Automating Construction of Machine Learning Models with Clinical Big Data: Proposal Rationale and Methods

Gang Luo[1], PhD; **Bryan L Stone**[2], MD, MS; **Michael D Johnson**[2], MD; **Peter Tarczy-Hornoch**[1, 3, 4], MD; **Adam B Wilcox**[1], PhD; **Sean D Mooney**[1], PhD; **Xiaoming Sheng**[5], PhD; **Peter J Haug**[6, 7], MD; **Flory L Nkoy**[2], MD, MS, MPH
[1]Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98109, USA
[2]Department of Pediatrics, University of Utah, 100 N Mario Capecchi Drive, Salt Lake City, UT 84113, USA
[3]Department of Pediatrics, Division of Neonatology, University of Washington, 1959 NE Pacific St, Seattle, WA 98195, USA
[4]Department of Computer Science and Engineering, University of Washington, 185 Stevens Way, Seattle, WA 98195, USA
[5]Department of Pediatrics, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84108, USA
[6]Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA
[7]Homer Warner Research Center, Intermountain Healthcare, 5121 South Cottonwood Street, Murray, UT 84107, USA
luogang@uw.edu, bryan.stone@hsc.utah.edu, mike.johnson@hsc.utah.edu, pth@uw.edu, abwilcox@uw.edu, sdmooney@uw.edu, xiaoming.sheng@utah.edu, peter.haug@imail.org, flory.nkoy@hsc.utah.edu

Corresponding author:
Gang Luo, PhD
Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98109, USA
Phone: 1-206-221-4596
Fax: 1-206-221-2671
Email: luogang@uw.edu

Abstract
Background: To improve health outcomes and cut healthcare costs, we often need to conduct prediction/classification using large clinical data sets, a.k.a. "clinical big data," e.g., to identify high-risk patients for preventive interventions. Machine learning has been proposed as a key technology for doing this. Machine learning won most data science competitions and could support many clinical activities, yet only 15% of hospitals use it for even limited purposes. Despite familiarity with data, healthcare researchers often lack machine learning expertise to directly use clinical big data, creating a hurdle in realizing value from their data. Healthcare researchers can work with data scientists with deep machine learning knowledge, but it takes time and effort for both parties to communicate effectively. Facing a U.S. shortage of data scientists and hiring competition from companies with deep pockets, healthcare systems have difficulty recruiting data scientists. Building and generalizing a machine learning model often requires hundreds to thousands of manual iterations by data scientists to select: a) hyper-parameter values and complex algorithms that greatly affect model accuracy, as well as b) operators and periods for temporally aggregating clinical attributes (e.g., whether a patient's weight kept rising in the past year). This process becomes infeasible with limited budgets.
Objective: This study's goal is to enable healthcare researchers to directly use clinical big data, make machine learning feasible with limited budgets and data scientist resources, and realize value from data.
Methods: This study will: 1) finish developing new software Auto-ML (Automated Machine Learning) to automate model selection for machine learning with clinical big data and validate Auto-ML on seven benchmark modeling problems of clinical importance, 2) apply Auto-ML and novel methodology to two new modeling problems crucial for care management allocation and pilot one model with care managers, and 3) perform simulations to estimate the impact of adopting Auto-ML on U.S. patient outcomes.
Results: We are currently writing Auto-ML's design document. We intend to finish our study in around five years.
Conclusions: Auto-ML will generalize to various clinical prediction/classification problems. With minimal help from data scientists, healthcare researchers can use Auto-ML to quickly build high-quality models. This will boost wider use of machine learning in healthcare and improve patient outcomes.

Keywords: Machine learning; automated temporal aggregation; automatic model selection; care management; clinical big data

1. Introduction
Barriers in using machine learning to realize value from clinical big data
To improve health outcomes and trim healthcare costs, we often need to perform prediction/classification using large clinical data sets, a.k.a. "clinical big data," e.g., to identify high-risk patients for preventive interventions. Machine learning has been

proposed as a key technology for doing this. Machine learning studies computer algorithms that learn from data, such as support vector machine, random forest, neural network, and decision tree [1]. Trials showed machine learning was used to help: a) lower 30-day mortality rate (odds ratio=0.53) in emergency department (ED) patients having community-acquired pneumonia [2], b) increase on-target hemoglobin values by 8.5-17% and reduce cardiovascular events by 15%, hospitalization days by 15%, blood transfusion events by 40-60%, expensive darbepoetin consumption by 25%, and hemoglobin fluctuation by 13% in end-stage renal disease patients on dialysis [3-6], c) reduce ventilator use by 5.2 days and healthcare cost by $1,500 per patient at a hospital respiratory care center [7], and d) lower healthcare cost in Medicare patients' last six months of life by 4.5% [8].

Machine learning could support many clinical activities, but only 15% of hospitals use it for even limited purposes [9]. Compared to statistical methods like logistic regression, machine learning poses less strict assumptions on distribution of data, can increase prediction/classification accuracy, in certain cases doubling it [10-12], and won most data science competitions [13]. Historically, machine learning was blamed for being a black box. A recent method can automatically explain any machine learning model's classification results with no accuracy loss [14, 15]. Yet, two hurdles remain in using machine learning in healthcare. First, despite familiarity with data, healthcare researchers often lack machine learning expertise to directly use clinical big data. Data scientists take years of training to gain deep machine learning knowledge. Healthcare researchers can work with them, but it takes time and effort for both parties to communicate effectively. Facing a U.S. shortage of data scientists estimated as high as 140,000+ by 2018 [16] and hiring competition from companies with deep pockets, healthcare systems have a hard time recruiting data scientists [17, 18]. As detailed below, developing a machine learning model often requires data scientists to spend extensive time on model selection, which becomes infeasible with limited budgets. Second, some healthcare systems such as Kaiser Permanente, Intermountain Healthcare (IH), University of Washington Medicine (UWM), Columbia University Medical Center, Veterans Health Administration, and University of Utah Health have teams devoted to data cleaning. Healthcare researchers can obtain cleaned data from these systems' enterprise data warehouses (EDWs). In other healthcare systems, one needs to laboriously clean data before machine learning. This is often done with the help of database programmers and/or master-level statisticians, who can also help with data pre-processing and are easier to find than data scientists with deep machine learning knowledge. This study addresses the first hurdle and focuses on automating machine learning model selection and temporal aggregation, an important type of data pre-processing.

Barrier 1: Data scientists are needed for choosing hyper-parameter values and algorithms

Each learning algorithm includes two categories of parameters: hyper-parameters a machine learning tool user manually sets prior to model training, and normal parameters automatically tuned in training the model (Table 1). Given a modeling problem such as predicting 30-day hospital readmission, an analyst manually constructs a model as follows. First, select an algorithm from many pertinent ones like the ~40 algorithms for classification included in Weka [19]. Second, set the values of the selected algorithm's hyper-parameters. Third, train the model to tune the normal parameters of the selected algorithm automatically. In case model accuracy is unsatisfactory, substitute the algorithm and/or hyper-parameter values and then re-train the model, while using some technique to avoid overfitting on the validation set [20-24]. This process is done over and over until the analyst runs out of time, has a model with good accuracy, or cannot improve further. If feature selection is considered, in each iteration the user also needs to choose a feature selection technique from many applicable ones and set its hyper-parameter values, making this process even more complex. Many possible combinations of hyper-parameter values and learning algorithms lead to hundreds to thousands of laborious and manual iterations to construct a model. These iterations need machine learning expertise, are typically done by a data scientist, and become a barrier [25].

Table 1. Two learning algorithms and their example normal parameters and hyper-parameters.

Learning algorithm	Example hyper-parameters	Example normal parameters
Support vector machine	regularization constant C, kernel to use, tolerance parameter, ε for round-off error, a polynomial kernel's degree	support vectors and their Lagrange multipliers
Random forest	number of independent variables to examine at each inner node of a classification and regression tree, number of trees	threshold value and input variable used at each inner node of a tree

Model accuracy is affected by choice of hyper-parameter values and learning algorithm. Thornton *et al*. [25] demonstrated for the 39 classification algorithms included in Weka, the impact on model accuracy averages 46% and can be up to 94%. Even considering a few popular algorithms like random forest and support vector machine, the impact is still above 20% on 2/3 of 21 benchmark data sets. The good choice changes by the particular modeling problem. Computer science researchers have investigated methods for automatically searching hyper-parameter values and algorithms [26]. Some methods can reach equal or better results compared to data scientists' manual tuning [27, 28]. But in case a large number of algorithms are examined, efforts like Auto-WEKA [25, 29-31], hyperopt-sklearn [28], and MLbase [32, 33] cannot effectively handle large data sets in reasonable time.

