McMaster eBusiness Research Centre

Using Big Data & Analytics to Predict Hospital Re-Admissions

By

Megan Nagpal and Reza Samavi

McMaster eBusiness Research Centre (MeRC)

WORKING PAPER No. 56
May 2016

USING BIG DATA & ANALYTICS TO PREDICT HOSPITAL RE-ADMISSIONS

By

Meghan Nagpal and Reza Samavi

MeRC Working Paper # 56

May 2016

DeGroote School of Business

McMaster University

Hamilton, Ontario, L8S 4M4

Canada

samavir@mcmaster.ca

LIST OF ABBREVIATIONS

CIHI Canadian Institute for Health Information
CIS Clinical Information System
DIKW Data-Information-Knowledge-Wisdom Pyramid
HER Electronic Health Record
ETL Extract, transform, load
M4CVD Mobile Machine-Learning Model for Monitoring Cardiovascular Disease
NLP Natural Language Processing
PPV Positive Predictive Value
SIEM Security Information and Event Management

1. INTRODUCTION

"Big data" is a term often used to describe very large data sets, ranging from terabytes to exabytes, which are often complex (Chen, Chiang, & Storey, 2012). SAS, an American developer of advanced analytical software, defines *big data* as a term to describe "*large volumes of data, both structured and unstructured, that inundates business on a day-to-day basis*". Industry analyst Doug Laney suggested that a big data system should carry the attributes of **volume, velocity,** and **variety** (SAS, n.d.). In addition to those three attributes of big data, SAS also considers **variability** and **complexity.** Table 1 describes each attribute as defined by SAS:

Table 1. Attributes of Big Data as defined by SAS

Attribute	Definition
Volume	Large amounts of data stored on an electronic system.
Velocity	Data streamed in a fast, and timely manner.
Variety	Data comes in all formats, structured and unstructured. Structured data could consist of numeric and/or binary values in a traditional database. Unstructured data could be any other form of input such as a text document or audio file.
Variability	Data can flow at inconsistent rates and there can be times where there are periodic peaks of data flow.
Complexity	Data can come from a variety of sources in which the complexities stem from matching, cleansing, and transforming the data.

SAS stresses that the amount of data is not important, but it is what organizations do with the data that is important. This is where analytics is added to the equation. As defined by IBM (Cortada, Gordon, & Lenihan, 2012), "*analytics is the systematic use of data and related business insights developed through applied analytical disciplines to drive fact-based decision making for planning, management, measurement, and learning. Analytics may be descriptive, predictive, or prescriptive."*

Currently, we are living in an era where data is readily available for businesses and services in almost every industry. The Data-Information-Knowledge-Wisdom (DIKW) Pyramid has historically been able to show the process in which data is transferred into wisdom in which decisions are made for the well-being of a population[1]. However, in this era of Big Data & Analytics, there has been a shift in which decisions are made through predictive-modelling through data as opposed to hypotheses' from knowledge (Batra, 2014). As mentioned in the quote by IBM (Cortada, Gordon, & Lenihan, 2012), analytics are descriptive, predictive, or prescriptive. This can be translated into a three-phased approach to utilizing analytics in which we answer the following questions:

[1] This is a hierarchy of how data is translated into wisdom. More information can be found on Wikipedia (https://en.wikipedia.org/wiki/DIKW_Pyramid)

1. *What can analytics tell us about our current state?*
2. *What can we predict about the future from these analytics?*
3. *What decisions can be made based off of the predictions made from these analytics?*

With the many technologies that are able to harness large volumes of data, the big data & analytics revolution stands to grow bigger in almost every industry (SAS, n.d.). Some examples of industries which are key beneficiaries of big data & analytics include banking, education, government, manufacturing, retail, media, small & midsize business, and health care. All industries benefit from big data for being able to analyze current trends, predicting future outcomes based on these trends, and being able to make decisions based off of these trends.

This paper specifically examines the role of big data in health care and will specifically examine the necessary technical & business requirements to build predictive models for identifying which patients are at risk for hospital re-admissions. The importance of predicting which patients are at risk for re-admissions allows healthcare providers to adjust discharge plans to minimize this risk.

This paper will first examine the role of big data & analytics in healthcare, give a brief history of big data and analytics, identify risk factors for hospital re-admission based on a literature review, and identify big data techniques and architectures required for such an analytical system from research publications. The benefit of creating an analytical system to predict which patients are most likely to be re-admitted to hospital following discharge is that healthcare providers can personalize care plans. Such a personalized approach can mitigate the risk of re-admission and potentially save costs for the healthcare system.

2. ROLE OF BIG DATA & ANALYTICS IN HEALTH CARE

McKinsey estimates that the application of Big Data & Analytics in healthcare can result in a savings of $300 million per a year in the United States (Raghupathi & Raghupathi, 2014). SAS has already identified health care as an industry which stands to gain from the use of big data. Beyond personalized patient care, predictive analytics can be used in health care settings for patient profiling, clinical operations, research & development, public health, evidence-based medicine, genomic analysis, remote monitoring, fraud prevention, and which patients are at risk for re-admission (Raghupathi & Raghupathi, 2014).

Referring to the five attributes of a big data system, Table 2 gives examples of how these attributes can be applied to the context of health care:

Table 2. Examples of health care applications of big data based on SAS' five attributes of a big data system

Attribute	Implications in the context of health care
Volume	With large adaptation of Electronic Health Records (EHR), there is plenty of opportunity to harness data from these records, as opposed to from paper records. As well, with the rise of smartphone applications and medical devices which send information to healthcare providers electronically, such as blood glucose monitors and pedometers, there is more opportunity to obtain data without the presence of a clinician (Centres for Medicare & Medicaid Services, 2012).

Velocity	EHRs and medical sensory devices give the opportunity for records to be transferred instantly and also allow for providers to make decisions in a timely manner. Timeliness also reduces medical errors, reduces duplicate lab requests, and reduces treatment delays (Centres for Medicare & Medicaid Services, 2012).
Variety	EHR systems can contain structured data such as contact information, demographical information, laboratory results, or immunization history (Centres for Medicare & Medicaid Services, 2012). Unstructured data can include medical notes, discharge summaries, and laboratory images & notes.
Variability	Hospital dashboards, integrated with EHRs may see peaks where they is higher flow of data. There may be periods of times where there is a peak patient flow, potentially in times of disease outbreak (Centres for Medicare & Medicaid Services, 2012).
Complexity	Health data is complex. Clinician teams can be working with different EHR systems among themselves, which could capture different types of information. Clinicians may also be observing data from medical devices and smartphone applications leading to a complex set of data (Centres for Medicare & Medicaid Services, 2012).

The availability of big data has resulted in a more consumer-driven society, in which analytics are able to help businesses predict the needs of their consumers. And with the implementation of Clinical Information Systems (CIS) which utilize Electronic Health Records (EHR) that can capture large volumes of data, health care is no exception to this new trend. Considering that there has been a shift towards patient-centered care and that patients have higher access to information than they have had in the past, big data & analytics has given the healthcare industry the ability to analyze current trends and to make predictions that can assist clinicians with providing personalized care for all patients, and potentially create better outcomes. Analytics does provide the opportunity to create benchmarks for new and effective treatments based on current trends. Ultimately, this allows providers to comply with treatments based on these benchmarks (Raghupathi & Raghupathi, 2013).

The benefit of analytics lies in the prescriptive outcomes made through descriptions of current trends. By current trends, data analysts are able to identify certain patterns among particular groups of patients and be able to prescribe a course of action to identify those problems, which is a prescriptive outcome. However, predictions are the bridge between current descriptions to prescriptive outcomes and this is why it is necessary that analytics provide accurate predictive models, or algorithms which could predict which patients are at risk for being re-admitted. Analytical techniques to create these predictive models will be discussed in this paper.

