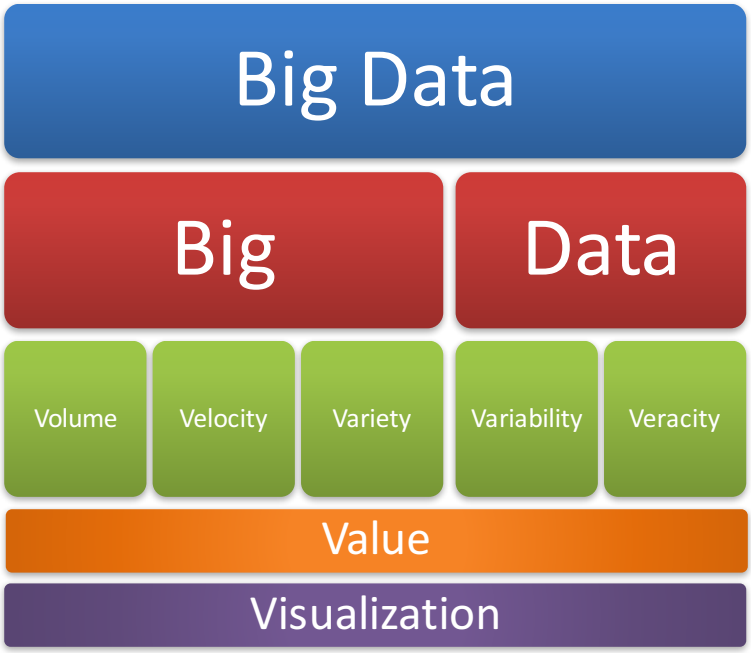


Figure 1: The various Vs associated with big data. Volume, velocity, and variety are unique features of big data that represent its bigness. Variability and veracity are characteristics of any type of data, including big data. The goal of big data analytics is to unearth the value hidden in the data, and appropriately visualize it to make informed decisions.

- *Velocity*: The speed with which new data is generated. The challenge here is to be able to effectively process the data in real-time. A good example of high velocity data source is Twitter, where more than 5000 tweets are posted every second.
- *Variety*: This refers to the heterogeneity in the data. For instance, many different types of healthcare data are generated and collected by different healthcare providers, such as electronic health records, x-rays, cardiograms, genomic sequence, etc. It is important to be able to derive insights by looking at all available heterogenous data in a holistic manner.
- *Variability*: The inconsistency in the data. This is especially important since the correct interpretation of the data can vary significantly depending on its context.
- *Veracity*: It refers to how trustworthy the data is. The quality of the insights resulting from analysis of any data is critically dependent on the quality of the data itself. Noisy data with erroneous values or lot of missing values can greatly hamper accurate analysis.
- *Visualization*: It means the ability to interpret the data and resulting insights. Visualization can be especially challenging for big data due to its other features as described above.
- *Value*: The goal of big data analytics is to discover the hidden knowledge from huge amounts of data, which is akin to finding a needle in a haystack, and can be extremely valuable. For example, big data analytics in healthcare can help enable personalized medicine by identifying optimal patient-specific treatments, which can potentially improve millions of lives, reduce waste of healthcare resources, and save billions of dollars in healthcare expenditure.

The first three Vs above distinguish big data from small data, and other Vs are characteristics of any type of data, including big data. Further, each application domain can also introduce its own nuances to the process of big data management and analytics. For example, in healthcare, the privacy and security of patients' data is of paramount importance, and compliance to HIPAA (Health Insurance Portability and

Accountability Act) and IRB (Institutional Review Board) protocols is necessary to work with many types of healthcare data. It is also worth noting here that although the size and scale of healthcare data is not as large as in some other domains of science like high-energy-physics or in business and marketing, but the shear complexity and variety in healthcare data becoming available nowadays requires the development of new big data approaches in healthcare. For example, there are electronic healthcare records (EHRs), medical images (e.g. mammograms), time-series data (e.g. ECG signals), textual data (doctor's notes, research papers), genome sequence and related data (e.g. SNPs).

So what can big data analytics do for a real world healthcare application? A variety of personalized information such as patients electronic health records is increasingly becoming available. What if we could intelligently integrate the hidden knowledge from such healthcare data during a real-time patient encounter to complement physicians expertise and potentially address the challenges of personalization, safe and cost-effective healthcare? Note that the challenge here is to make the insights patient-specific instead of giving generic population-wide statistics. Why is this important? Let us try to understand with the help of an example. The benefits of medical treatments can vary depending on one's expected survival, and thus not considering an individual patients prognosis can result in poor quality of care as well as non-optimal use of healthcare resources. Developing accurate prognostic models using all available information and incorporating them into clinical decision support could thus significantly improve quality of healthcare [2], both in terms of improving clinical decision support and enhancing informed patient consent. Development of accurate data-driven models can also have a tremendous economic impact. The Centers for Disease Control and Prevention estimates that there are more than 150,000 surgical-site infections annually [3], and it can cost $11,000 to $35,000 per patient, i.e., about $5 billion every year. Accurate predictions and risk estimation for healthcare outcomes can potentially avoid thousands of complications, resulting in improved resource management and significantly reduced costs. This requires development of advanced data-driven technologies that could effectively mine all available historical data, extract and suitably store the resulting insights and models, and make them available at the point of care in a patient-specific way.

In the rest of this chapter, we will see one such application of big data analytics on electronic healthcare records so as to make predictive models on it and discover interesting insights. In particular, we will take the example of lung cancer data from the Surveillance, Epidemiology, and End Results (SEER) program to build models of patient survival after 6 months, 9 months, 1 year, 2 years, and 5 years [4] and for conditional survival as well [5]. We will also see the application of association rule mining on this dataset for 5-year survival [6] and 5-year conditional survival [7]. Finally, we will discuss the online lung cancer outcome calculator that resulted from the described predictive analytics on SEER data, and conclude with some examples of big data analytics in other healthcare-related applications.

Big Data Analytics on SEER Lung Cancer Data

Lung (respiratory) cancer is the second most common cancer and the leading cause of cancer-related deaths in the USA. In 2012 alone, over 157,000 people in the United States died from lung cancer. The 5-year survival rate for lung cancer is estimated to be just 15% [8]. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an authoritative repository of cancer statistics in USA [9]. It is a population-based cancer registry covering about 26% of the US population and is the largest publicly available cancer dataset in USA. It collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix [8]. The SEER data attributes can be broadly categorized into demographic attributes, diagnosis attributes, treatment attributes, and outcome attributes (see Table 1). The presence of outcome attributes makes the SEER data very useful for doing predictive analytics and making models for cancer survival.