A hurdle to automatic search is the amount of time needed to assess on an entire data set a combination of hyper-parameter values and a learning algorithm. On a modern computer, it takes two days to train the champion ensemble model that won the Practice Fusion diabetes classification contest [34] one time on 9,948 patients with 133 input or independent variables, a.k.a. features. Even when disregarding ensembles of >5 base models, aborting long-running tests, and greatly limiting the hyper-parameter value search space (e.g., allowing ≤256 decision trees in a random forest) all impacting search result quality, >30 minutes are needed to test an average combination on 12,000 rows (data instances) with 784 attributes [35]. To ensure search result quality, automation efforts often test >1,000 combinations on the whole data set [35], leading to months of search time. On a data set with several dozen attributes and several thousand rows, a search can still take several days [25]. In reality, search time could be thousands of times longer even with a computer cluster for five reasons: 1) Model building is iterative. When a collection of clinical attributes yields low model accuracy, the analyst can include other attributes to boost accuracy. Every iteration takes a new search for hyper-parameter values and learning algorithms. 2) Frequently, ensembles of a large number of base models reach higher accuracy. The training time of an ensemble model rises proportionally to the number of base models. 3) Hyper-parameter values over a broad range are often used to achieve higher accuracy. The above champion ensemble model [34] uses 12 base models. Each random forest base model uses ≥15,000 decision trees. 4) Numerous rows, often from multiple healthcare systems, can reside in a data set. 5) Numerous attributes, e.g., derived from genomic or textual data, can exist in a data set. In a hospital without genomic data, a model for readmission prediction was built using 195,901 patients and 3,956 attributes already [36]. An algorithm's execution time rises proportionally to the number of attributes at a minimum and often superlinearly with the number of rows. Irrespective of whether search is done manually or automatically, a slow speed in search frequently causes a search to be terminated early, producing suboptimal model accuracy [35].

Barrier 2: Data scientists are needed for temporally aggregating clinical attributes

Numerous clinical attributes are documented over time needing aggregation prior to machine learning, e.g., weight at each patient visit is combined to check whether a patient's weight kept rising in the previous year. An aggregation period and operator pair (e.g., increasing trend, average, count, and maximum) needs to be specified for every attribute separately to compute an aggregate value. Usually, clinicians designate pairs and data scientists perform computation. Numerous pairs could be clinically meaningful. The ones that produce high accuracy change by the particular modeling problem and are usually not known in advance. Granted a modeling problem, the analyst picks one or more pairs for each attribute manually, then constructs a model. In case model accuracy is unsatisfactory, the analyst substitutes pairs for some attributes and re-constructs the model, while using some technique to avoid overfitting on the validation set [20-24]. This process between data scientists and clinicians is frequently repeated many times and becomes a barrier. No comprehensive aggregation operator list exists, demanding care to not omit effective operators.

Barrier 3: Data scientists are needed for generalizing models

A model built and accurate in a healthcare system often performs poorly and needs to be rebuilt for another system [37], with differing patients, practice patterns, and collected attributes impacting model selection [38, 39]. This needs data scientists and is a barrier, as a system often needs many models for diverse clinical activities.

As often quoted, McKinsey estimates proper use of clinical big data can bring up to >$300 billion in value to U.S. healthcare each year [16]. The achievable value is surely less, but still significant. To realize value from data, we need new approaches to enable healthcare researchers to directly use clinical big data, and make machine learning feasible with limited budgets and data scientist resources.

Our proposed software

To fill the gap, we will 1) finish developing open source software Auto-ML (Automated Machine Learning) to efficiently automate model selection for machine learning with clinical big data and validate Auto-ML on seven benchmark modeling problems of clinical importance, 2) apply Auto-ML and novel methodology to two new modeling problems crucial for care management allocation and pilot one model with care managers, and 3) perform simulations to estimate the impact of adopting Auto-ML on U.S. patient outcomes. We hypothesize that adopting Auto-ML will improve outcomes. Conceptually, Auto-ML will be an automated version of Weka [19] supporting automated temporal aggregation. With minimal help from data scientists, healthcare researchers can use Auto-ML to quickly build high-quality models. This expands the human resource pool for clinical machine learning and aligns with the industry trend of citizen data scientists, where an organization arms its talent with tools to do deep analytics [40]. Auto-ML can greatly reduce the time and cost required of scarce data scientists, busy clinicians, and computing resources in developing models, enable fast turnaround, and facilitate green computing. The faster a high-quality model gets built and deployed, the earlier it can bring outcome improvement. Auto-ML is not used to reach the maximum possible model accuracy in theory, which is hard to do in reasonable time. Instead, Auto-ML is used to quickly build high-quality models. If needed, data scientists and healthcare researchers can manually fine-tune them further.

Auto-ML will efficiently automate selection of feature selection techniques, hyper-parameter values, learning algorithms, and temporal aggregation operators and periods. Auto-ML will continuously show, as a function of time given for model selection, forecasted model accuracy as well as expected patient outcomes of model use. If trends are not promising, the user can abort, add more clinical attributes, and restart. Auto-ML is able to operate on a cluster of computers for scalable processing.

Gaps in patient identification for care management and our proposed solutions

To improve patient identification and outcomes for care management, Aim 2 will apply Auto-ML to two new modeling problems: 1) use a healthcare system's incomplete medical (clinical and/or administrative) data to find future high-cost, diabetic patients, and 2) use vast attributes in modern electronic medical records to find future hospital users in asthmatic patients.

Widely used for chronic diseases like asthma and diabetes, care management applies early interventions to high-risk patients to avoid high costs and health status decline [41-43]. In the U.S., 7.1 million children (9.6%) and 18.7 million adults (8%) [44] have asthma [45, 46]. Every year, asthma causes 1.8 million ED visits, 439,000 hospitalizations, $56 billion in cost [47], and 3,630 deaths [44]. Proper use of care management can cut down asthma exacerbations, trim cost by up to 15%, drop ED visits and hospital (re)admissions by up to 40%, and enhance quality of life, treatment adherence, and patient satisfaction by 30-60% [42, 48-54]. This impacts 63% of annual total asthma costs from asthma exacerbations [51, 55].

For care management to be effective within resource constraints, we should only enroll patients with the worst prognosis or anticipated to have the highest costs. Predictive modeling is widely used for care management [56] as the best method for finding high-risk patients [57], but current approaches have two gaps.

Scope gap: Often, a healthcare system has incomplete medical data on many of its patients, as a patient's complete data may spread across several healthcare systems [58, 59]. Typical models for predicting a patient's cost assume complete data [60-62]. A system usually does not apply models to patients on whom it possibly has incomplete data. As future high-cost patients are not found, care management is not used on them. This limits care management's scope of use to improve outcomes. UWM is seeking a way to fill the gap, notably for patients with diabetes. To do this, we will use a constraint to find patients who tend to get most of their care at UWM, use UWM's incomplete data to build a model, and apply it to them to facilitate care management.

Accuracy gap: Existing models for predicting hospital use (inpatient stay or ED visit) in asthmatic patients have low accuracy [63-68]. A typical model [65] missed 75% of future hospital users. 78% of patients in the high-risk group chosen by the model did not use hospitals in the next year. Two factors degrade accuracy. First, several dozen risk factors for hospital use in asthma are known, including age, gender, race/ethnicity, asthma medication use, prior healthcare use, comorbidities (ischemic heart disease, rhinitis, sinusitis, reflux, anxiety-depression, diabetes, cataracts, chronic bronchitis, and chronic obstructive pulmonary disease), allergies, lung function, number of asthma medication prescribers (as a measure of continuity of care), health insurance type, lab test results (total serum immunoglobulin E level, and eosinophil count), body mass index, smoking status, secondhand smoke exposure, the ratio of controller to total asthma medications, frequency of non-asthma visits, number of procedures, number of diagnoses, number of prescription drug claims, and asthma questionnaire results (frequency of asthma symptom occurrence, interference with normal activity, nighttime awakening, reliever use for symptom control, forced expiratory volume in 1 second (FEV1), peak expiratory flow rate, FEV1/FVC (forced vital capacity) ratio, asthma control test score, number of exacerbations last year, controller use, asthma-related acute care, asthma trigger reduction, and asthma medication) [55, 63, 65, 67-73]. Yet, a typical model uses <10 of these risk factors [63-67]. Existing models were built using data from either clinical trials or outdated electronic medical records gathering limited attributes [74]. No published model uses all known risk factors in modern electronic medical records gathering vast attributes [74]. Second, as with many diseases, many attributes predictive of hospital use in asthma are not found yet. If we could enroll 5% more of future hospital users in care management, we could avoid up to 8,780 hospitalizations and 36,000 ED visits for asthma each year. IH is seeking a way to fill the gap. To do this, we will use vast attributes in IH electronic medical records to build a model predicting hospital use in asthma. The attributes will cover many known risk factors for hospital use in asthma and be used to find new predictive factors.

Innovation

Our study is innovative for multiple reasons:

(1) With the new software that will be built in our project, for the first time, healthcare researchers with limited machine learning knowledge can quickly build high-quality machine learning models with minimal help from data scientists. The cost and time required of data scientists and clinicians in doing machine learning are greatly reduced. Also, it becomes possible to widely use machine learning in healthcare to realize value from clinical big data and improve patient outcomes. No existing software can greatly cut the long time required of data scientists in building and generalizing models.

(2) We will direct care management to more patients needing it more precisely than current approaches. For patients on whom it possibly has incomplete medical data, a healthcare system usually does not apply predictive models to find candidates for care management. Existing models for predicting hospital use in asthmatic patients were built mainly using a small set of patients (e.g., <1,000) or attributes (e.g., <10), creating a hurdle in finding many predictive attributes and their

interactions. Many known risk factors' predictive power for hospital use in asthma is unused. In contrast, we will expand the set of diabetic adults for whom predictive models and care management can be used. We will use many asthmatic children and attributes to build new, accurate models for hospital use. The attributes will cover many known risk factors for hospital use in asthma and be used to find new predictive factors. Our approaches to using incomplete data and vast attributes are new, with principles generalizable to many clinical applications.