3. HISTORY OF BIG DATA & ANALYTICS

The potential of Big Data & Analytics was realized in the early 2000's with the emergence of web-based search engines such as Google and Yahoo!, as well as with the rise of e-commerce sites such as Amazon (Chen, Chiang, & Storey, 2012). Initially, research centered on unstructured data delivered via the web which helped organizations such as Google Analytics observe patterns and

trends among web-users and helped organizations with website design, product placement optimization, market analysis, and product recommendations (Chen, Chiang, & Storey, 2012).

The emergence of social web applications such as blogs, forums, social networks, social media, and social games created a larger atmosphere to gain larger volumes of data from a diverse population in real time (Chen, Chiang, & Storey, 2012). The nature of businesses changed in the sense that marketing techniques involved a two-way approach with increased dialogue between customers and businesses (Chen, Chiang, & Storey, 2012) and therefore, allowing businesses to obtain more data on a larger scale and make predictions and recommendations accordingly.

The current decade sees a large increase of mobile devices including the use of smartphones and tablets, as well as other internet-enabled sensory devices that use technologies such as radio-frequency identification, barcode scanning, location detection, and motion sensing to name a few (Chen, Chiang, & Storey, 2012). The large volumes of data provide more sources to be able to understand the current state and be able to make predictions and recommendations based off of a subject matter.

4. IDENTIFYING RISK FACTORS FOR RE-ADMISSION

The Canadian Institution for Health Information (CIHI) uses health indicators, or single measures, which are reported regularly. These measures provide information about population health and/or a health system performance. This information is used by provincial/territorial governments, health authorities, and facilities track their performance and progress over time (Canadian Institute for Health Information, 2016). CIHI currently has over 100 indicators in its library, seven which specifically pertain to hospital re-admission. This paper will examine the predictors of four of these CIHI indicators; *all patients re-admitted to hospital* (30 day re-admission rate for medical, surgical, and obstetrics patients), *obstetric patients re-admitted to hospital* (30 day re-admission rate for obstetric patients), *patients aged 19 and under re-admitted to hospital* (30 day re-admission rate for pediatric patients), and *surgical patients re-admitted to hospital* (30 day re-admission rate following surgery).

4.1 All Patients Re-Admitted to Hospital

From 2013 – 2014, the overall Canadian national re-admission rate within 30 days was 8.9% (Canadian Institute for Health Information, 2016). This indicator measures the risk-adjusted re-admission rate following discharge within 30 days for medical, surgical, and obstetric patients; and patients aged 19 and over. Being able to identify and predict which patients are at-risk for re-admission can help clinicians tailor their course of care for their patient and create a customized discharge plan, suitable to the needs of a patient.

It is important to acknowledge that re-admission to hospital cannot always be prevented and it cannot always be predicted (Allaudeen, Vidyarthi, Maselli, & Auerbach, 2011). Healthcare providers cannot control the progression of the disease and whether or not patients follow their prescribed care plan upon discharge, which are both unavoidable factors towards re-admission (Rumball-Smith & Hider, 2009). However, even though some hospital re-admissions are unavoidable, numerous studies have suggested that it is an appropriate indicator for quality surveillance as it is a measure which is easy to track through hospital administrative data (Rumball-

Smith & Hider, 2009). The re-admission rate is influenced by the quality of care received in hospital, careful discharge planning, and the effectiveness of proper care transition and coordination, and community-based disease management programs (Canadian Institute for Health Information, 2016). One study of acute care patients among six hospitals in Toronto found that re-admitted patients were more likely to stay longer in hospital than they did for their initial stay and that 14% of re-admitted patients died during re-hospitalization (Gruneir, et al., 2011). In the United States, it was estimated that re-admissions to hospital cost Medicare $17.4 billion in 2004 (Jencks, Williams, & Coleman, 2009).

A literature review examined studies which identified risk factors of re-admissions among patient groups.

4.1.1 Demographical Risk Factors

Demographical risk factors for re-admission include age, gender, race, socioeconomic status, insurance status, and marital status. Studies suggest that patients above the age of 60 to be at higher risk of re-admission (Hasan, et al., 2010; Friedman & Basu, 2004; Jencks, Williams, & Coleman, 2009; Lagoe, Noetscher, & Murphy, 2001; Allaudeen, Vidyarthi, Maselli, & Auerbach, 2011). One reason, suggested by Hasan et al. (2010), is because patients in this age demographic tend to be at higher risk for co-morbidities. Gender is also a risk factor as men are more likely than women to experience a re-admission (Jencks, Williams, & Coleman, 2009; Lagoe, Noetscher, & Murphy, 2001). Race was another factor, as black patients are at higher risk for readmission (Friedman & Basu, 2004; Jencks, Williams, & Coleman, 2009; Allaudeen, Vidyarthi, Maselli, & Auerbach, 2011; Lagoe, Noetscher, & Murphy, 2001).

Socioeconomic status could be another risk factor for readmission. Hasan et al. (2010) identified that patients from lower income households were more likely to experience a re-admission, but their study did not suggest a correlation in their predictive model. However, Benbassat & Taragin (2000) and Arbaje et al. (2008) found that patients of lower socioeconomic status are likelier to be re-admitted to hospital after discharge.

Following socioeconomic status, insurance status is a risk for re-admission. Rumball-Smith & Hider (2009) suggest uninsured patients may be discharged prematurely, which in-turn could contribute to a re-admission. This indicator may be a risk factor in the United States but in Canada, hospital stays are covered under the universal, publicly funded system, so insurance status would not likely be a factor in pre-mature discharge. However, costs such as medical equipment, prescriptions, and outside hospital care are not necessarily covered through the universal system and patients would likely have to seek private insurance to cover these costs if required upon discharge.

Benbassat & Taragin (2000) and Arbaje et al. (2008) suggest that unmarried patients are more likely to be living alone and do not necessarily have the familial support following discharge, placing them at higher risk of re-admission.

4.1.2 Clinical Risk Factors

From a clinical perspective, re-admissions are sometimes unavoidable, especially in cases when the disease has progressed far and symptoms are difficult to predict (Rumball-Smith & Hider, 2009). However, there are certain clinical risk factors which can help identify patients who are at risk for re-admission.

Co-morbidities and chronic illness are major risk factors for re-admission (Allaudeen, Vidyarthi, Maselli, & Auerbach, 2011; Hasan, et al., 2010). It is important that patients receive quality care in-hospital to help manage co-morbidities and also receive an effective care plan upon discharge to minimize repercussions. However, an interesting find was that patients who stayed in long-term care facilities upon discharge were less likely to experience a re-admission than patients who were discharged directly to the home (Hasan, et al., 2010; Ashton & Wray, 1996; Rumball-Smith & Hider, 2009). This may be because these patients are able to receive immediate care following discharge unlike patients who are discharged directly to the home. Hasan et al. (2010) also found that hospital injuries, such as adverse drug events, also were a factor for predicting re-admission. These factors are a direct result of the quality of care received in hospital.

A longer length of the initial hospital stay and a higher number of hospitalizations in the last year was also found to be another clinical risk factor for determining hospital re-admission (Hasan, et al., 2010; Jencks, Williams, & Coleman, 2009; Lagoe, Noetscher, & Murphy, 2001). Jencks, Williams, & Coleman (2009) found that patients who experienced a re-admission had an initial hospital stay of 0.6 days longer than that of patients who were not re-admitted to hospital.
Certain morbidities put patients at higher risk for re-admission. End stage renal disease, heart disease, and diabetes were significant risk factors of re-admission (Jencks, Williams, & Coleman, 2009; Allaudeen, Vidyarthi, Maselli, & Auerbach, 2011; Lagoe, Noetscher, & Murphy, 2001; Gruneir, et al., 2011). In addition, Allaudeen, Vidyarthi, Maselli, & Auerbach (2011) identified cancer, weight loss, iron deficiency anemia, or use of high-risk medications such as steroids, narcotics, and anticholinergics as risk factors for re-admission. Allaudeen, Schnipper, Orav, Wachter, & Vidyarthi (2011) and Gruneir et al. (2011) found that pneumonia and gastrointestinal disorders such as bowel obstruction, gastroenteritis, cellulitis, and Clostridium difficile were also risk factors for re-admission.