Lung cancer survival prediction system

Till now we have seen what big data is and what big data analytics can do for healthcare applications. We have also had a brief introduction to SEER and what kind of data is present in the SEER database. So now let us dive deeper into what a typical workflow for predictive analytics looks like, with the specific example of lung cancer survival prediction on SEER data. Figure 2 depicts the overall end-to-end workflow. It is worth mentioning here that this workflow for predictive lung cancer outcome analytics is essentially a

Table 1: SEER data attributes

Type	*Examples*
Demographic	Age, gender, location, race/ethnicity, date of diagnosis
Diagnosis	Tumor primary site, size, extension, lymph node involvement
Treatment	Primary treatment, surgical procedure, radiation therapy
Outcome	Survival time, cause of death

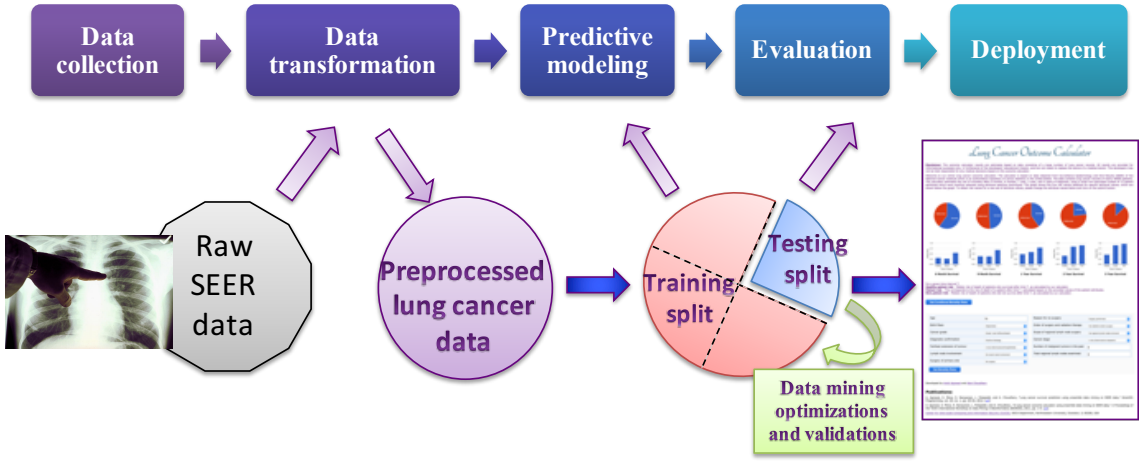


Figure 2: A typical workflow for predictive analytics, illustrated with the example of outcome prediction models for lung cancer using SEER data

healthcare adaptation of existing similar data science workflows in other domains, since most of the advanced techniques for big data management and analytics are invented in the field of computer science and more specifically high-performance data mining [10, 11], via applications in many different domains like business and marketing [12], climate science [13], materials informatics [14], and social media analytics [15], among many others. Here we will only focus on the healthcare application of developing a lung cancer survival prediction system. As shown in Figure 2, it has five stages described below.

Data collection

This is the obvious first step. Depending on the project, the kind of data required for it, and the license agreements associated with that data, this can be the easiest or the toughest step in the workflow. SEER has made it easy to get the 'SEER limited-use data' from their website on submitting a SEER limited-use data agreement form. It creates a personalized SEER research data agreement for every user that allows the use of the data for only research purposes. In particular, there must be no attempt to identify the individual patients in the database. Of course, the obvious identification information like patient name, SSN, etc. are excluded from the data released by SEER, but it still has demographic information like age, sex, race, which is very useful for research purposes, but should not be misused to try to identify patients in any way. Such compliance to HIPAA regulations is important to preserve patient privacy.

Data transformation

Once the data is available, the first step is to understand the data format and representation, and do any necessary transformations to make it suitable for modeling. Let us assume the data is in a row-column (spreadsheet) format, such as in the case of SEER data. Each row corresponds to a patient's medical record, and can also be referred to as an instance, data point, or observation. The columns are the attributes, such

as age, race, tumor size, surgery, outcome, etc. Data attributes can be of different types – numeric, nominal, ordinal, interval – and it is important to have the correct representation of each attribute for analysis, for which some data transformation might be necessary. More broadly, data transformation is needed to ensure the quality of the data ahead of modeling and remove or appropriately deal with noise, outliers, missing values, duplicate data instances, etc.

Data transformation is usually unsupervised, which means that it does not depend on the outcome or target attributes. For example, SEER encodes all attributes as numbers, and many of them are actually nominal, like marital status, where "1" represents "Single", "2" represents "Married", "3" represents "Separated", "4" represents "Divorced", "5" represents "Widowed", and "9" represents "Unknown". Numbers have a natural order and the operations of addition, subtraction, and division are defined, which may be fine for numeric attributes like "tumor size", but not for nominal attributes like marital status, sex, race, etc., Such attributes need to be explicitly converted to nominal for correct predictive modeling. Even numeric attributes need to examined carefully. For example, the tumor size attribute in SEER data gives the exact size of tumor in mm, if it is known. But in some cases, the doctor notes may say "less than 2cm", in which case it is encoded as "992", which could easily be misinterpreted as 992mm if not transformed appropriately. Another example of a unsupervised data transformation required in SEER data is to construct numeric survival time in months from the SEER format of YYMM, so that it can be modeled correctly.

The above data transformations are required due to the way SEER data is represented, and may be necessary for almost any project dealing with this data. But there are also problem-specific data transformations that may be necessary for building a model as originally intended. For example, if we are interested in building a predictive model for lung cancer survival, then we should only include those patient records where the cause of patients' death was lung cancer, which is given by the "cause of death" attribute. We also need to remove certain attributes from the modeling that directly or indirectly specify the outcome, e.g. cause of death, whether the patient is still alive. Further, for binary class prediction, we also need to derive appropriate binary attributes for survival time, e.g. 5-year survival.

There are also certain data transformation steps that could be supervised in some cases, meaning that they depend on the outcome attribute(s). Examples include feature selection/extraction, discretization, sampling, and all of these can be supervised or unsupervised. If they are supervised, they should in general be considered together with other supervised analytics so as to avoid over-fitting (more about this later).