(3) Our software will: a) automatically choose hyper-parameter values, feature selection techniques, and algorithms for a particular machine learning problem faster than existing methods. b) efficiently and automatically choose operators and periods for temporally aggregating clinical attributes. No such method currently exists. Longitudinal data analysis [75] models the dependent variable. In contrast, our temporal aggregation can use any function of independent variables. c) continuously show, as a function of time given for model selection, estimated patient outcomes of model use and forecasted model accuracy. For the first time, one can obtain feedback continuously throughout automatic model selection. d) enable fast turnaround. There is no such software at present.

(4) We will systematically compile the first list of regularly used operators for temporally aggregating clinical attributes. The list can be reused for future clinical data analysis studies. Using MapReduce [76] for distributed computing, we will provide the first implementation of many aggregation operators not offered by current big data software such as Hadoop [77] and Spark [78].

(5) We will estimate the impact of adopting our automated machine learning software on U.S. patient outcomes in two scenarios. No such estimate has ever been made. Our impact estimation method is new and can be applied to other scenarios and similar software.

In summary, this study is significant by making machine learning feasible with limited budgets and data scientist resources to help realize value from clinical big data and improve patient outcomes. The models that will be built for the two new modeling problems will help improve care management outcomes.

2. Methods

Auto-ML will be built atop current big data software, enabling it to operate on one computer or a cluster. Built atop the Hadoop distributed file system, Spark [78] is a major open source software system supporting MapReduce [76] for distributed computing. MLlib [79] is the accompanying machine learning library in Spark. Spark is able to perform machine learning >100 times quicker than Hadoop [80]. Auto-ML will be built using the Spark package as well as novel techniques to address current software's limitations.

Aim 1: Finish developing Auto-ML to automate model selection for machine learning with clinical big data and validate Auto-ML on seven benchmark modeling problems of clinical importance.

Figure 1 compares Auto-ML's approach of constructing models to the present one. Four steps are carried out sequentially during machine learning: temporally aggregate clinical attributes, choose hyper-parameter values, feature selection techniques, and algorithms, construct models, and assess models. The temporal aggregation step is optional, e.g., when no repeatedly recorded attribute exists. Auto-ML will use Spark as the basis for distributed computing. Auto-ML will be coded in Java so it can use open source software Spark and Weka, all having a Java application programming interface and/or coded in Java. The user will specify the storage location of the data set in Auto-ML's graphical input interface. Auto-ML will then put the data set into Spark prior to analysis.

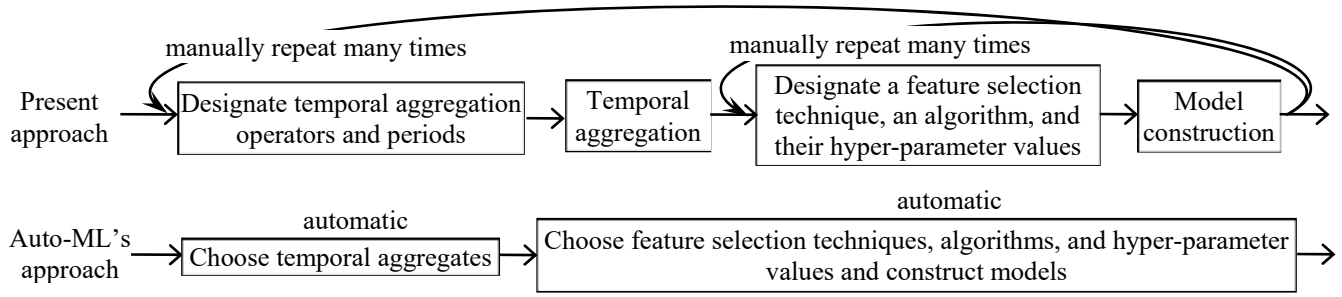

Figure 1. Auto-ML's approach of constructing machine learning models vs. the present one.

Auto-ML's machine learning functions

Auto-ML will integrate MLlib [79] and Weka's [19] machine learning functions by altering source code and/or invoking the Java application programming interfaces. As a broadly used machine learning toolkit, Weka includes many popular feature selection techniques and learning algorithms. distributedWekaSpark [81] is the distributed computing package of Weka for Spark that is able to operate on a computer cluster. MLlib is a distributed machine learning library in Spark implementing some techniques and algorithms supported by Weka. Auto-ML will support all techniques and algorithms available in Weka. Whenever possible, Auto-ML will use MLlib's code, which fuses with Spark better than distributedWekaSpark's code [81].

Weka's [19] graphical user interface covers feature selection (optional), model construction, and model assessment. In the input interface, the Weka user designates the dependent variable, independent variables, data file, learning algorithm, and the hyper-parameter values of the algorithm. After the user clicks the start button, Weka constructs a model and shows its performance measures. For machine learning, Auto-ML's graphical user interface will work similarly with two main differences. First, in Weka, the user must specify an algorithm prior to model building. Like Auto-WEKA [25], Auto-ML will use a hyper-parameter to represent the option of feature selection technique and automatically select the hyper-parameter values, technique, and algorithm. The user may override the choice of Auto-ML. Second, to facilitate the user to track automatic selection's progress, Auto-ML shows a curve presenting the highest accuracy reached over time. The user can terminate the process at any moment and obtain the most accurate model built. In the following, we outline the main techniques that we will use to build Auto-ML.

Aim 1.a: Devise a method to efficiently and automatically choose hyper-parameter values, feature selection techniques, and algorithms.

Our review paper [26] showed that few automatic selection methods [25, 28-31, 82] have been fully implemented, and can manage an arbitrary number of combinations of hyper-parameter values and many learning algorithms. All of these methods are similar to or based on the Auto-WEKA automatic selection approach [25], yet none of them can efficiently handle large data sets. To overcome current methods' inefficiency, we drafted a method based on Bayesian optimization for response surface to rapidly identify, for a specific modeling problem, a good combination of hyper-parameter values, a feature selection technique, and a learning algorithm when a large number of algorithms and techniques are examined [35, 83]. The method represents the option of technique as a special hyper-parameter, proceeds in stages, and conducts progressive sampling [84], filtering, as well as fine-tuning to rapidly shrink the search space. We do fast trials on a small sample taken from the data set to drop unpromising combinations early, reserving resources to fine-tune promising ones. A combination is promising when a model built using it and the sample reaches an error rate below a beginning threshold. Then, we decrease the threshold, enlarge the sample, test and adjust combinations, and cut the search space several times. At the last stage, we find an effective combination using the full data set.

More specifically, at each stage our method uses a training sample and a validation sample. They have no overlap and contain data instances randomly chosen from the data set. We keep the validation sample the same and expand the training sample across stages (Figure 2). At the first stage, we start from a small training sample. For each learning algorithm, we evaluate a fixed number of random hyper-parameter value combinations, if any, as well as its default one. To evaluate a combination, we use it, the training sample, and algorithm to construct a model, then use the validation sample to assess the model's error rate. We identify and remove unpromising algorithms based on the test results. At each subsequent stage that is not the last one, we enlarge the training sample. For each remaining algorithm, we construct a separate regression model, use a Bayesian optimization for response surface approach to choose several new hyper-parameter value combinations, and test them. We identify and remove additional unpromising algorithms based on the test results. At the last stage, we do some final tests on the full data set to come up with the ultimate search result.

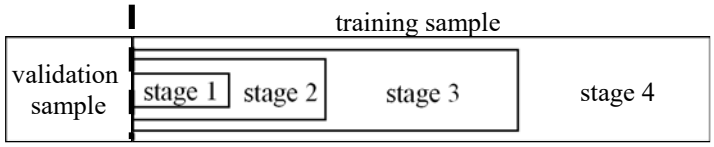

Figure 2. Progressive sampling adopted in our draft automatic model selection method.

Our draft method needs further optimization for efficiency and effectiveness. To do this, we will expand the draft method to include multiple optimization techniques: the seven outlined in our design paper [24] and the six described as follows.

Technique 1: Use two validation samples to help avoid overfitting. At each stage except for the last one, our draft method [35, 83] uses the same validation sample containing a moderate number of data instances to perform many tests. This could lead to overfitting to the validation sample [20-23] that will misguide future search. To help avoid overfitting, we will use two validation samples of equal size with as little overlap as possible, and reduce the frequency of revealing information about the

second validation sample [23]. When the data set has enough data instances, the two validation samples will have no overlap. For a combination of hyper-parameter values and a learning algorithm, we use it and the training sample to construct a model and assess the model's error rate twice, once on either validation sample. Intuitively, the two error rates would be roughly the same in the absence of overfitting. If the error rate on the first validation sample is higher than a specific threshold (e.g., in the top 50% of the error rates on the first validation sample of all combinations tested so far at this stage), we use it as the combination's error rate estimate. Regardless of its exact value, a high error rate estimate will guide future search to avoid the combination's neighborhood. If the threshold is not exceeded, we compare the error rate on the first validation sample with that on the second. If the former is not lower than the latter by a certain threshold (e.g., 5%), we use the former as the combination's error rate estimate. Otherwise, we use the latter as the combination's error rate estimate, as overfitting to the first validation sample is likely to have occurred.

The above approach uses the same two validation samples across different stages. Alternatively, if the data set contains many data instances, we can use a different validation sample at each stage. Each time we arrive at a new stage, we redo sampling to obtain a new validation sample. This also helps avoid overfitting to the same validation sample that is repeatedly used. We will compare the two approaches and choose the one that performs better.

Technique 2: Use multiple feature selection techniques concurrently to drop unpromising features early. Feature selection and model building time rises proportionally to the number of features at a minimum. Doing a test is slow when many features exist in the data set. To tackle this issue, we previously proposed before doing tests, applying a feature selection technique to the data set (or a large sample of it) and rapidly dropping features not likely to have high predictive power [24]. Yet, like the no free lunch theorem [85] shows, no technique can guarantee good performance in all cases. Relying on a single technique can be risky, causing predictive features to be dropped erroneously. To reduce the risk, we will use multiple techniques concurrently. A feature is dropped only if at least a certain number of these techniques all regard it as unpromising.