4.2 Obstetric Patients Re-admitted to Hospital

Overall, the national re-admission rate for obstetric patients within 30 days of discharge was 2% in 2013 (Canadian Institute for Health Information, 2016). Despite the low re-admission rate, it is important to predict which patients are at-risk for re-admission to ensure that there are better outcomes for both mother and child upon delivery. It has been found that women who deliver through caesarian were found more likely to be re-admitted to hospital within 30 days of discharge compared to women who delivered vaginally (Liu, et al., 2002; Liu, et al., 2005). Risk factors for re-admission following a caesarian delivery include pelvic injury, post-partum hemorrhaging, major puerperal infection, and obstetric complications (Liu, et al., 2005).

Interestingly, Liu et al. (2002) found that re-hospitalized women who delivered through caesarian and were discharged within two days were at higher risk of being re-admitted likely because of it being a pre-mature re-admission. Liu et al. (2002) also found that women discharged after five days were also at risk of re-admission likely due to having more complications following delivery.

4.3 Patients Aged 19 and Younger Re-Admitted to Hospital

Overall, the national re-admission rate for patients aged 19 and younger within 30 days of discharge was 6.7% in 2013 (Canadian Institute for Health Information, 2016). Generally, demographical risk factors for re-admission of pediatric patients are similar to those of general medical patients. Black race, older age, and insurance coverage were all risk factors for pediatric re-admission (Feudtner, et al., 2009). Interestingly, Feudtner et al. (2009), found that females were more likely to experience a re-admission than males. From a clinical perspective, longer length of stay, previous hospitalizations, chronic conditions and co-morbidities were also seen as risk factors for re-admission. Specific conditions pertaining the pediatric re-admission include asthma and extremely low birth weight at infancy (Feudtner, et al., 2009).

4.4 Surgical Patients Re-admitted to Hospital

Overall, the national re-admission rate for surgical patients within 30 days of discharge was 6.9% in 2013 (Canadian Institute for Health Information, 2016). Surgical risk factors for re-admission are also similar to that of general medicine patients in regards to longer length of hospital stay and having co-morbidities (Kiran, et al., 2004; Kassin, et al., 2012). Longer hospital stay may be a reflection of prolonged recovery from surgery, which can impact length of hospital stay. Specific co-morbidities associated with surgical re-admission include disseminated cancer, dyspnea, and post-operative open wound (Kassin, et al., 2012)**.** Certain procedures had higher chances of re-admission which include pancreatectomy, colectomy, and liver resection (Kassin, et al., 2012).

5. PREDICTING RE-ADMISSION AND THE LACE ALGORITHM

Researchers have developed algorithms to predict hospital re-admissions. The LACE index score is an algorithm which can help predict which patients are at risk for being re-hospitalized (Gruneir, et al., 2011). The LACE algorithm was derived through a study of 4812 patients discharged from eleven hospitals in Ontario between 2002 and 2006 (van Walraven, et al., 2010). This paper will focus specifically in the LACE index score. In the study, van Walraven et al. (2010) used 48 patient-level and administrative variables which may contribute as a factor towards re-admission. Logical regression was performed to determine which variables were independently associated with re-admission. Four variables were identified as shown in Table 3:

Table 3 LACE Index Definition

Mnemonic	Definition
L	Length of hospital stay in days
A	Acuity of admission
C	Co-morbidity of patient based on Charlson Index Score (Charlson, Pompei, Ales, & Mackenzie, 1987)
E	Emergency department admissions in the past six months

The LACE score is calculated by assigning a score based on the length of stay, whether or not the patient was admitted through emergency, the number and acuity of morbidities, and the number of emergency room visits over the past month, calculated from a range of 0 to 19. A score of 10 or

higher would indicate that the patient is at high risk for re-admission. (Refer to Appendix I to view the exact scoring algorithm.) Internal validations of the LACE index score was done with comparisons among half of the participants in the study conducted by van Walraven et al. (2010). External validations were done by randomly selecting patients discharged into the community from Ontario hospitals using the Discharge Abstract Database (van Walraven, et al., 2010).

Studies in Ontario have suggested that the LACE Index is a useful predictor of which patients are at highest risk of re-admission to hospital, but more research is needed to understand the validity of this score as it is not entirely accurate (Gruneir, et al., 2011; van Walraven, Wong, Forster, & Hawken, 2012; van Walraven, et al., 2010). At most, the LACE Index is a moderate predictor with a c-statistic of 0.61 (Middleton, Lakhanpal, Price, & Butler, 2015). Nonetheless, an analytical system could use this algorithm as a foundation to creating predictive models.

5.1 Predicting Re-admission through the PARR-30 Index

The PARR-30 Index is another algorithm used to predict hospital re-admissions within 30 days of discharge. It is used as an alternative to the LACE index in hospitals in the United Kingdom (Bardsley, Georghiou, Billings, & Blunt, 2012). The PARR-30 index was developed through a series of logistical regressions on variables which are considered to be risk factors of hospital re-admissions (Billings, et al., 2012). It assigns a positive predictive value (PPV), which is a percentage of patients likely to experience a re-admission (Billings, et al., 2012). At a threshold of 50% to classify a high risk patient, Billings et al. (2012) found that 59.2% of patients identified as high-risk were subsequently re-admitted within 30 days. An interesting technique the study conducted by Billings et al. (2012) is that a business case was developed to guide providers with developing interventions to prevent re-admissions. Using case studies also gives clinicians practice with implementing the algorithm and frequent practice can allow them to mentally evaluate if a patient is likely to experience a re-admission during the initial hospital stay. However, the use of automated predictive models does eliminate this need.

Appendix II contains the variables and the scoring algorithm for the PARR-30 index. While the PARR-30 index has been shown to be an effective scoring algorithm for predicting hospital re-admissions, this paper will focus solely on the LACE algorithm.

6. REQUIREMENTS FOR AN HER SYSTEM TO PREDICT HOSPITAL RE-ADMISSIONS

From the identified risk factors of re-admission and the mnemonics of the LACE algorithm, Table 4 describes the necessary requirements of an EHR system to predict hospital re-admissions.

Table 4 Requirements of an EHR system to predict hospital re-admissions

Requirement of EHR System	Recommendation
Stores basic demographical patient information (e.g. age, gender, contact, race, income, marital status, insurance plan)	Certain demographical factors place patients at risk for re-admission. Tracking these data helps identify which patients are at higher risk.
Medical history and co-morbidities, entered in checkboxes and free text	Certain morbidities place patients at higher risk for re-admission. Checkboxes allow for structured data, which is easier for analysis. Free text allows the entering of information not otherwise captured by the checkboxes. A good big data system contains structured and unstructured information.
Patients can fill out information prior to an unexpected admission to hospital	Patients would have already entered sensitive information, such as race or income level, prior to admission. This type of information can be difficult for a triage nurse to gauge upon admission. Allows patients to have control and awareness of their own health.
Hospital administrative data (length of stay, previous hospitalizations)	Key pieces of information necessary for LACE computations.
Clinical information	Practitioner should track all processes during hospitalization and a careful discharge plan. Some information could be in the form of checkboxes, and other information in the form of free text.
Integration & Interoperability	EHR system should be integrated across a regional healthcare network so that medical data is captured should patient seek re-admission at another hospital

6.1 EFFECTIVENESS OF PROPOSED SOLUTION

As mentioned, the LACE Index is a moderately effective predictor of hospital re-admissions (Gruneir, et al., 2011; van Walraven, Wong, Forster, & Hawken, 2012; van Walraven, et al., 2010). The data elements mentioned above all provide structured data so that a system can calculate a

patient's LACE score. However, the above proposed solution would only work in perfect conditions. There are many drawbacks to this solution as described below.