Predictive modeling

Once appropriate data transformation has been performed and the data is ready for modeling, we can employ supervised data mining techniques for feature selection and predictive modeling. Caution needs to be exercised here to appropriately split the data into training and testing sets (or use cross validation), else the model may be subject to overfitting and give over-optimistic accuracy. If the target attribute is numeric (e.g. survival time) regression techniques can be used for predictive modeling, and if it is categorical (e.g. whether a patient survived at least five years) classification techniques can be used. Some techniques are capable of doing both regression and classification. Further, there also exist several ensemble learning techniques that can combine the results from base learners in different ways, and in some cases have shown to improve accuracy and robustness of the final model. Table 2 lists some of the popular predictive modeling techniques.

Evaluation

Traditional statistical methods such as logistic regression are typically evaluated by building the model on the entire available data, and computing prediction errors on the same data, and it has been a common practice in statistical analysis of medical data as well for many years. Although this approach may work well in some cases, it is nonetheless prone to over-fitting, and thus can give over-optimistic accuracy. It is easy to see that a data-driven model can, in principle "memorize" every single instance of the dataset and thus result in 100% accuracy on the same data, but will most likely not be able to work well on unseen data. For this reason, advanced data-driven techniques that usually result in black-box models need to be evaluated on data that the model has not seen while training. A simple way to do this is to build the model only on random half of the data, and use the remaining half for evaluation. This is called the train-test split setting for model evaluation. Further, the training and testing halves can then also be swapped for another round

Table 2: Popular predictive modeling algorithms

Modeling Technique	*Brief description*
Naive Bayes	A probabilistic classifier based on Bayes theorem
Bayesian network	A graphical model that encodes probabilistic conditional relationships among variables
Logistic regression	Fits data to a sigmoidal S-shaped logistic curve
Linear regression	A linear least-squares fit of the data w.r.t. input features
Nearest-neighbor	Uses the most similar instance in the training data for making predictions
Artificial neural networks	Uses hidden layer(s) of neurons to connect inputs and outputs, edge weights learnt using back propagation (called deep learning if more than two layers)
Support vector machines	Based on the Structural Risk Minimization, constructs hyperplanes multidimensional feature space
Decision table	Constructs rules involving different combinations of attributes
Decision stump	A weak tree-based machine learning model consisting of a single-level decision tree
J48 (C4.5) decision tree	A decision tree model that identifies the splitting attribute based on information gain/gini impurity
Alternating decision tree	Tree consists of alternating prediction nodes and decision nodes, an instance traverses all applicable paths
Random tree	Considers a randomly chosen subset of attributes
Reduced error pruning tree	Builds a tree using information gain/variance and prunes it using reduced-error pruning to avoid over-fitting
AdaBoost	Boosting can significantly reduce error rate of a weak learning algorithm
Bagging	Builds multiple models on bootstrapped training data subsets to improve model stability by reducing variance
Random subspace	Constructs multiple trees systematically by pseudo-randomly selecting subsets of features
Random forest	An ensemble of multiple random trees
Rotation Forest	Generates model ensembles based on feature extraction followed by axis rotations

of evaluation and the results combined to get predictions for all the instances in the dataset. This setting is called two-fold cross validation, as the dataset is split into two parts. It can further be generalized to k-fold cross validation, where the dataset is randomly split into k parts. $k-1$ parts are used to build the model and the remaining one part is used for testing. This process is repeated k times with different test splits, and the results combined to get predictions for the all the instances in the dataset using a model that did not see them while training. Leave-one-out cross validation (LOOCV) is a special case of the more generic k-fold cross validation, with $k = N$, the number of instances in the dataset. LOOCV is commonly used when the dataset is not very large. To predict the target attribute for each data instance, a separate predictive model is built using the remaining $N-1$ data instances, and the whole process is repeated for each data instance. The resulting N predictions can then be compared with the N actual values to calculate various quantitative metrics for accuracy. In this way, each of the N instances is tested using a model that did not see it while training, thereby maximally utilizing the available data for model building. Cross validation is a standard evaluation setting to eliminate any chances of over-fitting. Of course, k-fold cross validation necessitates building k models, which may take a long time on large datasets.

Comparative assessments of how close the models can predict the actual outcome are used to provide an evaluation of the models' predictive performance. Many binary classification performance metrics are usually used for this purpose such as accuracy, precision, recall/sensitivity, specificity, area under the ROC curve, etc.

1. **c-statistic (AUC)**: The ROC (Receiver operating characteristic) curve is a graphical plot of true positive rate and false positive rate. The area under the ROC curve (AUC or c-statistic) is one of the most effective metric for evaluating binary classification performance, as it is independent of the probability cutoff and measures the discrimination power of the model.

2. **Overall accuracy**: It is the percentage of predictions that are correct. For highly unbalanced classes where the minority class is the class of interest, overall accuracy by itself may not be a very useful indicator of classification performance, since even a trivial classifier that simply predicts the majority class would give high values of overall accuracy.

$$Overall\ accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

   where TP is the number of true positives (hits), TN is number of true negatives (correct rejections), FP is number of false positives (false alarms), and FN is number of false negatives (misses).

3. **Sensitivity (Recall)**: It is the percentage of positive labeled records that were predicted positive.

Recall measures the completeness of the positive predictions.

$$Sensitivity = \frac{TP}{(TP + FN)}$$

4. **Specificity**: It is the percentage of negative labeled records that were predicted negative, thus measuring the completeness of the negative predictions.

$$Specificity = \frac{TN}{(TN + FP)}$$

5. **Positive predictive value (Precision)**: It is the percentage of positive predictions that are correct. Precision measures the correctness of positive predictions.

$$Positive\ predictive\ value = \frac{TP}{(TP + FP)}$$

6. **Negative predictive value**: It is the percentage of negative predictions that are correct, thereby measuring the correctness of negative predictions.

$$Negative\ predictive\ value = \frac{TN}{(TN + FN)}$$

7. **F-measure**: It is not too difficult to have a model with either good precision or good recall, at the cost of each other. F-measure combines the two measures in a single metric such that it is high only if both precision and recall are high.

$$F - measure = \frac{2.precision.recall}{(precision + recall)}$$

Deployment

After the predictive models have been constructed and properly evaluated, they need to be deployed appropriately to make the resulting healthcare insights available to various stakeholders at the point of care. For the lung cancer survival prediction project, the predictive models were incorporated in a web-tool that allows users to enter patient attributes and get patient-specific risk values. More details about the lung cancer outcome calculator are described later in this chapter.