Technique 3: At the first stage for each learning algorithm, ensure a minimum number of tests conducted on every feature evaluator and feature search method. Every feature selection technique adopts a feature evaluator as well as a feature search method [25]. At the first stage for no learning algorithm, our draft method guarantees the number of tests conducted on every feature evaluator or feature search method. Without enough tests, we cannot tell how well a feature evaluator or feature search method works with the algorithm. To tackle this issue, at the first stage for each algorithm, we will check the number of tests conducted on every feature evaluator and feature search method. If the number for a feature evaluator or feature search method is smaller than a specific threshold (e.g., 3), we conduct more tests for the feature evaluator or feature search method to make up the difference. This approach can be adopted for several other components of a data analytic pipeline [86] like handling imbalanced classes and missing values.

Technique 4: Share information on the best few results obtained so far among different learning algorithms. Our draft method conducts a separate set of tests for every algorithm. When conducting tests for an algorithm, we may find a combination of a feature selection technique and its hyper-parameter values with superior performance. Yet, the combination may not be tested together with other algorithms, as its information is not shared with them. This can degrade the ultimate search result's quality. To tackle this issue, we will share information on the best few results obtained so far among different algorithms. At the end of each stage except for the last one, we identify a pre-chosen number n_1 (e.g., 3) of combinations of algorithms, techniques, and hyper-parameter values that achieve the lowest error rates among all combinations examined so far. Then we extract the corresponding n_2 combinations of techniques and their hyper-parameter values. Typically, $n_2=n_1$. Occasionally, n_2 can be $<n_1$, as the same combination of a technique and its hyper-parameter values may appear in more than one of the n_1 combinations. At the next stage, for each remaining algorithm, we ensure each of the n_2 combinations of techniques and their hyper-parameter values is tested by adding additional tests, if needed.

Technique 5: For a data set with relatively few data instances, dynamically allocate its data instances between the training and validation samples across stages. A data set with relatively few data instances can still be large if it contains many features. In this case, our draft method uses a fixed portion of it as the validation sample, which includes a small number of data instances. Because of insufficient testing, the error rate estimates obtained on the trained models can be non-robust, degrading the ultimate search result's quality. To tackle this issue, we will dynamically allocate the data instances in the data set between the training and validation samples across stages. At each stage except for the last one, we give all data instances that are in the data set, but not in the training sample to the validation sample. With more data instances in the validation sample, the error rate estimates obtained on the trained models can be more robust. Krueger *et al*. [87] used a similar approach to perform fast cross-validation to select a good hyper-parameter value combination for a given learning algorithm and modeling problem.

Technique 6: Consider distances between hyper-parameter value combinations when choosing randomly sampled combinations for testing. At each stage that is neither the first nor the final one, for each remaining learning algorithm, our draft method performs one or more rounds of Bayesian optimization. In each round, several new and randomly sampled combinations are chosen out of many for testing and used to adjust the regression model. For the regression model to guide search well, the combinations chosen for testing need to have a reasonable coverage of the hyper-parameter space rather than all reside in a small region. To achieve this, we will attempt to ensure that each randomly sampled combination chosen for testing is away from each other combination chosen for testing by at least a specific distance. The distance threshold may decrease over stages.

Aim 1.b: Devise a method to efficiently and automatically choose operators and periods for temporally aggregating clinical attributes.

Our design paper [24] outlines our method for automating the process of temporally aggregating clinical attributes. We will flesh out our method's technical details. Our automation method needs disease-specific knowledge on aggregation operators and periods compiled by clinicians and stored in Auto-ML. Various medical data sets use differing schemas, medical coding systems, and medical terminologies, forming a hurdle in applying pre-compiled knowledge. To tackle this, the automated temporal aggregation function of Auto-ML demands the data set, except for the dependent variable, to comply with the OMOP (Observational Medical Outcomes Partnership) common data model [88] and its linked standardized terminologies [89]. Since OMOP standardizes administrative and clinical attributes from ≥10 large U.S. healthcare systems [90, 91], Auto-ML can be adopted for data sets from those systems. We intend to include support for the PCORnet [92] and i2b2 common data models [93] in the future.

Aim 1.c: Continuously show, as a function of time given for model selection, forecasted model accuracy and projected patient outcomes of model use.

During automatic selection, to be more useful and user-friendly, Auto-ML will show projected patient outcomes of model use and forecasted model accuracy as a function of time given for model selection (Figure 3). Our design paper [24] outlines our method for doing this. We will flesh out our method's technical details and write a user manual for Auto-ML.

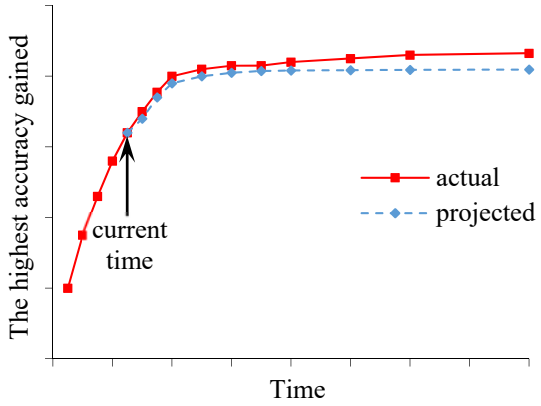

Figure 3. The highest model accuracy gained by Auto-ML over time.

Aim 1.d: Validate Auto-ML on seven benchmark modeling problems.

We will perform tests with healthcare researchers using seven modeling problems and data sets that we worked on before. Each problem uses a different data set from a distinct healthcare system. We chose these problems because they are on common diseases, are clinically important, and have readily accessible data sets. Auto-ML can be used for other clinical activities.

Subject recruitment: Via announcing in our institution's email lists and personal contact, we will recruit 25 healthcare researchers from the University of Washington School of Medicine with ~2,500 faculty members, most doing healthcare research. These healthcare researchers would regard their familiarity with medical data at MD level, but their machine learning knowledge below the level taught in a typical machine learning course for computer science undergraduates. We will conduct purposeful sampling to ensure enough variability [94]. All test participants will have fulfilled UWM's required training for information security and privacy policy. Participants will receive pseudonyms linking their responses to questions for privacy protection. After giving consent, each will get a copy of task description, Auto-ML's user manual, and the metadata document detailing the attributes in the data set. Upon task completion, each will receive $200 as compensation for participation.

Computing environment: We will perform all of our experiments on a HIPAA (Health Insurance Portability and Accountability Act)-compliant computer cluster at the University of Washington. After obtaining proper authorization, all test participants and research team members at the University of Washington can access the cluster using their university computers.

Modeling problem 1: Predict diagnosis of type 2 diabetes in adult patients in the next year.

Data set and patient population: The clinical and administrative data set is de-identified and publicly available from Practice Fusion's diabetes classification contest [15, 34], containing previous 3-year (2009-2012) records as well as the labels of 9,948 adult patients from all U.S. states in the following year. 1,904 of these patients had a diagnosis of type 2 diabetes in the following year. The data set comes from an electronic medical record vendor's EDW, includes repeatedly recorded attributes, and covers patient demographics, allergies, diagnoses, immunizations, medications, smoking status, lab results, and vital signs. We will put this data set in the OMOP common data model form with its linked standardized terminologies.

Model information: The dependent variable is in the following year, whether a patient had a diagnosis of type 2 diabetes. 2/3 of patients will be randomly selected and put into the training set to construct models. The remaining 1/3 of patients will form the test set for assessing model performance. We will use the area under the receiver operating characteristic curve (AUC) [19] performance metric.

Modeling problems 2-7: Each of the six problems uses a distinct, de-identified, and publicly available data set from the University of California, Irvine machine learning repository [95] to perform a task: 1) Arcene: classify mass-spectrometric data into cancer vs. normal patterns; 2) Arrhythmia: classify 12-lead electrocardiogram recordings into one of 16 groups about cardiac arrhythmia; 3) Cardiotocography: classify fetal cardiotocograms into one of three fetal states; 4) Diabetic Retinopathy Debrecen: use features obtained from the Messidor image set to detect whether an image includes signs of diabetic retinopathy; 5) Mammographic Mass: use Breast Imaging Reporting and Data System attributes and patient age to separate benign from malignant mammographic masses; 6) Parkinson Speech: use sound recordings to identify Parkinson's disease patients. No data set has repeatedly recorded attributes needing temporal aggregation. The repository [95] includes a detailed description of the problems and data sets. For each data set, 2/3 of it will be randomly selected and put into the training set to construct models. The remaining 1/3 of it will form the test set for assessing model performance. We will use the accuracy metric suitable for multi-class classification.

Build models: We are familiar with the literature on the seven modeling problems. For each problem, our data scientist Dr. Luo will work with the clinicians in our team and manually build a machine learning model with as high accuracy as possible. This accuracy will serve as the gold standard reflecting current best practice of model building. Each of the 25 recruited healthcare researchers will be randomly given a problem and use Auto-ML to build models for it.

Performance evaluation & sample size justification: We will test the hypothesis that ≥60% of healthcare researchers can use Auto-ML to achieve model accuracy ≥95% of the gold standard. When 60% of healthcare researchers can actually achieve model accuracy ≥95% of the gold standard, a sample size of 25 healthcare researchers produces a one-sided 95% lower confidence limit of 42%.