The first assumption is that there exists a system where patients can enter in their personal information prior to a perceived hospital admission. Following this assumption is the assumption that patients would actually take the time to enter in their information prior to an unexpected hospital admission. Unless there is an incentive for patients to enter in their data, it is unlikely that a significant number of patients would. Furthermore, data entered into the system could be skewed because it would be dependent on factors such as access to a computer or access to a regular physician.

Another assumption is that healthcare providers will be able to enter in structured data in real-time during a hospital admission. However, it is easier for physicians to dictate their charts on paper, and have them appended into an EHR (Middleton, Lakhanpal, Price, & Butler, 2015). This creates unstructured data, which may not even be in typed form (e.g. scanned paper records).

Finally, the elements required for a LACE prediction are not necessarily delivered in real-time. Healthcare providers are more likely to enter in the symptoms a patient is feeling, as opposed to a diagnosis upon initial hospitalization. As well, there is no actual way to know a patient's length of stay in hospital until after they are discharged from hospital (Middleton, Lakhanpal, Price, & Butler, 2015).

Considering that the purpose of predicting a hospital re-admission is to improve patient care during the initial hospitalization, a predictive analytical system should be able to predict a patient's length of stay and also be able to predict a diagnosis based on the symptoms provided. Indicators, such as socioeconomic status, can also be predicted through an analytical system.

The LACE index may be an efficient algorithm for predicting hospital re-admissions after discharge, but clearly there are deficiencies in predicting hospital re-admission during the patient's initial stay. The goal of the system should be to predict which patients are at-risk for re-admission during the initial hospital stay. While the LACE index is a good foundation for predicting hospital re-admissions, health researchers can utilize the powers of big data to form new algorithms which can predict hospital re-admissions in real-time. For example, predictive models could predict a patient's length of stay during the initial hospitalization by analyzing the patient's current symptoms and prognosis. This could predict a LACE score before discharge. Moreover, predictions of re-admission can be made based on the acuity and severity of a patient's symptoms. As symptoms change, the likelihood of re-admission also changes. More information about predictive modelling is explained in Section 7.2.

7. TECHNIQUES FOR ANALYZING BIG DATA

Data Lakes[2] store large amounts of data in a central location and can serve as asset to analysts when creating predictive models. Internally, a hospital-wide data lake could include paper charts, notes from physicians and nurses, laboratory orders and results, discharge reports, or referral reports from a previous point of care (Middleton, Lakhanpal, Price, & Butler, 2015). This data lake may also contain patient data, hospital data, and data on diagnoses' (Raghupathi & Raghupathi, 2013). Externally, a data lake may contain data from other sources such as benchmarks, government data, WHO data, insurance companies, patient databases, public health systems, and data from external hospitals and/or clinics (Raghupathi & Raghupathi, 2013). These already provide a rich source of information which can be used to form a data lake.

The main benefit of data lakes is that it allows for data from different sources to be stored in a centralized location and, as a result, allows for analysts to tackle big data projects (Gartner, 2014). However, this big lake also creates a huge pool of raw data, sometimes without any metadata, requiring data scientists to devote a significant amount of effort to clean the data to an understandable format. In healthcare settings, one must assume that it is impossible to have all data be "perfect". There will be pieces of data which can be considered to be "dirty", which could carry the characteristics (but not limited to) of being incorrectly entered, in the wrong format, or pieces of information being missing. Gartner estimates that 10% of IT costs goes towards data integration and quality, with most of these costs attributing to human labour (Kruse, Papotti, & Naumann, 2015). Especially without metadata, health data analysts are left with the task of piecing data from scratch (Gartner, 2014). It is important for health data analysts to factor in the cost of cleaning up and structuring data. There are tools and techniques which do allow for analysis of unstructured data, but one should not underestimate the effort required to structure data into some standardized format prior to performing these techniques. Some techniques to analyze data are discussed below.

Natural Language Processing (NLP) systems mine unstructured text into a structured format for further analysis. Machine learning involves the creation of predictive models, based off of data analysis. Data mining is the ability to extract patterns from large datasets.

A procedure to develop a predictive model using the data in the data lake would be to first mine the unstructured data into structured data using NLP. For known rules, unstructured text could be mapped to structured data through rule-based NLP systems. Otherwise, a statistic-based NLP system could use statistics to map unstructured data. After the process of NLP, machine learning techniques can be used to create predictive models from this data. Following the development of these predictive models, data mining can be used to find associations among clusters of data to give recommendations from these associations. Figure 1 shows a diagram of this procedure with each arrow representing the sequence of events:

[2] A *Data Lake* holds a large amount of raw data in its natural format
Source: http://searchaws.techtarget.com/definition/data-lake

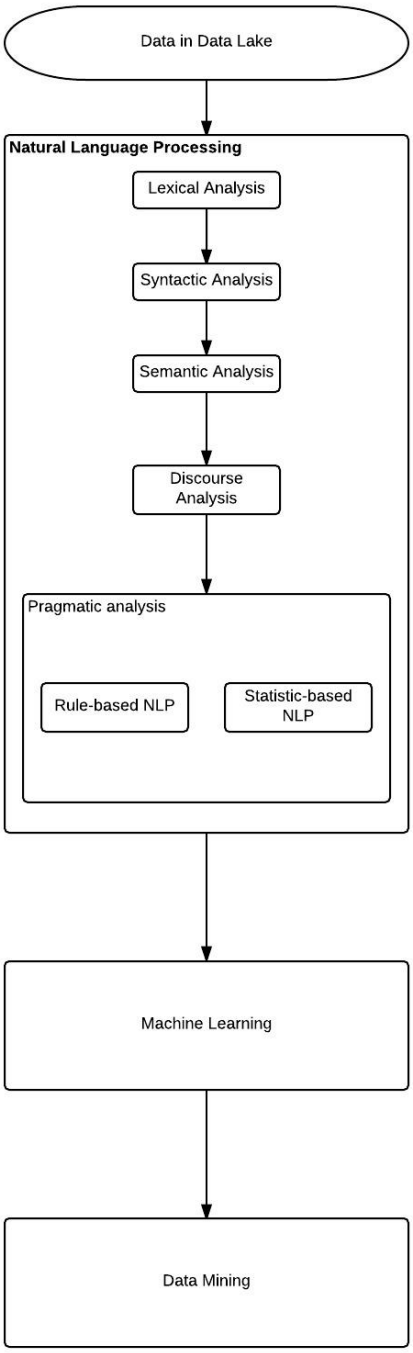


Figure 1 Procedure for creating predictive algorithm from Data Lake

7.1 Natural Language Processing (NLP)

Natural Language Processing allows a big data system to mine unstructured text into a structured format to help provide further analysis. Given that EHR's do have the ability to store information in the form of text boxes, it is almost guaranteed that a hospital data lake would contain unstructured text. NLP would be essential in regards to being able to analyze current trends in data and being able to essentially make accurate predictions. As mentioned, information is more likely to be dictated and transcribed in an unstructured format into an EHR system and NLP has the capability to potentially parse this information in real-time.

In the case of predicting hospital re-admissions, NLP can be useful by identifying patient symptoms in an unstructured block of text and being able to predict a likely diagnosis for the symptoms. Also, being able to process discharge notes can be useful for prescribing care plans for patients in the future.

NLP may also have the ability to extract certain demographic characteristics of a patient. Phrases captured in an EHR such as "cannot afford" or "came with spouse" or "recently travelled to <insert country> to visit family" can offer insight into the characteristics of a patient such as their income, marital status, or race, which an EHR system may not actually capture in a structured format.