Conditional survival prediction

Survival prediction from time of diagnosis can be very useful as we have seen till now, but for patients who have already survived a period of time since diagnosis, conditional survival is a much more clinically relevant and useful measure, as it tries to incorporate the changes in risk over time. Therefore, the above-described lung cancer survival prediction system was adapted to create additional conditional survival prediction models. Since 5-year survival rate is the most commonly used measure to estimate the prognosis of cancer, the conditional survival models were designed to estimate patient-specific risk of mortality after five years of diagnosis of lung cancer, given that the patient has already survived for 3 months, 6 months, 12 months, 18 months, and 24 months.

In order to construct a model for estimating mortality risk after five years of diagnosis of patients already survived for time T, only those patients were included in the modeling data that survived at least time T. Note that this is equivalent to taking the data used in the calculator to build 5-year survival prediction model, and removing the instances where the survival time was less than T. Thus, five new datasets were created for five different values of T (3 months, 6 months, 12 months, 18 months, and 24 months), and the same binary classification techniques were used to build five new models.

Association rule mining

Association rule mining is useful to discover patterns in the data. In contrast with predictive modeling where one is interested in predicting the outcome for a given patient, here one is interested in bottom-up discovery of associations among the attributes. If a target attribute is specified, such association rule mining can help identify segments (subsets of data instances) in the data defined by specific attributes' values such that those segments have extreme average values of the target attribute. Note that this is tantamount to the inverse question of retrieval in databases, where one gives the segment definition in terms of attribute values, and the database system returns the segment, possibly along with the average value of the target attribute in that segment. However, such database retrieval cannot automatically discover segments with extreme average values of the target attribute, which is exactly what association rule mining can do. Let us take the example of the SEER dataset to make it clear. In this case, we have patient attributes including an outcome/target attribute (survival time). Let us say the average survival time in the data is t_{avg}. It would then be of interest to automatically discover from the data under what conditions – as defined by the combination of patient attribute-values – is the survival time t'_{avg} significantly greater or significantly lower than t_{avg}. Similarly, if the target attribute is nominal like 5-year-survival (whether or not a patient survived for at least five years), and the fraction of survived patients in the entire dataset is f, then it would be interesting to find segments where this fraction f' is significantly higher or lower than f.

Illustrative data mining results on SEER data

We now present some examples of the results of above-described big data analytics on lung cancer EHR data from SEER. In [5], the SEER November 2008 Limited-Use Data files [9] were used, which was released in April 2009. It had a follow-up cutoff date of December 31, 2006, i.e., the patients were diagnosed and followed-up up to this date. Data was selected for the patients diagnosed between 1998 and 2001. Since the follow-up cutoff date for the SEER data in study was December 31, 2006 and the goal of the project was to predict survival up to five years, data of 2001 and before was used. Also, since several important attributes were introduced to the SEER data in 1998 (like RX Summ-Surg Site 98-02, RX Summ-Scope Reg 98-02, RX Summ-Surg Oth 98-02, Summary stage 2000 (1998+)), data of 1998 and after was used. There were a total of 70,132 instances of patients with cancer of the respiratory system between 1998 and 2001, and there were 118 attributes in the raw data from SEER.

The SEER-related preprocessing resulted in modification and splitting of several attributes, many of which were found to have significant predictive power. In particular, 2 out of 11 newly created (derived) attributes were within the top 13 attributes that were eventually selected to be used in the lung cancer outcome calculator. These were a) the count of regional lymph nodes that were removed and examined by the pathologist; and b) the count of malignant/in-situ tumors. These attributes were derived from 'Regional Nodes Examined' and 'Sequence Number-Central' respectively from raw SEER data, both of which had nominal values encoded within the same attribute, with the latter also encoding non-malignant tumors. After performing various steps of data transformation and feature selection, the data was reduced to 46,389 instances of lung cancer patients and 13 attributes (excluding the outcome attribute).

Predictive analytics

For predictive analytics, binary outcome attributes for 6-month, 9-month, 1-year, 2-year, and 5-year survival were derived from survival time. The dataset of 5-year survival was subsequently filtered to generate five new datasets for modeling conditional survival after five years of diagnosis, given that the patient has already survived 3 months, 6 months, 12 months, 18 months, and 24 months.

Many predictive modeling techniques were found to give good accuracy measures that were statistically indistinguishable with the best accuracy. From amongst those, we chose the model based on alternating decision trees with additional logistic modeling on top for better calibration. Ten-fold cross validation was used to estimate the accuracy of all the ten models. Table 3 presents the results for all the models (only accuracy and AUC included here for simplicity), along with the distribution of survived and not-survived patients in the data used to build the corresponding model.

Table 3: Model classification performance (10-fold cross-validation)

Model	*%Survived*	*%Not survived*	*%Model Accuracy*	*AUC*
$5yr$	12.8	87.2	91.8	0.924
$2yr$	23.4	76.6	85.6	0.859
$1yr$	40.2	59.8	74.5	0.796
$9mon$	48.8	51.2	71.0	0.779
$6mon$	60.1	39.9	69.8	0.765
$5yr \mid 3mon$	16.9	83.1	89.8	0.912
$5yr \mid 6mon$	21.4	78.6	87.3	0.900
$5yr \mid 12mon$	31.9	68.1	82.1	0.875
$5yr \mid 18mon$	43.9	56.1	78.1	0.850
$5yr \mid 24mon$	54.9	45.1	76.1	0.830

Association rule mining

For association rule mining analysis, all missing/unknown values were removed, since we are interested in finding segments with precise definitions in terms of patient attributes. The survival time (in months) was chosen as the target attribute for the HotSpot algorithm. The dataset had 13,033 instances, 13 input patient attributes, and 1 target attribute. The average survival time in the entire dataset (t_{avg}) was 24.45 months. So it would be interesting to find segments of patients where the average survival time is significantly higher than or significantly lower than 24.45 months. Two independent analyses were performed to find segments in which average survival time was higher and lower than overall average survival, represented in the form of association rules. $Lift$ of a rule/segment is a multiplicative metric that measures the relative improvement in the target (here survival time) as compared to the average value of the target across the entire dataset.