User feedback: When model construction is finished, we will use both open-ended and semi-structured questions to survey the 25 healthcare researchers. As detailed in our design paper [83], we will obtain quantitative outcome measures covering model accuracy, time on task, self-efficacy for constructing machine learning models with clinical big data, satisfaction, trustworthiness, adequacy, and quality of documentation. The questionnaire will contain a text field for gathering comments on Auto-ML. We will refine and finalize Auto-ML by considering suggestions from those comments. We will perform a user satisfaction survey using the System Usability Scale (SUS), a widely used industry standard [96, 97] on overall satisfaction rating for products.

Analysis: We will use the accepted inductive approach endorsed by Patton *et al*. [94, 98] to do qualitative analysis. We will put the 25 healthcare researchers' textual comments into ATLAS.ti, a qualitative analysis software tool [99]. The research team will independently highlight quotations related to the issue of using Auto-ML. We will examine quotations, categorize them into pre-codes, and merge them into categories in multiple iterations. We will synthesize categories to find general themes. Quantitative analyses will include adding the scores in the SUS and presenting every quantitative outcome measure's descriptive statistics.

Aim 2: Apply Auto-ML and novel methodology to two new modeling problems crucial for care management allocation and pilot one model with care managers.

We will apply Auto-ML to two modeling problems for care management allocation, to which our institutions are seeking solutions. Both use data sets having repeatedly recorded attributes. We will put the data sets in the OMOP common data model

form with its linked standardized terminologies. We will use the same computing environment and recruiting method mentioned in Aim 1.d. We will recruit two healthcare researchers not engaged in Aim 1.d. Each will be randomly given a problem and use Auto-ML to build models for it. Upon task completion, each will receive $200 as compensation for participation.

Modeling problem 8: Use vast attributes in modern IH electronic medical records to predict hospital use in asthmatic children in the next year.

Patient population: IH pediatric patients (age 0-17) in 2005-2016 with asthma, identified by Schatz *et al*.'s method [63, 100, 101] as having 1) at least one diagnosis code of asthma (ICD-9 (International Classification of Diseases, Ninth Revision) 493.xx or ICD-10 (International Classification of Diseases, Tenth Revision) J45/J46.*) or 2) ≥2 "asthma-related medication dispensings (excluding oral steroids) in a one-year period, including β-agonists (excluding oral terbutaline), inhaled steroids, other inhaled anti-inflammatory drugs, and oral leukotriene modifiers."

Data set: By running Oracle database SQL (Structured Query Language) queries, our contracted IH data analyst will extract from the IH EDW a de-identified, clinical and administrative data set, encrypt it, and securely transfer it to a HIPAA-compliant computer cluster for secondary analysis. For each of the last five years, the data cover ~27,000 asthmatic children. The data set is the electronic documentation of ~95% of pediatric care in Utah [102, 103] and includes ~400 attributes partially listed in our paper [14]. These attributes cover many known risk factors for hospital use in asthma and can be used to find new predictive factors.

Model information: The dependent variable is whether an asthmatic patient incurred hospital use (inpatient stay or ED visit) with a primary diagnosis of asthma (ICD-9 493.xx or ICD-10 J45/J46.*) in the following year [14, 63, 64]. As outcomes need to be computed for the following year, we effectively have 11 years' IH data. We will construct models using the data in the first 10 years, and acquire a model's accuracy estimate via testing on the data in the 11th year. This mirrors future use of the model in practice. We will use the AUC [19] performance metric.

Modeling problem 9: Use UWM's incomplete data to predict individual diabetic adults' costs in the next year.

Patient population: UWM adult patients (age≥18) in 2012-2016 with diabetes, identified by the method in Neuvirth *et al*. [104] as having one or more hemoglobin A1c test results ≥6.5%.

Data set: A UWM data analyst will run SQL Server database SQL queries to extract from the UWM EDW a de-identified, clinical and administrative data set, encrypt it, and securely transfer it to a HIPAA-compliant computer cluster for secondary analysis. The data cover ~28,000 diabetic adults per year. Other details of the data set are similar to those in modeling problem 8.

Model information: The dependent variable is a diabetic patient's total allowed cost to UWM in the following year [60, 61]. Allowed costs are less inflated than billed costs and less subject to variation due to member cost-sharing than net incurred claims [60, page 45]. We will adopt the medical consumer price index [105] to convert all costs to 2016 dollars to handle inflation. As outcomes need to be computed for the following year, we effectively have four years' UWM data. We will construct models using the data in the first three years, and acquire a model's accuracy estimate via testing on the data in the fourth year. This mirrors future use of the model in practice. We will use the R^2 performance metric [61].

To fill the scope gap mentioned in the introduction, we will use a constraint to find patients who tend to get most of their care at UWM. Intuitively, it is easier to identify future high-cost patients among them than among others. We will use UWM's incomplete data to build a cost prediction model and apply it to them. Regardless of his/her total future cost at non-UWM facilities, a patient who will incur high cost at UWM can be a candidate for care management. By care managing future high-cost patients identified by the model, we will expand the scope of using care management to improve outcomes. The principle of our approach to using incomplete data generalizes to many other clinical applications.

Several candidate constraints exist: 1) the patient had ≥2 visits to UWM in the past year, 2) the patient has a UWM primary care physician and lives within 5 miles of a UWM hospital, and 3) the patient saw a primary care physician or endocrinologist at UWM in the past year and lives within 60 miles (~1 hour of driving distance) of a UWM hospital. UWM primary care physicians tend to make referrals within UWM. Endocrinologists often serve some of the same roles as primary care physicians. Usually, a patient incurs high cost because of hospital use. As patients living far away from UWM hospitals are less likely to use them, UWM tends to have less of these patients' medical data. We will refine the three candidate constraints and investigate others. To select the constraint to be used, we will use PreManage data that UWM has on all of its patients. PreManage is Collective Medical Technologies Inc.'s commercial product providing encounter and diagnosis data on inpatient stays and ED visits at many U.S. hospitals [106]. PreManage data cover 105 (~94% of) hospitals in Washington, including the 4 hospitals of UWM. Using UWM data and grouper models like the Clinical Classifications Software system to group diagnosis codes and reduce features [60], we will build two models, one for estimating an inpatient stay's allowed cost and another for estimating an ED visit's allowed cost based on patient demographics and diagnosis data. We will use UWM patient demographics data, PreManage diagnosis data, and the two models to estimate the allowed cost of a UWM patient's every non-UWM inpatient

stay and ED visit reflected by PreManage encounter data. By aggregating the estimated costs of individual non-UWM inpatient stays and ED visits, we will assess each UWM patient's portion of cost spent at non-UWM hospitals and use the portions to evaluate every candidate constraint. If a healthcare system does not have enough data to make the two models reasonably accurate, it can use the average costs of an inpatient stay and ED visit to assess each patient's portion of cost spent at external hospitals. If a system has an insurance plan's complete claim data on some of its patients, it can use the data similarly.

Performance evaluation & sample size justification: For each of the two new modeling problems, we will test the hypothesis that healthcare researchers are able to use Auto-ML to achieve higher model accuracy than existing approaches. We will regard Aim 2 partly successful if we accept the hypothesis in only one problem, and completely successful if we accept the hypothesis in both problems.

For modeling problem 8, we will compare the accuracies reached by the model built by the healthcare researcher and the model in Schatz *et al* [65]. The first model is built using Auto-ML and vast attributes in modern IH electronic medical records. The second model depicting the existing approach was built using a few known risk factors for hospital use in asthma. Using vast attributes can increase prediction accuracy [107]. We will accept the hypothesis when the first model reaches higher AUC than the second one by ≥0.05. Existing predictive models for hospital use in asthma usually achieve an AUC far <0.8 [63-68]. Assuming these two models' prediction results have a correlation coefficient of 0.6 for both classes and performing a two-sided Z-test at a significance level of 0.05, a sample size of n=561 data instances per class provides 90% power to find a discrepancy of 0.05 between the two models' AUCs. The IH data in the 11th year include about 27,000 asthmatic children, offering enough power to test our hypothesis. Using many patients is essential for improving prediction accuracy, although only a small sample size is needed to show statistical significance.

For modeling problem 9, we will compare the accuracies gained by two models. The patient cohort includes those satisfying the chosen constraint. The first model is built by the healthcare researcher using Auto-ML and clinical and administrative data. The second model depicting the existing approach is a commercial claims-based one available at UWM achieving an R^2 <20%. Although the second model was not designed for such use, we will apply it to the patient cohort on whom UWM possibly has incomplete data, which is better than the normal practice of making no predictions. Adding clinical data can increase prediction accuracy [108]. We will accept the hypothesis when the first model reaches higher R^2 than the second one by ≥5%. Using an F-test at a significance level of 0.05 and under the assumption of the existence of 20 features from clinical data in addition to ≤300 features used in the second model, a sample size of 443 patients provides 90% power to identify an increase of 5% in R^2 from 20%. Using the second candidate constraint, we estimate that the patient cohort will cover ~22% of diabetic adult patients at UWM. The fourth year's UWM data include ~28,000 diabetic adults, offering enough power to test our hypothesis.

Pilot with care managers: We will pilot the model the healthcare researcher will build for modeling problem 9 with UWM care managers. As an UWM operational project, we are working on this modeling problem and have access to ~25 UWM care managers. Via announcing in their email lists and personal contact, we will recruit 5 care managers. We will conduct purposeful sampling to ensure enough variability [94]. All test participants will give consent and have fulfilled UWM's required training for information security and privacy policy. Participants will receive pseudonyms linking their responses to questions for privacy protection. Upon task completion, each will receive $200 as compensation for participation.