There are five steps to NLP as defined by Tutorials Point (2016), described in Table 5:

Table 5 Steps to Natural Language Processing

#	Step	Definition
1	Lexical Analysis	Analyzing the structure of the text in regards to paragraphs, sentences, and words.
2	Syntactic Analysis (Parsing)	Analyzes the arrangement of the words in regards to grammatical correctness in the English language.
3	Semantic Analysis	Analyzes the dictionary meaning of the text.
4	Discourse Analysis	Analyzes the meaning of the sentence prior to the current sentence as the meaning of the first sentence impacts the meaning of the next one.
5	Pragmatic Analysis	Uses real-world knowledge to re-interpret on what the parser actually meant.

7.1.1 Rule-based Natural Language Processor

A rule-based NLP uses rules written by a group of experts to map the unstructured data to those rules. While they are expensive to implement because of the consultation needed by numerous experts, it is easier to pinpoint errors in the system because human intervention can update these rules (Wolniewicz, n.d.). In a clinical setting, a diagnosis could be made based on data from NLP, being matched to the rules.

7.1.2 Statistic-based Natural Language Processor

A statistic-based NLP uses statistics across a large group of data to map the unstructured data based on what was inputted previously. They are cheaper to implement as expert opinion is not necessarily needed to map this information, but there is a higher chance that there will be inaccuracies as errors could be difficult to map (Wolniewicz, n.d.). In a clinical setting, a diagnosis could be made based on statistical data of a symptom, as opposed to using a fixed set of rules. It is not uncommon to see a hybrid-based NLP approach in a clinical-setting, in which both rules and statistics are utilized (Wolniewicz, n.d.).

7.2 Machine Learning

Machine learning uses data analysis to automate analytical model building (SAS, n.d.). It makes predictions based off of iterative computations over a wide range of data. With a large data lake, the possibilities are endless. Machine learning can utilize not only internal data from the hospital to create algorithms and predictive models for hospital re-admissions, but also use external benchmarks and government records to enhance these algorithms.

The benefit in machine learning lies in regards to the fact that predictions are able to be made in real-time, which is a big advantage over the LACE algorithm, which makes a prediction after discharge. As mentioned in Section 6.1, a patient's length of stay can be predicted during the initial hospitalization, based off of the severity of their symptoms. This provides real-time predictions of the likelihood of re-admission.

Machine learning utilizes the data from EHRs which are both structured and unstructured information, with the aid of NLPs. In the case of predicting hospital re-admissions, algorithms to predict a patient's diagnosis which would come useful for calculating a patient's *Charlson index score*, can be generated through machine learning and NLP. Furthermore, predictive models from machine learning can predict a patient's *length of stay* based on symptoms. Both *length of stay* and *Charlson index score* are mnemonics of the LACE algorithm. Moreover, as the severity and acuity of a patient's symptoms change during their course of stay, predictive modelling can assess a patient's likelihood to be re-admitted based on their progress during hospitalization. As a patient's symptoms change during hospitalization, their likelihood to be re-admitted changes and predictive modelling can give real-time feedback of the likelihood of re-admission. Care plans can be adjusted based off of this feedback.

Furthermore, the LACE algorithm does not acknowledge other key risk factors for re-admission including hospital injuries, place of discharge, and demographical attributes, and it does not necessarily classify special patients, such as surgical or obstetric patients. Machine learning has the capability to use information from all sources to draw conclusions, even if the information pertains to predictors that are indirectly related to admissions. Researchers are already studying ways to use machine learning predict adverse drug events (Gurulingappa, Mateen-Rajpu, & Toldo, 2012; Liu & Chen, 2013), which would also be useful in predicting hospital re-admissions (Hasan, et al., 2010). Another example of an indirect benefit of machine learning would be to use government data on income levels by postal code and make predictions on a patient's income level based on their address.

Machine learning has the capability to create iterative algorithms which can predict which patients are at risk of re-admission and these algorithms can evolve over time. This is beneficial because risk factors can change over time and an effective predictor today may not be effective in a decade.

Other domains of healthcare rely on machine learning to build predictive models and cross-collaboration of findings can be useful for building predictive models. A proposed system, Mobile Machine-Learning Model for Monitoring Cardiovascular Disease (M4CVD), utilizes physiological signals from wearable sensors, combined with clinical data from health records to monitor patients at risk of cardiovascular disease (Boursalie, Samavi, & Doyle, 2015). The M4CVD uses machine learning algorithms to output a binary value for patients of either possessing a continued risk for cardiovascular disease, or not. An experiment found that the M4CVD was able to classify a patient's risk with 90.5% accuracy (Boursalie, Samavi, & Doyle, 2015). Devices such as the M4CVD can also be used by clinicians to understand the correlation between cardiovascular disease and re-admission.

7.3 Data Mining

Data mining can serve as an effective bridge between predictive and prescriptive analytics. It is a set of techniques which extract patterns from large datasets using statistics and machine learning to make predictions and recommendations (McKinsey Global Institute, 2011). Data mining techniques consist of storing classes of information, group them into clusters, identify associations, and anticipate trends through sequential patterns (Anderson School of Management, n.d.). These techniques are used by many retailers and marketing organizations to make predictions about customer usages. Examples would be how Netflix can recommend which movies to watch based on previous viewing habits or how banks can recommend certain financial services based on a customer's financial history and spending habits.

These techniques can be utilized with machine learning and NLP in a clinical setting. Different data classifications could pertain to demographics, clinical information, and hospital administrative data. Clusters of information can be grouped based on certain demographic types or certain morbidities. Associations could be determined based on these clusters, an example would be the income level of men who have been previously admitted to hospital. The data mining system could use sequential patterns to determine the likelihood of re-admission based off of the discoveries from the previous techniques. It could potentially recommend an effective care plan to the discharging practitioner to prevent re-admission.

8. ARCHITECTURAL TECHNIQUES FOR BIG DATA

With data lakes containing large amounts of raw data, there lacks a form of oversight or governance (Gartner, 2014). Governance allows for data to be managed by a body or council, define a set of procedures, and a plan to execute these procedures (Rouse, 2007). Along with proper governance, architectural techniques allow for data to be stored in a safe and efficient manner and allow for governance to execute a strategy for analysis, depending on the needs of their clients.

Examples of architectural techniques include cloud computing, distributed computing, data warehousing, and must include security & privacy strategy. Cloud computing allows for data computations to take place over a distributed network. Distributed computing allows for tasks to be divided among multiple computers so that they are computed in parallel. Data warehousing, which uses the extract, transform, and load technique, allows unstructured data to be stored and

allows for direct reporting and querying. Security & privacy architectures continue to be examined, considering the large volume of data.

8.1 Cloud Computing

Cloud computing involves having a distributed system utilize computing resources to deliver services over a network (McKinsey Global Institute, 2011). Considering that hospital re-admissions could be missed if a patient seeks care at a different hospital, it is important that a regional hospital system is able to share data to be able to coordinate an effective care strategy for the patient. A regional hospital system could utilize a common network to store their data and perform computations as cloud does provide common management of a resource pool (IBM, n.d.). Considering the sensitive nature of health data, it is best that this network be hosted and managed internally, with local hospital computers being able to connect to this internal regional network.

On a cloud, computations are done in parallel, with data stored in clusters, due to the large volumes of data which are stored on the cloud (Hashem, et al., 2015). Hadoop is an example of a software which allows for distributed computations across clusters of computers.

8.2 Distributed Computing

Distributed networks allow for computations to take place in parallel. Tasks are divided over multiple computers on the same network and tasks are computed in parallel (McKinsey Global Institute, 2011). This process allows for tasks to be computed efficiently in both time and cost and are more reliable as multiple devices can make up for one device failing (McKinsey Global Institute, 2011). Given that hospitals are large organizations and that predictions ideally should be made in real time, a system which predicts hospital re-admissions should use distributed computing to save on time and cost.