For association rule mining analysis on conditional survival data, a new dataset was constructed using only the cases in which the patient survived at least 12 months from the time of diagnosis. The conditional survival dataset had 6,788 instances, the same 13 input patient attributes and 1 target attribute. The average survival time in the conditional survival dataset was 42.54 months. So, the above analysis was repeated on the conditional survival dataset with $t_{avg} = 42.54$.

Tables 4 and 5 present the non-redundant association rules obtained with 'higher' and 'lower' mode respectively. Tables 6 and 7 present the same for the conditional survival dataset.

Lung cancer outcome calculator

The web-tool is available at http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator, and uses the following 13 attributes:

1. **Age at diagnosis**: Numeric age of the patient at the time of diagnosis of lung cancer.

2. **Birth place**: The place of birth of the patient. There are 198 options available to select for this attribute (based on the values observed in the SEER database).

3. **Cancer grade**: A descriptor of how the cancer cells appear and how fast they may grow and spread. Available options are - well-differentiated, moderately differentiated, poorly differentiated, undifferentiated, and undetermined.

4. **Diagnostic confirmation**: The best method used to confirm the presence of lung cancer. Available options are - positive histology, positive cytology, positive microscopic confirmation (method unspecified), positive laboratory test/marker study, direct visualization, radiology, other clinical diagnosis, and unknown if microscopically confirmed.

5. **Farthest extension of tumor**: The farthest documented extension of tumor away from the lung, either by contiguous extension (regional growth) or distant metastases (cancer spreading to other organs

Table 4: Non-redundant association rules denoting segments where average survival time is significantly higher than 24.45 months

Segment description	*Avg. survival time*	*Segment size*	*Lift*
The tumor is well-differentiated and localized, regional lymph nodes examined is between 4 and 17, age of the patient at time of diagnosis is less than 79, current tumor is patient's first or second tumor, and resection of lobe/bilobectomy is performed by the surgeon	68.18	100	2.79
The tumor is localized, age of patient is between 39 and 52, number of regional lymph nodes examined is between 1 and 14, and resection of lobe/bilobectomy is performed by the surgeon	68.11	100	2.79
Tumor is well-differentiated, number of regional lymph nodes examined is less than 15, resection of lobe/bilobectomy is performed, and regional lymph nodes are removed	66.83	101	2.73
Tumor is localized, age of patient is between 41 and 52, tumor is confined to one lung, and resection of lobe/bilobectomy is performed	66.26	111	2.71
Patient is born in Hawaii, patient's age is less than 76, there is no lymph node involvement, and resection of lobe/bilobectomy is performed	64.98	106	2.66
Tumor is localized, patient is born in Hawaii, patient's age is less than 83, and surgery is performed	63.96	101	2.62
Tumor is well-diffentiated, number of lymph nodes examined is between 7 and 18, there is no lymph node involvement, and patient's age is less than 81	63.86	101	2.61
Tumor is localized, patient is born in Connecticut, tumor is confined to one lung, number of lymph nodes examined is greater than 2, and resection of lobe/bilobectomy is performed	63.10	103	2.58
Tumor is well-differentiated, there is no lymph node involvement, patient's age is less than 76, and intrapulmonary/ipsilateral hilar/ipsilateral peribronchial nodes are removed	62.16	100	2.54
Tumor is localized (confined to one lung), patient is born in Hawaii and is less than 82 years old	60.38	101	2.47
Tumor is localized (confined to one lung), patient is born in Hawaii, and cancer is confirmed by positive histology	60.18	103	2.46
Tumor is localized, patient is born in California, and resection of lobe/bilobectomy is performed by the surgeon	58.71	100	2.40

Table 5: Non-redundant association rules denoting segments where average survival time is significantly lower than 24.45 months

Segment description	*Avg. survival time*	*Segment size*	*Lift*
Tumor has metastasized and is poorly differentiated, lymph nodes are involved in metastasis, and no lymph nodes are removed	5.21	100	4.69
Tumor has metastasized and is poorly differentiated, no surgery is performed, and the patient is born in Hawaii	5.67	110	4.31
Tumor has metastasized, no surgery is performed, cancer is confirmed by positive histology, and patient is born in Hawaii	5.73	128	4.26
Tumor has metastasized, surgery is contraindicated and not performed, and cancer is confirmed by positive histology	5.78	132	4.23
Pleural effusion has taken place, tumor is poorly differentiated, subcarinal/carinal/mediastinal/tracheal/aortic/ pulmonary ligament/pericardial lymph nodes are involved, and no surgery is performed	7.53	205	3.25
Pleural effusion has taken place, cancer is confirmed by positive cytology, surgery is not recommended and hence not performed	8.60	112	2.84

far from primary site through bloodstream or lymphatic system). There are 20 options available to select for this attribute. The original SEER name for this attribute is 'EOD extension'.

6. **Lymph node involvement**: The highest specific lymph node chain that is involved by the tumor. Cancer cells can spread to lymph nodes near the lung, which are part of the lymphatic system (the system that produces, stores, and carries the infection-fighting-cells). This can often lead to metastases. There are 8 options available for this attribute. The original SEER name for this attribute is 'EOD Lymph Node Involv'.

Table 6: Non-redundant association rules denoting segments in the conditional survival dataset where average survival time is significantly higher than 42.54 months

Segment description	*Avg. survival time*	*Segment size*	*Lift*
Tumor is well-differentiated and localized, patient's age is less than 71, less than 13 regional lymph nodes are examined, and resection of lobe/bilobectomy is performed	72.92	104	1.71
Tumor is well-differentiated and localized (confined to one lung), patient's age is less than 71, surgery is performed, less than 8 regional lymph nodes are examined	72.50	103	1.70
Tumor is well-differentiated, patient's age is less than 84, regional lymph nodes are removed, no lymph node involvement, no radiation therapy, and resection of lobe/bilobectomy is performed	71.95	100	1.69
Tumor is localized (confined to one lung), patient's age is between 41 and 52, surgery is performed, and resection of lobe/bilobectomy is performed	69.66	105	1.64
Tumor is well-differentiated, patient's age is less than 79, no lymph node involvement, between 5 and 9 regional lymph nodes are examined	68.44	100	1.61
Tumor is localized (confined to one lung), patient's age is less than 77, patient is born in Connecticut, and resection of lobe/bilobectomy is performed	67.99	119	1.60
Patient's age is less than 76, patient is born in Hawaii, no lymph node involvement, and resection of lobe/bilobectomy is performed	67.81	101	1.59
Patient's age is less than 75, patient is born in California, no lymph node involvement, and resection of lobe/bilobectomy is performed	65.37	102	1.54
Tumor is localized, no regional lymph nodes removed, and resection of lobe/bilobectomy is performed	62.14	102	1.46