Table 2. The dependent variable list.

Variable	Description
Impact on enrollment decision	Response to the question: Will the prediction result and automatically generated explanations change your enrollment decision on the patient?
Usefulness of the prediction result	Response to the question: How useful is the prediction result? Rating is on a 1-7 Likert-type scale, with anchors of "not at all" (1) and "very useful" (7).
Usefulness of the automatically generated explanations	Response to the question: How useful are the automatically generated explanations? Rating is on a 1-7 Likert-type scale, with anchors of "not at all" (1) and "very useful" (7).
Trustworthiness of the prediction result	Response to the question: In your opinion, how much clinical sense does the prediction result make? Rating is on a 1-7 Likert-type scale, with anchors of "not at all" (1) and "completely" (7).
Trustworthiness of the automatically generated explanations	Response to the question: In your opinion, how much clinical sense do the automatically generated explanations make? Rating is on a 1-7 Likert-type scale, with anchors of "not at all" (1) and "completely" (7).

We will use our previously developed method [15] to automatically explain the model's prediction results. For each care manager, we will randomly select 20 UWM diabetic adult patients, half of whom the model predicts will incur a cost >$30,000. The care manager is unaware of any of these patients' outcome in the next year. For each patient, we will first show the care

manager the historical, de-identified patient attributes, then the prediction result and automatically generated explanations, and finally survey him/her using both open-ended and semi-structured questions. The questions will cover whether the prediction result and explanations will change his/her enrollment decision on the patient, their usefulness, and their trustworthiness as shown in Table 2. The questionnaire will contain a text field for gathering comments. We will analyze collected information in a similar way to Aim 1.d.

For modeling problem 8, medication order and refill information is needed for identifying asthma. The IH data set contains this because IH has its own health insurance plan. If too much refill information is missed at IH, data from the all-payer claims database [109] will be used. For modeling problem 9, in our ongoing UWM operational project, we have used ~30 attributes and ~6,000 patients to build a basic cost prediction model, which achieved an R^2 close to that of the commercial claims-based model. Since the healthcare researcher will use many more attributes and patients that should increase model accuracy, we expect the cost prediction model built by him/her to achieve higher R^2 than the claims-based model.

Although using a constraint to fill the scope gap partially addresses UWM data's incompleteness, UWM still has incomplete medical data on some of its patients satisfying the constraint. For each such diabetic patient, the dependent variable of the patient's total allowed cost to UWM is only part of the patient's total allowed cost to all systems. The patient's features are computed from incomplete data. Both factors may create difficulty for significantly improving R^2. If this occurs, we will revise the dependent variable to a diabetic patient's total allowed cost to UWM or reflected by PreManage data. On average, the revised dependent variable is closer to the patient's total allowed cost to all systems than the original one. Recall that based on UWM patient demographics and PreManage diagnosis data, we will use two models to estimate the allowed cost of the patient's every non-UWM inpatient stay and ED visit reflected by PreManage encounter data. We will supplement UWM data with PreManage data to make patient data more complete for computing patient features. This approach of using PreManage data and revising the dependent variable can be adopted to improve the accuracy of predicting future hospital use.

For either new modeling problem, if one healthcare researcher fails to build a reasonably accurate model, we will recruit another healthcare researcher.

Aim 3: Perform simulations to estimate the impact of adopting Auto-ML on U.S. patient outcomes.

To determine Auto-ML's value for future clinical deployment, we will estimate the impact of adopting Auto-ML on U.S. patient outcomes. Trials showed that machine learning helped drop 30-day mortality rate in ED patients with community-acquired pneumonia (risk ratio≈odds ratio=0.53, as the mortality rate is «1) [2], and cut hospitalization days by 15% in end-stage renal disease patients on dialysis [3]. We will use these two scenarios to demonstrate our simulation method. Our method generalizes to other scenarios and similar software. We will use the same computing environment mentioned in Aim 1.d. We first discuss the scenario of ED patients with community-acquired pneumonia.

Estimate outcomes: The outcome is 30-day mortality. We will use the latest, de-identified, and publicly available Nationwide Emergency Department Sample (NEDS) database [110] including visit information from ~20% of U.S. EDs. Consider the case with Auto-ML. The likelihood L that an ED can successfully use machine learning for this scenario = $p_1 \times p_2$. p_1 is the probability a healthcare researcher in the ED can build a high-quality machine learning model for this scenario using Auto-ML. p_2 is the probability the ED can successfully deploy the model if it can be built. Using Aim 1.d's test results on whether healthcare researchers can use Auto-ML to achieve model accuracy ≥95% of the gold standard, we will conservatively estimate p_1's minimum and maximum values, e.g., by fitting a normal distribution and using its 2.5 and 97.5 percentile points. Based on his extensive experience with deploying models [2], Dr. Haug will conservatively estimate p_2's minimum and maximum values. For each of p_1 and p_2, we will adopt five levels going from the minimum to the maximum value for sensitivity analysis. The middle level is the default one and used for hypothesis testing.

For each ED in the NEDS database, we will retrieve the annual number of patients with community-acquired pneumonia. We will simulate whether or not the ED can successfully use machine learning for this scenario based on the likelihood L. If success/not success, for each ED patient with community-acquired pneumonia, we will simulate whether the patient will die or not based on the 30-day mortality rate reported in the paper [2] when using/not using machine learning. The overall outcome estimate combines the expected outcomes for all patients and EDs. The patients' discharge weights in the NEDS database will be used to obtain national estimates from sample data in the database. We will handle the case without Auto-ML similarly by simulating not using machine learning.

Outcome evaluation & sample size justification: Outcomes achieved with and without Auto-ML will be compared. We will test the primary hypothesis that using Auto-ML will be linked to reduced mortality. In the most conservative case assuming a proportion of discordant pairs of 10%, a sample size of 1,152 patients provides 90% power to notice an odds ratio of 0.53 [2] using a two-sided McNemar test at a significance level of 0.05. Each year, community-acquired pneumonia incurs 1.5 million ED patient visits [111], giving adequate power to test the hypothesis. To acquire the whole range of possible outcomes, we will

do sensitivity analysis by changing the levels of the probabilities p_1 and p_2, 30-day mortality rate, and rate reduction gained by machine learning.

The scenario of end-stage renal disease patients on dialysis will be handled similarly, with the following main differences. The outcome is number of hospitalization days. The healthcare unit is dialysis facility. For each U.S. dialysis facility, we will obtain its latest annual total number of hospitalization days and patient count from dialysisdata.org [112] to fit a Poisson distribution. For each dialysis patient in the facility, we will simulate his/her annual number of hospitalization days using the distribution, as is often done in the literature [113]. We will test the secondary hypothesis that using Auto-ML will be linked to reduced hospitalization days.

If the results from a single simulation run appear too skewed, we will conduct multiple runs and then average their results.

Ethics approval

We have already acquired institutional review board approvals from UWM and IH for our study.

3. Results

Our paper [35] describes our draft method for automating machine learning model selection. The paper shows that compared to the modern Auto-WEKA automatic selection method [25], on six medical and 21 non-medical benchmark data sets, on average our draft method reduced search time by 28 fold, classification error rate by 11%, and standard deviation of error rate due to randomization by 36%. On each of these data sets, our draft method can finish the search process in 12 hours or less on a single computer. The results obtained on the medical data sets are similar to those obtained on the non-medical data sets. The healthcare researchers in the Veterans Affairs Salt Lake City Health Care System have used our draft method successfully for a clinical research project [114]. One purpose of this study is to improve the draft method so that it can handle larger data sets more efficiently and effectively.

At present, we are writing Auto-ML's design document. We intend to finish this study in around five years.

4. Discussion

Auto-ML will generalize to various clinical prediction/classification problems, as its design relies on no special property of a specific data set, patient population, or disease. Auto-ML will be tested on nine modeling problems and data sets, each from a distinct healthcare system. By providing support for common data models (e.g., OMOP [88]) and their linked standardized terminologies adopted by a large number of systems, Auto-ML can be used to construct models if attributes required to solve a problem are accessible in a structured data set or in one of those common data models. This enables data integration and facilitates building models with data from multiple systems. To help users decide whether any data quality issues need to be handled before modeling, Auto-ML will show the numbers of attribute values outside reasonable ranges and numbers of missing values of non-repeatedly recorded attributes.

The gaps in scope and accuracy mentioned in the introduction exist in many clinical applications. The principles of our approaches to using incomplete medical data and vast attributes generalize to many other clinical applications beyond the two on care management listed in the introduction.

In summary, our new software is designed to efficiently automate machine learning model selection and temporal aggregation of clinical attributes. By making machine learning feasible with limited budgets and data scientist resources, our new software will help realize value from clinical big data and improve patient outcomes. The models that will be built for the two new modeling problems will help improve care management outcomes.

Acknowledgments

We thank E. Sally Lee, Xinran Liu, Xueqiang Zeng, and Nickolas Robison for helpful discussions.

Authors' contributions

GL was mainly responsible for the paper. He conceptualized and designed the study, performed the literature review, and wrote the paper. BS, MJ, PT, AW, SM, PH, and FN offered feedback on miscellaneous medical issues, contributed to conceptualizing the presentation, and revised the paper. XS took part in conceptualizing and writing the statistical analysis sections.

Conflicts of interest

None declared.