8.3 Data Warehousing & ETL

Data warehouses are specialized databases which store large amounts of structured data for reporting and analysis (McKinsey Global Institute, 2011). Data warehouses allow for data from multiple sources and algorithms to be integrated into one repository for fast analysis and direct querying (Widom, 1995). Data warehouses uses extract, transform, load (ETL) software tools which extract data from outside sources, transform them as needed for the operation, and load them into the database or data warehouse (McKinsey Global Institute, 2011). Considering that information to predict hospital re-admissions would come from multiple sources, data warehousing using ETL tools would serve as an ideal platform.

8.4 Security & Privacy

Given the sensitive nature of health information, privacy and security of information must be taken into consideration when designing an analytical system. There are many organizations researching security architectures for big data given the fact that it is difficult to protect big data due to its large volume and complexity (Cardenas, Manadhata, & Rajan, 2013). The first thing an organization must do is identify each data source in a repository and who has access to this information and then classify it by sensitivity (Tankard, 2012). Controls should be set by the principle of least privilege[3] and access should be monitored and logged (Tankard, 2012). Security Information and Event Management (SIEM) technologies can monitor network traffic. Finally, any information

[3] Allow users to have access to the minimum amount of necessary information. More information at http://searchsecurity.techtarget.com/definition/principle-of-least-privilege-POLP

that is not needed should be disposed of appropriately to minimize the volume of data. Ways to appropriately dispose data is by encryption, tokenization, or data masking, as these techniques make that data unreadable for those without the keys to unlock it (Tankard, 2012).

9. CONCLUSION & RECOMMENDATIONS

As we are heading into the mid-point of the 2010 decade, there will be an increase in the use of big data & analytics across many different industries. The ability to store large volumes of data from a variety of different sources provides businesses with rich data lakes to perform analysis on this data. An analytical system has the ability to **describe the current state, make predictions based on the previous descriptions,** and **prescribe solutions based on these predictions.**

With the increased adaptation of EHRs, healthcare is no exception to this trend. To leverage big data & analytics, an EHR system must:

- Contain a large **volume** of data
- Allow for information to be given in a timely manner, ensuring **velocity**

The reality of EHR systems is that they already:

- Have a **variety** of data, structured and unstructured
- Are **varied** and experience peaks during certain times
- Are **complex**, as information is coming from a variety of different sources

CIHI tracks hospital re-admission rates (within 30 days of initial admission) as an indicator of the quality of care that a patient receives in hospital. While all re-admissions cannot be prevented, it is important to minimize the incidence of re-admission to ensure that patients are able to recover following a hospital admission. If an EHR system can predict which patients are at risk for re-admission to hospital, clinicians can tailor the care specifically to each patient to prevent a re-admission. Risk factors for re-admission can be based on demographical factors such as gender, age, race, marital status, and income level or on clinical factors such as disease progression, chronic illness or co-morbidities, types of morbidities, length of hospital stay, and the number of previous admissions to hospital.

The LACE Index Score is a moderately accurate way of predicting hospital re-admissions. There are limitations, however, in regards to the fact that the LACE Index score only can make predictions after a discharge, which does not necessarily impact the quality of care during the index admission. The LACE Index score also neglects other important risk factors such as socioeconomic and demographics. Moreover, important information such as symptoms and morbidities, are not necessarily stored in a structured format, which would be required in the LACE algorithm.

A big data system, which utilizes a data warehouse and also utilizes NLP & machine learning can make predictive models based on information entered into an EHR to predict which patients are at risk for re-admission. Data mining can also assist clinicians with creating an effective care plan upon discharge, to mitigate the risk of re-admission. Ideally, such a system should be hosted on a private, secure, internal cloud, managed by a regional health system with computations done in parallel.

Finally, it should be noted that health data analysts play an important role in developing predictive models through EHR data. A significant amount of effort is invested towards structuring raw data for analysis. Effective data governance allows for efficient analytical procedures to suit the needs of clients and also allows data to be stored to protect its safety & privacy. Big data has strong potential to be used in healthcare settings, and it is the way to move forward in predicting hospital re-admissions.

APPENDIX I

Reproduced from:
Niewiadomski, E. (n.d.). *How to calculate the LACE risk score*. Retrieved April 25, 2016, from Besler Consulting: http://www.besler.com/lace-risk-score/

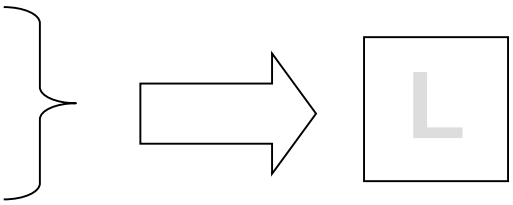
MR#_____
UNIT_____
DOS_____

LACE Index Scoring Tool for Risk Assessment of Hospital Readmission

Step 1. Length of Stay

Length of stay (including day of admission and discharge): _____ days

Length of stay (days)	Score (circle as appropriate)
1	1
2	2
3	3
4-6	4
7-13	5
14 or more	7



Step 2. Acuity of Admission

Was the patient admitted to hospital via the emergency department?
If yes, enter "3" in Box A, otherwise enter "0" in Box A

Step 3. Comorbidities

Condition (definitions and notes on reverse)	Score (circle as appropriate)	
Previous myocardial infarction	+1	If the TOTAL score is between 0 and 3 enter the score into Box C. If the score is 4 or higher, enter 5 into Box C
Cerebrovascular disease	+1	
Peripheral vascular disease	+1	
Diabetes without complications	+1	
Congestive heart failure	+2	
Diabetes with end organ damage	+2	
Chronic pulmonary disease	+2	
Mild liver or renal disease	+2	
Any tumour (including lymphoma or leukemia)	+2	
Dementia	+3	
Connective tissue disease	+3	
AIDS	+4	
Moderate or severe liver or renal disease	+4	
Metastatic solid tumour	+6	
TOTAL		

C

Step 4. Emergency department visits

How many times has the patient visited an emergency department in the six months prior to admission (not including the emergency department visit immediately preceding the current admission)? _____

Enter this number or 4 (whichever is smaller) in Box E

E

Add numbers in Box L, Box A, Box C, and Box E to generate LACE score and enter into box below.

LACE

LACE Score Risk of Readmission: ≥ 10 High Risk

Condition	Definition and/or notes
Previous myocardial infarction	Any previous definite or probable myocardial infarction
Cerebrovascular disease	Any previous stroke or transient ischemic attack (TIA)

Peripheral vascular disease	Intermittent claudication, previous surgery or stenting, gangrene or acute ischemia, untreated abdominal or thoracic aortic aneurysm
Diabetes without microvascular complications	No retinopathy, nephropathy or neuropathy
Congestive heart failure	Any patient with symptomatic CHF whose symptoms have responded to appropriate medications
Diabetes with end organ damage	Diabetes with retinopathy, nephropathy or neuropathy
Chronic pulmonary disease	??
Mild liver or renal disease	Cirrhosis but no portal hypertension (i.e., no varices, no ascites) OR chronic hepatitis Chronic Renal Disease
Any tumour (including lymphoma or leukemia)	Solid tumours must have been treated within the last 5 years; includes chronic lymphocytic leukemia (CLL) and polycythemia vera (PV)_
Dementia	Any cognitive deficit??
Connective tissue disease	Systemic lupus erythematosus (SLE), polymyositis, mixed connective tissue disease, moderate to severe rheumatoid arthritis, and polymyalgia rheumatica
AIDS	AIDS-defining opportunistic infection or CD4 < 200
Moderate or severe liver or renal disease	Cirrhosis with portal hypertension (e.g., ascites or variceal bleeding) End stage Renal Disease, Hemodialysis or Peritoneal Dialysis
Metastatic solid tumour	Any metastatic tumour

APPENDIX II – CALCULATION OF THE PARR-30 SCORE

Variables to calculate re-admission risk as per the PARR-30 index:

[Table below reproduced from:
Billings, J., Blunt, I., Stevenson, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *British Medical Journal Open*.]