Table 7: Non-redundant association rules denoting segments in the conditional survival dataset where average survival time is significantly less than 42.54 months

Segment description	*Avg. survival time*	*Segment size*	*Lift*
Tumor is undifferentiated and has metastasized, subcarinal/ carinal/mediastinal/tracheal/ aortic/ pulmonary ligament/ pericardial lymph nodes are involved, no regional lymph nodes are removed, and no surgery is performed	17.18	100	2.48
Tumor is spread, surgery not recommended, patient is born in Iowa	20.28	137	2.10
Tumor is spread and undifferentiated, surgery not recommended, subcarinal/ carinal/mediastinal/tracheal/ aortic/ pulmonary ligament/ pericardial lymph nodes are involved, and cancer is confirmed by positive histology	20.35	124	2.09
Pleural effusion has taken place, and tumor is poorly differentiated	22.96	101	1.85

7. **Type of surgery performed**: The surgical procedure that removes and/or destroys cancerous tissue of the lung, performed as part of the initial work-up or first course of therapy. There are 25 options available for this attribute, like cyrosurgery, fulguration, wedge resection, laser excision, pneumonectomy, etc. The original SEER name for this attribute is 'RX Summ-Surg Prim Site'.

8. **Reason for no surgery**: The reason why surgery was not performed (if not). Available options are - surgery performed, surgery not recommended, contraindicated due to other conditions, unknown reason, patient or patient's guardian refused, recommended but unknown if done, and unknown if surgery performed.

9. **Order of surgery and radiation therapy**: The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation. Available options are - no radiation and/or surgery, radiation before surgery, radiation after surgery, radiation both before and after surgery, intraoperative radiation therapy, intraoperative radiation with other radiation given before/after surgery, and sequence unknown but both surgery and radiation were given. The original SEER name for this attribute is 'RX Summ-Surg/Rad Seq'.

10. **Scope of regional lymph node surgery**: It describes the removal, biopsy, or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event. There are 8 options available for this attribute. The original SEER name for this attribute is 'RX Summ-Scope Reg 98-02'.

11. **Cancer stage**: A descriptor of the extent to which the cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis, etc. Available options are - in situ (noninvasive neoplasm), localized (invasive neoplasm confined to the lung), regional (extended neoplasm), distant (spread neoplasm), and unstaged/unknown. The original SEER name for this attribute is 'Summary Stage 2000 (1998+)'.

12. **Number of malignant tumors in the past**: An integer denoting the number of malignant tumors in the patient's lifetime so far. This attribute is derived from the SEER attribute 'Sequence Number-Central', which encodes both numeric and categorical values for both malignant and benign tumors within a single attribute. As part of the preprocessing, the original SEER attribute was split into numeric and nominal parts, and the numeric part was further split into 2 attributes representing number of malignant and benign tumors respectively.

13. **Total regional lymph nodes examined**: An integer denoting the total number of regional lymph nodes that were removed and examined by the pathologist. This attribute was derived by extracting the numeric part of the SEER attribute 'Regional Nodes Examined'.

Figure 3 shows a screenshot of the lung cancer outcome calculator. This calculator is widely accessed from more than 15 countries, including many medical schools and hospitals. A previous version of this calculator were presented in [4]. The current calculator incorporates faster models as described in this chapter, and has a redesigned interface. It allows the user to enter values for the above-described 13 attributes and get patient-specific risk. For all the ten models, it also shows the distribution of survived and not survived patients in the form of pie charts. Upon entering the patient attributes on the website, the patient-specific risk calculated by all the ten models is depicted along with the healthy and sick patient risk, which are essentially the median risk of death of patients who actually survived and did not survive respectively, as calculated by the corresponding model. It generates bar charts corresponding to each of the ten models, and each of them has three bars. The middle bar denotes the patient-specific risk, and the left (right) bars denote the healthy (sick) patient risk. The patient-specific risk is thus put in context of the healthy and sick patient risk for an informative comparison.

Any data-driven tool like this in the field of healthcare has a disclaimer about its use, stating that it is meant to complement and not replace the advice of a medical doctor. Many such calculators are becoming popular in healthcare.

Other applications of big data analytics in healthcare

We will conclude with a sampling of some other applications of big data in healthcare. There has been abundant work on mining electronic health records in addition to what is described in this chapter. Some of these include mining data from a particular hospital [16], ACS NSQIP (American College of Surgeons National Surgical Quality Improvement Program) [17], and UNOS (United Network for Organ Sharing) [18].

Apart from electronic health records, a very important source of healthcare data is social media. We are in the midst of a revolution in which, using social media, people interact, communicate, learn, influence and make decisions. This data includes multi-way communications and interactions on social media (e.g., Facebook, Twitter), discussion forums and blogs in the area of health-care, public-health and medicine. The emergence and ubiquity of online social networks has enriched this data with evolving interactions and communities at mega-scale and people are turning to social media for various kinds of health-care guidance and knowledge, including proactive and preventive care. Patients with like conditions - often chronic conditions, such as flu, cancer, allergy, multiple sclerosis, diabetes, arthritis, ALS, etc. find patients with the same condition on these networking sites and in public forums. And these virtual peers can very much become a key guiding source of data unlike in the past, when all information emanated from physicians. This big data, being produced in social media domain offers a unique opportunity for advancing, studying

the interaction between society and medicine, managing diseases, learning best practices, influencing policies, identifying best treatment, and in general, to empower people. It thus has numerous applications in public health informatics, and we are already seeing several studies in this domain [19, 20, 21].

Technological advances in sensors, micro- and nano-electronics, advanced materials, mobile computing, etc. have had an immense impact towards enabling future Internet of Things (IoT) applications in several fields including healthcare. We are currently witnessing a rapid adoption of wearable devices under the IoT paradigm for a variety of healthcare applications [22]. These wearable and implantable sensors along with smart phones that are ubiquitously used all over the world form another source of healthcare big data, and provide unprecedented opportunities for continuous healthcare monitoring and management.