Abbreviations

AUC: area under the receiver operating characteristic curve
Auto-ML: Automated Machine Learning
ED: emergency department
EDW: enterprise data warehouse
FEV1: forced expiratory volume in 1 second
FVC: forced vital capacity
HIPAA: Health Insurance Portability and Accountability Act
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
IH: Intermountain Healthcare
NEDS: Nationwide Emergency Department Sample
OMOP: Observational Medical Outcomes Partnership
SQL: Structured Query Language
SUS: System Usability Scale
UWM: University of Washington Medicine

References

1. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer; 2009. ISBN:038777243X
2. Dean NC, Jones BE, Jones JP, Ferraro JP, Post HB, Aronsky D, Vines CG, Allen TL, Haug PJ. Impact of an electronic clinical decision support tool for emergency department patients with pneumonia. Ann Emerg Med 2015;66(5):511-20. PMID:25725592
3. Barbieri C, Molina M, Ponce P, Tothova M, Cattinelli I, Ion Titapiccolo J, Mari F, Amato C, Leipold F, Wehmeyer W, Stuard S, Stopper A, Canaud B. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. Kidney Int 2016;90(2):422-9. PMID:27262365
4. Brier ME, Gaweda AE, Dailey A, Aronoff GR, Jacobs AA. Randomized trial of model predictive control for improved anemia management. Clin J Am Soc Nephrol 2010;5(5):814-20. PMID:20185598
5. Gaweda AE, Aronoff GR, Jacobs AA, Rai SN, Brier ME. Individualized anemia management reduces hemoglobin variability in hemodialysis patients. J Am Soc Nephrol 2014;25(1):159-66. PMID:24029429
6. Gaweda AE, Jacobs AA, Aronoff GR, Brier ME. Model predictive control of erythropoietin administration in the anemia of ESRD. Am J Kidney Dis 2008;51(1):71-9. PMID:18155535
7. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online 2013;12 Suppl 1:S4. PMID:24565021
8. Hamlet KS, Hobgood A, Hamar GB, Dobbs AC, Rula EY, Pope JE. Impact of predictive model-directed end-of-life counseling for Medicare beneficiaries. Am J Manag Care 2010;16(5):379-84. PMID:20469958
9. Jvion's latest predictive analytics in healthcare survey finds that advanced predictive modeling solutions are taking a strong foothold in the industry. http://chimecentral.org/jvion-releases-findings-latest-predictive-analytics-healthcare-survey/. Archived at http://www.webcitation.org/6oOiQpFqo
10. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer; 2013. ISBN:1461468485
11. Axelrod RC, Vogel D. Predictive modeling in health plans. Disease Management & Health Outcomes 2003;11(12):779-87. doi:10.2165/00115677-200311120-00003
12. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLoS One 2014;9(2):e88225. PMID:24520356
13. Kaggle homepage. https://www.kaggle.com/. Archived at http://www.webcitation.org/6oOiI4j9t
14. Luo G, Stone BL, Sakaguchi F, Sheng X, Murtaugh MA. Using computational approaches to improve risk-stratified patient management: rationale and methods. JMIR Res Protoc 2015;4(4):e128. PMID:26503357
15. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2. PMID:26958341
16. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute report. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation. Archived at http://www.webcitation.org/6oOjj3GpL
17. Auffray C, Balling R, Barroso I, Bencze L, Benson M, Bergeron J, Bernal-Delgado E, Blomberg N, Bock C, Conesa A, Del Signore S, Delogne C, Devilee P, Di Meglio A, Eijkemans M, Flicek P, Graf N, Grimm V, Guchelaar HJ, Guo YK, Gut IG, Hanbury A, Hanif S, Hilgers RD, Honrado Á, Hose DR, Houwing-Duistermaat J, Hubbard T, Janacek SH, Karanikas H, Kievits T, Kohler M, Kremer A, Lanfear J, Lengauer T, Maes E, Meert T, Müller W, Nickel D, Oledzki P, Pedersen B, Petkovic M, Pliakos K, Rattray M, I Màs JR, Schneider R, Sengstag T, Serra-Picamal X, Spek W, Vaas LA,

van Batenburg O, Vandelaer M, Varnai P, Villoslada P, Vizcaíno JA, Wubbe JP, Zanetti G. Making sense of big data in health research: Towards an EU action plan. Genome Med 2016;8(1):71. PMID:27338147
18. Hoskins M. Common big data challenges and how to overcome them. Big Data 2014;2(3):142-3. PMID:27442494
19. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann; 2016. ISBN:0128042915
20. Rao RB, Fung G. On the dangers of cross-validation. an experimental evaluation. Proc. SDM 2008:588-96. doi:10.1137/1.9781611972788.54
21. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 2010;11:2079-107.
22. Reunanen J. Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research 2003;3:1371-82.
23. Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. Generalization in adaptive data analysis and holdout reuse. Proc. NIPS 2015:2350-8.
24. Luo G. PredicT-ML: A tool for automating machine learning model building with big clinical data. Health Inf Sci Syst 2016;4:5. PMID:27280018
25. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. Proc. KDD 2013:847-55. doi:10.1145/2487575.2487629
26. Luo G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw Model Anal Health Inform Bioinform 2016;5:18. doi:10.1007/s13721-016-0125-6
27. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Proc. NIPS 2012:2960-8.
28. Komer B, Bergstra J, Eliasmith C. Hyperopt-sklearn: automatic hyperparameter configuration for scikit-learn. Proc. SciPy 2014:33-9.
29. Kotthoff L, Thornton C, Hoos H, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. Journal of Machine Learning Research 2017;18(25):1-5.
30. Salvador MM, Budka M, Gabrys B. Towards automatic composition of multicomponent predictive systems. Proc. HAIS 2016:27-39. doi:10.1007/978-3-319-32034-2_3
31. Salvador MM, Budka M, Gabrys B. Automatic composition and optimisation of multicomponent predictive systems. https://arxiv.org/abs/1612.08789.
32. Kraska T, Talwalkar A, Duchi JC, Griffith R, Franklin MJ, Jordan MI. MLbase: a distributed machine-learning system. Proc. CIDR 2013.
33. Sparks ER, Talwalkar A, Haas D, Franklin MJ, Jordan MI, Kraska T. Automating model search for large scale machine learning. Proc. SoCC 2015:368-80. doi:10.1145/2806777.2806945
34. Practice Fusion diabetes classification homepage. https://www.kaggle.com/c/pf2012-diabetes. Archived at http://www.webcitation.org/6oOiaVvU5
35. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. http://pages.cs.wisc.edu/~gangluo/progressive_sampling.pdf. Archived at http://www.webcitation.org/6rz22eiKY
36. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proc. KDD 2015:1721-30. doi:10.1145/2783258.2788613
37. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc 2014;21(4):699-706. PMID:24481703
38. Borsi JP. Hypothesis-free search for connections between birth month and disease prevalence in large, geographically varied cohorts. AMIA Annu Symp Proc 2016:319-25. PMID:28269826
39. Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc. AMIA Symp 2000:923-7. PMID:11080019
40. Hickins M. Citizen data scientists unite! http://www.forbes.com/sites/oracle/2016/10/03/citizen-data-scientists-unite. Archived at http://www.webcitation.org/6oOjBNVtO
41. Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB, Blumenthal D. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. J Gen Intern Med 2007;22 Suppl 3:391-5. PMID:18026807
42. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014;33(1):124-31. PMID:24395944
43. Nelson L. Lessons from Medicare's demonstration projects on disease management and care coordination. https://www.cbo.gov/sites/default/files/112th-congress-2011-2012/workingpaper/WP2012-01_Nelson_Medicare_DMCC_Demonstrations_1.pdf. Archived at http://www.webcitation.org/6agZcFxD1