Table 1 Summary of variables* included in model, and their coefficients, SE and significance

Variable	Coefficient	SE	Significance
Patient age (squared)	6e-5	0	<0.001
Number of emergency hospital discharges in the last year	0.121	0.002	<0.001
Whether there had been a prior emergency hospital discharge in the past 30 days	0.526	0.012	<0.001
Whether the current admission was an emergency admission	0.556	0.011	<0.001
Index of multiple deprivation band for the place of residence (lower super output area)	0.021–0.102	0.013–0.018	<0.001–0.142
History in the prior 2 years (from any hospital episode statistics primary or secondary diagnostic field) of 11 major health conditions drawn from the Charlson comorbidity index			
Congestive heart failure	0.095	0.018	<0.001
Peripheral vascular disease	0.104	0.022	<0.001
Chronic pulmonary disease	0.224	0.012	<0.001
Diabetes with chronic complications	0.146	0.032	<0.001
Renal disease	0.198	0.018	<0.001
Metastatic cancer with solid tumour	0.276	0.024	<0.001
Other malignant cancer	0.507	0.015	<0.001
Moderate/severe liver disease	0.267	0.049	<0.001
Other liver disease	0.213	0.031	<0.001
Haemiplegia or paraplegia	0.106	0.033	0.001
Dementia	0.047	0.026	0.071
Hospital-specific variable (range of values in appendix 1)	−0.976 to 0.308	0.043 to 0.206	<0.001 to 0.966
Constant	−2.918	0.032	0

*Full details of the model and definitions are available from http://www.nuffieldtrust.org.uk.

Each variable is divided into three sections:

[Information presented below has been reproduced from:
Nuffield Trust. (n.d.). *How to implement the PARR-30 model: required data and algorithm*. Retrieved April 25, 2016, from Nuffield Trust: http://www.nuffieldtrust.org.uk/how-implement-parr-30-model-required-data-and-algorithm]

I. **General information about the patient:**
   - The NHS organization code of the hospital in which they are being treated;
   - Their age;
   - Their postcode (which is used to assign a deprivation level to the patient).

II. **Information about the patient's history of emergency admissions:**
   - Number of (non-obstetrics and gynaecology) emergency admissions the patient had in the last year;
   - Whether or not the patient had any (non-obstetrics and gynaecology) emergency admissions in the last 30 days;
   - Is the current admission an emergency admission?

III. **Information about the patient medical history:**
Does the patient have a history of each of the following conditions (yes/no)?
- Congestive heart failure
- Peripheral vascular disease
- Dementia
- Chronic pulmonary disease
- Other liver disease
- Other malignant cancer
- Metastatic cancer with solid tumour
- Moderate/severe liver disease
- Diabetes with chronic complications
- Hemiplegia or paraplegia
- Renal disease

For each of the 14 items of information from sections II and III, multiply them by a model coefficient. The questions that have yes/no answers are applied as yes = 1 and no =0. Coefficients for all items in the model are published in the appendix to our *BMJ Open* paper.
The three items of general information about the patient (section 1) are handled a little differently and need a bit of processing first:

- The NHS organization code is used to look up a coefficient for that trust (each organization has a different one).
- The patient's age needs to be squared before it is fed into the model and multiplied by its coefficient.
- Lastly, the patient's postcode is used to look up the deprivation level of their local area. This is done by mapping from the patient's postcode to lower super output area (LSOA), then looking up that LSOA's deprivation score in the2007 Index of Multiple Deprivation (IMD).

The PARR-30 model uses six bands of deprivation, and the band in which the LSOA deprivation falls will point to the deprivation coefficient to be used. Of course, if the patient's LSOA is already known there is no need to gather the postcode. The Office for National Statistics has a search tool that can help.
When all the information has been multiplied by the respective coefficients, the results are summed and a constant term added. This final result can be converted to a percentage using:

exp(risk_score)/(1+exp(risk_score))

Worked Example
A fictional 83-year-old woman from a relatively deprived part of London is about to be discharged from a large teaching hospital in London. Her home post code is E1 5AA. She received an emergency admission linked to her chronic obstructive pulmonary disease 7 days ago.

Though she has not been in hospital within the last month, she did have two discharges following emergency admissions in the previous year. The patient also has a history of congestive heart failure and peripheral vascular disease.

Two things have to be looked up: the NHS code for Barts and The London NHS Trust is 'RNJ', and the associated coefficient is 0.1171. The patient's LSOA is Tower Hamlets 013C, and the 2007 IMD score was 34.84. This places it in our 25 to 40 deprivation band, for which the coefficient is 0.023915.

The patient's risk score is then:

Information	Value	Coefficient	Value * Coefficient
Hospital code	n/a	0.1171	0.117
Squared code	6889	0.000060581	0.417
Deprivation	n/a	0.066135	0.066
Number of admissions last year	2	0.12145	0.243
Admission in last month	0	0.5258	0
Current admission is emergency/un-planned	1	0.55648	0.556
Congestive heart failure	1	0.09498	0.095
Chronic pulmonary disease	1	0.22433	0.224
Peripheral vascular disease	1	0.10425	0.104
Constant	-2.91821	1	-2.918
		Total	**-1.096**
		Probability	**25.1%**

REFERENCES

1. Allaudeen, N., Schnipper, J. L., Orav, E. J., Wachter, R. M., & Vidyarthi, A. M. (2011). Inability of Providers to Predict Unplanned Readmissions. *Journal of General Internal Medicine, 26*(7), 771 - 776.
2. Allaudeen, N., Vidyarthi, A., Maselli, J., & Auerbach, A. (2011). Redefining Readmission Risk Factors for General Medicine Patients. *Journal of Hospital Medicine, 6*(2), 54 - 60.
3. Anderson School of Management. (n.d.). *Data Mining: What is Data Mining?* . Retrieved from University of California Los Angeles: http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
4. Arbaje, A. I., Wolffe, J. L., Yu, Q., Powe, N. R., Andersen, G. F., & Boult, C. (2008). Postdischarge Environmental and Socioeconomic Factors and the Likelihood of Early Hospital Readmission Among Community-Dwelling Medicare Beneficiaries. *The Gerontologist, 48*(4), 495 - 504.
5. Ashton, C. M., & Wray, N. P. (1996). A Conceptual Framework for the Study of Readmission as an Indicator of Quality of Care. *Social Science & Medicine, 43*(11), 1533 - 1541.
6. Baillie, C. A., Van Zanbergen, C., Tait, G., Hanish, A., Leas, B., French, B., . . . Unscheid, C. A. (2014). The Readmission Risk Flag: Using the Electronic Health Record toAutomatically Identify Patients at Risk for 30-Day Readmission. *Journal of Hospital Medicine, 8*(12), 689 - 695.
7. Bardsley, M., Georghiou, T., Billings, J., & Blunt, I. (2012, August 10). *Predicting risk of hospital readmission with PARR-30*. Retrieved April 25, 2016, from Nuffield Trust: http://www.nuffieldtrust.org.uk/our-work/projects/predicting-risk-hospital-readmission-parr-30
8. Batra, S. (2014). Big Data Analytics and its Reflections on DIKW Hierarchy. *Review of Management, 4*(1/2), 5 - 17.
9. Benbassat, J., & Taragin, M. (2000). Hospital Readmissions as a Measure of Quality of Health Care. *Archives of Internal Medicine, 160*(8), 1074 - 1081.
10. Billings, J., Blunt, I., Stevenson, A., Georghiou, T., Lewis, G., & Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *British Medical Journal Open*.
11. Boursalie, O., Samavi, R., & Doyle, T. E. (2015). M4CVD: Mobile Machine Learning Model for Monitoring Cardiovascular Disease. *Procedia Computer Science, 63*, 384 - 391.
12. Canadian Institute for Health Information. (2016). *All Patients Readmitted to Hospital*. Retrieved March 26, 2016, from Your Health System: http://yourhealthsystem.cihi.ca/inbrief/?lang=en#!/indicators/006/all-patients-readmitted-to-hospital/;mapC1;mapLevel2;/