The field of genomics is another area where big data analytics can play an important role. It is well recognized that in genomics and life sciences, almost everything is based on complex sequence-structure-function relationships, which are far from being well understood. With genomic sequencing becoming progressively easier and affordable, we have arrived at a point in time where huge amounts of biological sequence data have become increasingly available, thanks to the advent of Next Generation Sequencing (NGS). Functional interpretation of genomic data is the major task in fundamental life science. Research results in this area in turn feed research in other important areas such as cell biology, genetics, immunology and disease-oriented fields. There has been a lot of work in bioinformatics on sequence data in terms of computationally mining the genomic sequences for interesting insights such as homology detection [23, 24]. Furthermore, biological sequencing data also ushers an era of personal genomics enabling individuals to have their personal DNA sequenced and studied to allow more precise and personalized ways of anticipating, diagnosing and treating diseases on an individual basis (precision medicine). Genome assembly and sequence mapping techniques [25, 26] form the first step of this process by compiling the overlapping reads into a single genome. While it is a fact that personalized medicine is becoming more and more common, it is nonetheless in its infancy and we are still far from realizing the dream of personalized medicine by optimally utilizing the flood of genomic data that we are able to collect now. Clearly, computational sequence analysis techniques are critical to unearth the hidden knowledge from such genomic sequence data, and big data analytics is expected to play a big role in that. For further reading on big data analytics in genomics, the following articles are recommended [27, 28, 29].

Summary

Big data has become a very popular term denoting huge volumes of complex datasets generated from various sources at a rapid rate. This big data potentially has immense hidden value that needs to be discovered by means of intelligently-designed analysis methodologies that can scale for big data, and all of that falls in the scope of big data analytics. In this chapter, we have looked at some of the big data challenges in general, and also what they mean in context of healthcare. As an example on big data mining in healthcare, some recent works dealing with the use of predictive analytics and association rule mining on lung cancer data from SEER were discussed, including a lung cancer outcome calculator that has been deployed as a result of this analytics. Finally, we also briefly looked at a few other healthcare-related areas where big data analytics is playing an increasingly vital role.

Acknowledgments

The authors would like to thank the SEER program to make the limited-use data available for the works described in this chapter.

References

[1] T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[2] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, 2015.

[3] S. S. Magill, J. R. Edwards, W. Bamberg, Z. G. Beldavs, G. Dumyati, M. A. Kainer, R. Lynfield, M. Maloney, L. McAllister-Hollod, J. Nadle, S. M. Ray, D. L. Thompson, L. E. Wilson, and S. K. Fridkin, "Multistate point-prevalence survey of health care-associated infections," *New England Journal of Medicine*, vol. 370, no. 13, pp. 1198–1208, 2014. PMID: 24670166.

[4] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on seer data," in *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (BIOKDD)*, (New York, NY, USA), pp. 1–9, ACM, 2011.

[5] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on seer data," *Scientific Programming*, vol. 20, no. 1, pp. 29–42, 2012.

[6] A. Agrawal and A. Choudhary, "Identifying hotspots in lung cancer data using association rule mining," in *2nd IEEE ICDM Workshop on Biological Data Mining and its Applications in Healthcare (BioDM)*, pp. 995–1002, 2011.

[7] A. Agrawal and A. Choudhary, "Association rule mining based hotspot analysis on seer lung cancer data," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 2, no. 2, pp. 34–54, 2011.

[8] L. A. G. Ries and M. P. Eisner, *Cancer of the lung*, ch. 9, pp. 73–80. National Cancer Institute, SEER Program, 2007.

[9] SEER, "Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) limited-use data (1973-2006)." National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, 2008. released April 2009, based on the November 2008 submission.

[10] A. Agrawal, M. Patwary, W. Hendrix, W.-k. Liao, and A. Choudhary, *High performance big data clustering*, pp. 192–211. IOS Press, 2013.

[11] Y. Xie, D. Palsetia, G. Trajcevski, A. Agrawal, and A. Choudhary, "Silverback: Scalable association mining for temporal data in columnar probabilistic databases," in *Proceedings of 30th IEEE International Conference on Data Engineering (ICDE), Industrial and Applications Track*, pp. 1072–1083, 2014.

[12] Y. Xie, D. Honbo, A. Choudhary, K. Zhang, Y. Cheng, and A. Agrawal, "Voxsup: a social engagement framework," pp. 1556–1559, ACM, 2012. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD) (Demo paper).

[13] A. R. Ganguly, E. Kodra, A. Agrawal, A. Banerjee, S. Boriah, S. Chatterjee, S. Chatterjee, A. Choudhary, D. Das, J. Faghmous, P. Ganguli, S. Ghosh, K. Hayhoe, C. Hays, W. Hendrix, Q. Fu, J. Kawale, D. Kumar, V. Kumar, W.-k. Liao, S. Liess, R. Mawalagedara, V. Mithal, R. Oglesby, K. Salvi, P. K. Snyder, K. Steinhaeuser, D. Wang, and D. Wuebbles, "Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques," *Nonlinear Processes in Geophysics*, vol. 21, pp. 777–795, 2014.

[14] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science," *APL Materials*, vol. 4, no. 053208, pp. 1–10, 2016.

[15] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Honbo, A. Agrawal, and A. Choudhary, "Muses: a multilingual sentiment elicitation system for social media data," *IEEE Intelligent Systems*, vol. 99, pp. 1541–1672, 2013.

[16] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," *Journal of the American Medical Informatics Association*, vol. 20, pp. e118–e124, 2013. JSM and AA are co-first authors.

[17] A. Agrawal, R. Al-Bahrani, R. Merkow, K. Bilimoria, and A. Choudhary, "Colon surgery outcome prediction using acs nsqip data," in *Proceedings of the KDD Workshop on Data Mining for Healthcare (DMH)*, pp. 1–6, 2013.

[18] A. Agrawal, R. Al-Bahrani, J. Raman, M. J. Russo, and A. Choudhary, "Lung transplant outcome prediction using unos data," in *Proceedings of the IEEE Big Data Workshop on Bioinformatics and Health Informatics (BHI)*, pp. 1–8, 2013.

[19] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: Demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1474–1477, 2013.

[20] Y. Xie, Z. Chen, Y. Cheng, K. Zhang, A. Agrawal, W.-k. Liao, and A. Choudhary, "Detecting and tracking disease outbreaks by mining social media data," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2958–2960, 2013.