44. Asthma. http://www.cdc.gov/nchs/fastats/asthma.htm. Archived at http://www.webcitation.org/6agaQMYxr
45. Akinbami LJ, Moorman JE, Liu X. Asthma prevalence, health care use, and mortality: United States, 2005-2009. Natl Health Stat Report 2011;(32):1-14. PMID:21355352
46. Akinbami LJ, Moorman JE, Bailey C, Zahran HS, King M, Johnson CA, Liu X. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS Data Brief 2012;(94):1-8. PMID:22617340
47. Asthma in the US. http://www.cdc.gov/vitalsigns/asthma/. Archived at http://www.webcitation.org/6oOjKVf75
48. Levine SH, Adams J, Attaway K, Dorr DA, Leung M, Popescu P, Rich J. Predicting the financial risks of seriously ill patients. California HealthCare Foundation. http://www.chcf.org/publications/2011/12/predictive-financial-risks. Archived at http://www.webcitation.org/6oOjQCINu
49. Rubin RJ, Dietrich KA, Hawk AD. Clinical and economic impact of implementing a comprehensive diabetes management program in managed care. J Clin Endocrinol Metab 1998;83(8):2635-42. PMID:9709924
50. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999;103(3 Pt 1):436-40. PMID:10069877
51. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in Medicaid. Pediatrics 2000;105(5):1029-35. PMID:10790458
52. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42.
53. Beaulieu N, Cutler DM, Ho K, Isham G, Lindquist T, Nelson A, O'Connor P. The business case for diabetes disease management for managed care organizations. Forum for Health Economics & Policy 2006;9(1):1-37.
54. Dorr DA, Wilcox AB, Brunker CP, Burdon RE, Donnelly SM. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. J Am Geriatr Soc 2008;56(12):2195-202. PMID:19093919
55. Forno E, Celedón JC. Predicting asthma exacerbations in children. Curr Opin Pulm Med 2012;18(1):63-9. PMID:22081091
56. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-36. PMID:15451964
57. Curry N, Billings J, Darin B, Dixon J, Williams M, Wennberg D. Predictive Risk Project Literature Review. London: King's Fund. http://www.kingsfund.org.uk/sites/files/kf/field/field_document/predictive-risk-literature-review-june2005.pdf. Archived at http://www.webcitation.org/6aga2XBFC
58. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010;170(22):1989-95. PMID:21149756
59. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc. 2011;2011:409-16. PMID:22195094
60. Duncan I. Healthcare Risk Adjustment and Predictive Modeling. Winsted, CT: ACTEX Publications Inc.; 2011. ISBN:1566987695
61. Ash A, McCall N. Risk assessment of military populations to predict health care cost and utilization. http://www.rti.org/pubs/tricare_riskassessment_final_report_combined.pdf. Archived at http://www.webcitation.org/6aga7wYZC
62. Iezzoni LI. Risk adjustment for measuring health care outcomes, 4th ed. Chicago, IL: Health Administration Press; 2013. ISBN:1567934374
63. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003;9(8):538-47. PMID:12921231
64. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999;36(4):359-70. PMID:10386500
65. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004;10(1):25-32. PMID:14738184
66. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998;157(4 Pt 1):1173-80. PMID:9563736
67. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, Sylvia JM, Weiss ST, Celedón JC. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010;138(5):1156-65. PMID:20472862
68. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, Wenzel SE, Weiss ST. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006;28(6):1145-55. PMID:16870656
69. Schatz M. Predictors of asthma control: what can we modify? Curr Opin Allergy Clin Immunol 2012;12(3):263-8. PMID:22517290
70. Stanford RH, Shah MB, D'Souza AO, Schatz M. Predicting asthma outcomes in commercially insured and Medicaid populations. Am J Manag Care 2013;19(1):60-7. PMID:23379745

71. Hyland ME, Whalley B, Halpin DM, Greaves CJ, Seamark C, Blake S, Pinnuck M, Ward D, Hawkins A, Seamark D. Frequency of non-asthma GP visits predicts asthma exacerbations: an observational study in general practice. Prim Care Respir J 2012;21(4):405-11. PMID:22836742
72. Crawford AG, Fuhr JP Jr, Clarke J, Hubbs B. Comparative effectiveness of total population versus disease-specific neural network models in predicting medical costs. Dis Manag 2005;8(5):277-87. PMID:16212513
73. Coyle YM. Predictors of acute asthma relapse: strategies for its prevention. J Asthma 2003;40(3):217-24. PMID:12807164
74. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016;Suppl 1:S48-61. PMID:27199197
75. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis, 2nd ed. Hoboken, NJ: Wiley; 2011. ISBN:0470380276
76. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Proc. OSDI 2004:137-50.
77. White T. Hadoop: The Definitive Guide, 4th ed. Sebastopol, CA: O'Reilly Media; 2015. ISBN:1491901632
78. Karau H, Konwinski A, Wendell P, Zaharia M. Learning Spark: Lightning-Fast Big Data Analysis. Sebastopol, CA: O'Reilly Media; 2015. ISBN:1449358624
79. Meng X, Bradley J, Yuvaz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai D, Amde M, Owen S, Xin D, Xin R, Franklin M, Zadeh R, Zaharia M, Talwalkar A. MLlib: machine learning in Apache Spark. Journal of Machine Learning Research 2016;17(34):1-7.
80. Xin RS, Rosen J, Zaharia M, Franklin MJ, Shenker S, Stoica I. Shark: SQL and rich analytics at scale. Proc. SIGMOD 2013:13-24. doi:10.1145/2463676.2465288
81. Mining big data using Weka 3. http://www.cs.waikato.ac.nz/ml/weka/bigdata.html. Archived at http://www.webcitation.org/6oOkT63lx
82. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Proc. NIPS 2015:2944-52.
83. Luo G. MLBCD: A machine learning tool for big clinical data. Health Inf Sci Syst 2015;3:3. PMID:26417431
84. Provost FJ, Jensen D, Oates T. Efficient progressive sampling. Proc. KDD 1999:23-32. doi:10.1145/312129.312188
85. Wolpert DH. The lack of a priori distinctions between learning algorithms. Neural Computation 1996;8(7):1341-90. doi:10.1162/neco.1996.8.7.1341
86. Zhang Y, Bahadori MT, Su H, Sun J. FLASH: fast Bayesian optimization for data analytic pipelines. Proc. KDD 2016:2065-74. doi:10.1145/2939672.2939829
87. Krueger T, Panknin D, Braun ML. Fast cross-validation via sequential testing. Journal of Machine Learning Research 2015;16(1):1103-55.
88. Observational Medical Outcomes Partnership (OMOP) Common Data Model homepage. http://omop.org/CDM. Archived at http://www.webcitation.org/6agamjByZ
89. Observational Medical Outcomes Partnership (OMOP) vocabularies. http://omop.org/Vocabularies. Archived at http://www.webcitation.org/6oOkOnGg5
90. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60. PMID:22037893
91. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-8. PMID:26262116
92. The National Patient-Centered Clinical Research Network (PCORnet) Common Data Model homepage. http://www.pcornet.org/resource-center/pcornet-common-data-model/. Archived at http://www.webcitation.org/6oOkB07wd
93. Informatics for Integrating Biology and the Bedside (i2b2) Design Document Data Repository (CRC) Cell. https://www.i2b2.org/software/files/PDF/current/CRC_Design.pdf. Archived at http://www.webcitation.org/6oOkHsUCx
94. Patton MQ. Qualitative Research & Evaluation Methods, 3rd ed. Thousand Oaks, CA: SAGE Publications; 2001. ISBN:0761919716
95. University of California, Irvine machine learning repository. http://archive.ics.uci.edu/ml/. Archived at http://www.webcitation.org/6oOkdrOjg
96. Brooke J. SUS - A quick and dirty usability scale. http://hell.meiert.org/core/pdf/sus.pdf, 1996. Archived at http://www.webcitation.org/6oOmoIihR
97. Tullis T, Albert W. Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics, 2nd ed. Waltham, MA: Morgan Kaufmann; 2013. ISBN:0124157815
98. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. American Journal of Evaluation 2006;27(2):237-46. doi:10.1177/1098214005283748

99. ATLAS.ti qualitative analysis software. http://www.atlasti.com/index.html. Archived at http://www.webcitation.org/6oOjqY7rj
100. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012;50 Suppl:S30-5. PMID:22692256
101. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006;41(10):962-71. PMID:16871628
102. James BC, Savitz LA. How Intermountain trimmed health care costs through robust quality improvement efforts. Health Aff 2011;30(6):1185-91. PMID:21596758
103. Byington CL, Reynolds CC, Korgenski K, Sheng X, Valentine KJ, Nelson RE, Daly JA, Osguthorpe RJ, James B, Savitz L, Pavia AT, Clark EB. Costs and infant outcomes after implementation of a care process model for febrile infants. Pediatrics 2012;130(1):e16-24. PMID:22732178
104. Neuvirth H, Ozery-Flato M, Hu J, Laserson J, Kohn MS, Ebadollahi S, Rosen-Zvi M. Toward personalized care management of patients at risk: the diabetes case study. Proc. KDD 2011:395-403. doi:10.1145/2020408.2020472
105. Consumer Price Index - Measuring price change for medical care in the CPI. http://www.bls.gov/cpi/cpifact4.htm. Archived at http://www.webcitation.org/6agaajPCS
106. PreManage of Collective Medical Technologies Inc. http://collectivemedicaltech.com/what-we-do-2/premanage/. Archived at http://www.webcitation.org/6oOi8en1e
107. Sun J, Hu J, Luo D, Markatou M, Wang F, Edabollahi S, Steinhubl SE, Daar Z, Stewart WF. Combining knowledge and data driven insights for identifying risk factors using electronic health records. AMIA Annu Symp Proc. 2012;2012:901-10. PMID:23304365
108. Selby JV, Karter AJ, Ackerson LM, Ferrara A, Liu J. Developing a prediction rule from automated clinical databases to identify high-risk patients in a large population with diabetes. Diabetes Care 2001;24(9):1547-55. PMID:11522697
109. The APCD (all-payer claims database) Council homepage. http://www.apcdcouncil.org/. Archived at http://www.webcitation.org/6agaeGTXm
110. The Nationwide Emergency Department Sample (NEDS) database homepage. https://www.hcup-us.ahrq.gov/db/nation/neds/nedsdbdocumentation.jsp. Archived at http://www.webcitation.org/6oOi2Tqjz
111. Kyriacou DN, Yarnold PR, Soltysik RC, Self WH, Wunderink RG, Schmitt BP, Parada JP, Adams JG. Derivation of a triage algorithm for chest radiography of community-acquired pneumonia patients in the emergency department. Acad Emerg Med 2008;15(1):40-4. PMID:18211312
112. https://www.dialysisdata.org/. Archived at http://www.webcitation.org/6oOhv8xXR
113. Arora P, Kausz AT, Obrador GT, Ruthazer R, Khan S, Jenuleson CS, Meyer KB, Pereira BJ. Hospital utilization among chronic dialysis patients. J Am Soc Nephrol 2000;11(4):740-6. PMID:10752533
114. Divita G, Luo G, Tran LT, Workman TE, Gundlapalli AV, Samore MH. General symptom extraction from VA electronic medical notes. Stud Health Technol Inform 2017.