13. Canadian Institute for Health Information. (2016). *Indicators*. Retrieved March 16, 2016, from Canadian Institute for Health Information: https://www.cihi.ca/en/health-system-performance/performance-reporting/indicators
14. Cardenas, A. A., Manadhata, P. K., & Rajan, S. P. (2013). Big Data Analytics for Security. *IEEE Computer and Reliability Societies*.
15. Centres for Medicare & Medicaid Services. (2012, March 26). *Electronic Health Records.* Retrieved March 8, 2016, from Centres for Medicare & Medicaid Services: https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html?redirect=/ehealthrecords/
16. Charlson, M. E., Pompei, P., Ales, K. L., & Mackenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of chronic diseases, 40*(5), 373-383.
17. Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impac. *MIS Quarterly, 36*(4), 1165 - 1188.
18. Cortada, J. W., Gordon, D., & Lenihan, B. (2012). *The value of analytics in healthcare.* Somers, NY: IBM Institute for Business Value.
19. Feudtner, C., Levin, J. E., Srivastava, R., Goodman, D. M., Slinom, A. D., Sharma, V., . . . Hall, M. (2009). How well can hospital readmission be predicted in a cohort of hospitalized children? A retrospective, multicenter study. *Pediatrics, 123*(1), 286 - 293.
20. Friedman, B., & Basu, J. (2004). The Rate and Cost of Hospital Readmissions for Preventable Conditions. *Medical Care Research & Review, 61*(2), 225- 240.
21. Gartner. (2014, July 28). *Gartner Says Beware of the Data Lake Fallacy*. Retrieved April 25, 2016, from Gartner: http://www.gartner.com/newsroom/id/2809117
22. Gruneir, A., Dhalla, I. A., Van Walraven, C., Fischer, H. D., Camacho, X., Rochon, P. A., & Anderson, G. M. (2011). Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. *Open Medicine, 5*(2), 104 -111.
23. Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics, 3*(15).
24. Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, C. M., Kaboli, P. J., Auerbach, A. D., . . . Schnipper, J. L. (2010). Hospital Readmission in General Medicine Patients: A Prediction Model. *Journal of General Internal Medicine, 25*(3), 211 - 219.
25. Hashem, I. A., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98 - 115.
26. IBM. (n.d.). *What is cloud computing?* Retrieved from IBM Cloud: http://www.ibm.com/cloud-computing/what-is-cloud-computing.html
27. Jencks, S. M., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among Patients in the Medicare Fee-for-Service Program. *New England Journal of Medicine, 360*(14), 1418 - 1428.
28. Kassin, M. T., Owen, R. M., Perez, S. D., Leeds, I., Cox, J. C., Schnier, K., . . . Sweeney, J. F. (2012). Risk Factors for 30-Day Hospital Readmission among General Surgery Patients. *Journal of the American College of Surgeons, 215*(3), 322 - 330.

29. Kiran, R. P., Delaney, C. P., Senagore, A. J., Steel, M., Garafalo, T., & Fazio, V. W. (2004). Outcomes and Prediction of Hospital Readmission after Intestinal Surgery. *Journal of the American College of Surgeons, 198*(6), 877 - 883.
30. Kruse, S., Papotti, P., & Naumann, F. (2015). Estimating Data Integration and Cleaning Effort. *18th International Conference on Extending Database Technology.* Brussels, Belgium.
31. Lagoe, R. A., Noetscher, C. M., & Murphy, M. P. (2001). Hospital Readmission: Predicting the Risk. *Journal of Nursing Care Quality, 15*(4), 69 - 83.
32. Liu, S., Heaman, M., Joseph, K. S., Liston, R. M., Huang, L., Sauve, R., & Kramer, M. S. (2005). Risk of Maternal Postpartum Readmission Associated with Mode of Delivery. *Obstetrics & Gynecology, 105*(4), 836 - 842.
33. Liu, S., Heaman, M., Kramer, M. S., Demissie, K., Wen, S. W., & Marcoux, S. (2002). Length of hospital stay, obstetric conditions at childbirth, and maternal readmission: a population-based cohort study. *American journal of obstetrics and gynecology, 187*(3), 681 - 687.
34. Liu, X., & Chen, H. (2013). AZDrugMiner: An information extraction system for mining patient-reported adverse drug events in online patient forums. *Smart Health*, 134 - 150.
35. McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity.*
36. Middleton, J., Lakhanpal, R., Price, S., & Butler, D. (2015, November 16). *Improving the Efficacy of Predictive Models.* Retrieved March 16, 2016, from HIMSS Clinical & Business Intelligence: http://s3.amazonaws.com/rdcms-himss/files/production/public/FileDownloads/2015-11-10_Article_Improving%20the%20Efficacy%20of%20Predictive%20Models_FINAL.pdf
37. Niewiadomski, E. (n.d.). *How to calculate the LACE risk score*. Retrieved April 25, 2016, from Besler Consulting: http://www.besler.com/lace-risk-score/
38. Nuffield Trust. (n.d.). *How to implement the PARR-30 model: required data and algorithm*. Retrieved April 25, 2016, from Nuffield Trust: http://www.nuffieldtrust.org.uk/how-implement-parr-30-model-required-data-and-algorithm
39. Raghupathi, W., & Raghupathi, V. (2013). An overview of health analytics. *Journal of Health & Medical Informatics, 4*(132).
40. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2*(1), 3.
41. Rouse, M. (2007, July). *Data Governance*. Retrieved April 25, 2016, from Tech Target: http://searchdatamanagement.techtarget.com/definition/data-governance
42. Rumball-Smith, J., & Hider, P. (2009, February). The validity of readmission rate as a marker of the quality of hospital care, and a recommendation for its definition. *The New Zealand Medical Journal, 122*(1289), 63 - 70.
43. SAS. (n.d.). *Big Data: What it is and Why it Matters.* Retrieved March 8, 2016, from SAS - The Power to Know: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

44. SAS. (n.d.). *Machine Learning: What it is & why it matters?* Retrieved March 16, 2016, from SAS: http://www.sas.com/en_us/insights/analytics/machine-learning.html
45. Tankard, C. (2012). Big data security. *Network Security*.
46. Tutorials Point. (2016). *AI - Natural Language Processing.* Retrieved March 16, 2016, from Tutorials Point: http://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm
47. van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., . . . Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal, 182*(6), 551 - 557.
48. van Walraven, C., Wong, J., Forster, A. J., & Hawken, S. (2012). Predicting post-discharge death or readmission: deterioration of model performance in population having multiple admissions per patient. *Journal for Evaluation in Clinical Practice, 19*(6), 1012 - 1018.
49. Widom, J. (1995). Research Problems in Data Warehousing. *Proceedings of the fourth international conference on Information and knowledge management* (pp. 25 - 30). Association for Computer Machinery.
50. Wolniewicz, R. (n.d.). *Computer-assisted coding and natural language processing.* Retrieved March 16, 2016, from 3M Health Information Systems: http://s3.amazonaws.com/rdcms-himss/files/production/public/HIMSSorg/Content/files/HIS_Comp_Assist-coding_and_NLP_Whitepaper-3M_May%202015.pdf

McMaster University
1280 Main St. W. DSB A202
Hamilton, ON
L8S 4M4

Tel: 905-525-9140 ext. 23956
Fax: 905-521-8995
Email: ebusiness@mcmaster.ca
Web: http://merc.mcmaster.ca