[21] K. Lee, A. Agrawal, and A. Choudhary, "Mining social media streams to improve public health allergy surveillance," in *Proceedings of IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM)*, pp. 815–822, 2015.

[22] J. Andreu-Perez, D. R. Leff, H. Ip, and G.-Z. Yang, "From wearable sensors to smart implants—toward pervasive and personalized healthcare," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2750–2762, 2015.

[23] A. Agrawal and X. Huang, "Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 1, pp. 194–205, 2011.

[24] A. Agrawal and X. Huang, "Psiblast_pairwisestatsig: Reordering psi-blast hits using pairwise statistical significance," *Bioinformatics*, vol. 25, no. 8, pp. 1082–1083, 2009.

[25] X. Huang and A. Madan, "Cap3: A dna sequence assembly program," *Genome research*, vol. 9, no. 9, pp. 868–877, 1999.

[26] S. Misra, A. Agrawal, W.-k. Liao, and A. Choudhary, "Anatomy of a hash-based long read sequence mapping algorithm for next generation dna sequencing," *Bioinformatics*, vol. 27, no. 2, pp. 189–195, 2011.

[27] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, *et al.*, "Big data: The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

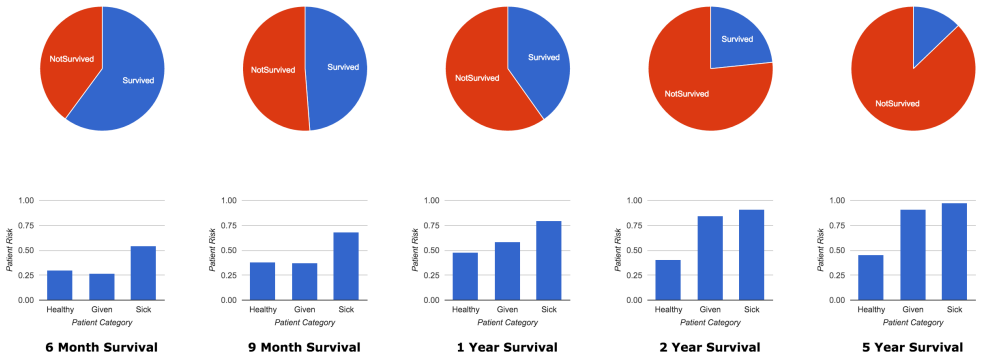
[28] A. ODriscoll, J. Daugelaite, and R. D. Sleator, "big data, hadoop and cloud computing in genomics," *Journal of biomedical informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[29] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

Lung Cancer Outcome Calculator

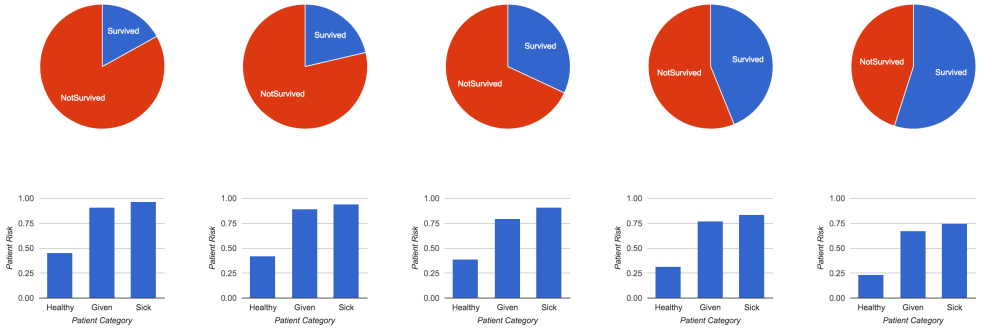
Disclaimer: The outcome calculator results are estimates based on data consisting of a large number of lung cancer records. All results are provided for informational purposes only, in furtherance of the developers' educational mission, and are not meant to replace the advice of a medical doctor. The developers may not be held responsible for any medical decisions based on this outcome calculator.

Welcome to our online lung cancer outcome calculator. The calculator is based on data obtained from Surveillance Epidemiology and End Results (SEER) of the National Cancer Institute which is an authoritative repository of cancer statistics in the United States. The data contains lung cancer records of nearly 50000 patients. The calculator estimates the risk of mortality after 6 months, 9 months, 1 year, 2 year, and 5 years of diagnosis, using a small non-redundant subset of 13 patient attributes which were carefully selected using attribute selection techniques. The graph shows the five risk values obtained for specific attribute values, which are shown below the graph. To obtain risk values for a new set of attribute values, please change the attribute values below and click on the submit button.



6 Month Survival | **9 Month Survival** | **1 Year Survival** | **2 Year Survival** | **5 Year Survival**

5 Year Conditional Survival



Given 3 Months Survival | **Given 6 Months Survival** | **Given 12 Months Survival** | **Given 18 Months Survival** | **Given 24 Months Survival**

For a given time interval T,
Healthy patient risk - Median risk of death of patients who survived after time T, as calculated by our calculator.
Patient risk - This corresponds to the risk of death of a patient after time T, calculated based on the provided values of the patient attributes.
Sick patient risk - Median risk of death of patients who did not survive after time T, as calculated by our calculator.

Age	75	Reason for no surgery	Surgery performed
Birth Place	Afghanistan	Order of surgery and radiation therapy	No radiation and/or surgery
Cancer grade	Grade I (well-differentiated)	Scope of regional lymph node surgery	No regional lymph nodes removed
Diagnostic confirmation	Positive histology	Cancer stage	In situ (Noninvasive neoplasm)
Farthest extension of tumour	In situ (Noninvasive/intraepithelial)	Number of malignant tumors in the past	0
Lymph node involvement	No lymph node involvement	Total regional lymph nodes examined:	0
Surgery of primary site	No surgery		

Get Mortality Risks

Developed by Ankit Agrawal and Alok Choudhary

Publications:

A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," Scientific Programming, vol. 20, no. 1, pp. 29-42, 2012. [url]

A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "A lung cancer outcome calculator using ensemble data mining on SEER data," in Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (BIOKDD), 2011, pp. 1-9. [url]

Center for Ultra-scale Computing and Information Security (CUCIS), EECS Department, Northwestern University, Evanston, IL 60208, USA

Figure 3: Screenshot of the Lung Cancer Outcome Calculator (available at http://info.eecs.northwestern.edu/LungCancerOutcomeCalculator).