Reflex Testing using Machine Learning in the Clinical Laboratory

Kyle Belanger

8/2/2022

Table of contents

[Introduction 3](#_Toc115118100)

[Purpose and Research Statement 5](#_Toc115118101)

[Significance 5](#_Toc115118102)

[Literature Review 7](#_Toc115118103)

[A Brief Primer on Machine Learning 7](#_Toc115118104)

[Supervised vs Unsupervised Learning 8](#_Toc115118105)

[Machine Learning Workflow 9](#_Toc115118106)

[Machine Learning in the Clinical Laboratory 10](#_Toc115118107)

[Reflex Testing 11](#_Toc115118108)

[Chapter 3 12](#_Toc115118109)

[Proposed Study Set Up 12](#_Toc115118110)

[References 13](#_Toc115118111)

# Introduction

The early 20th century marked the beginning of a quality movement in hospitals and laboratories that began with physicians and healthcare workers. In the early part of the century, many hospitals started reorganizing their laboratories to be headed by biochemists. Professional organizations emerged as self-regulating groups that helped ensure the skills and knowledge of laboratory professionals would pass the scrutiny of the hospitals that employed them (Berger, 1999). An American Medical Association survey later showed that 48% of U.S. hospitals had clinical laboratories by 1923 (Berger, 1999). Before 1960, almost all testing in the laboratory was performed using manual methods. In the mid-1960s, a limited amount of automated analyzers became available, allowing for more rapid testing and running multiple tests at the same time (Park & Kricka, 2017). Since these early days of automation in the last fifty years, the clinical laboratory has rapidly expanded automation techniques. These include pre-packaged ready-to-use reagents, automated dispensing, incubation and measurement, automated sample processing (e.g., total laboratory automation systems, one- and two-dimensional bar codes, radio frequency identification tags), multiplexing tests from a single sample (e.g., microarrays), automated data processing (e.g., reference range, alert value comparisons, quality control assessment), automated interpretation (e.g., auto-verification), image analysis (e.g., automated peripheral blood smear morphology - CellaVision, whole slide scanning in surgical pathology), and mobile or static robots to operate analyzers (Park & Kricka, 2017). This rise in automation in the clinical laboratory has also led to the need for more advanced computer systems to go along with the advances in instrument technology. Over the past few decades, LISs have evolved from relatively narrow, often arcane, or home-grown systems into sophisticated systems that are more user-friendly and support a broader range of functions and integration with other technologies that laboratories deploy (Henricks, 2015). Modern LISs consist of complex, interrelated computer programs and infrastructure that support laboratories’ vast array of information-processing needs. LISs have functions in all phases of patient testing, including specimen and test order intake, specimen processing and tracking, support of analysis and interpretation, and report creation and distribution. In addition, LISs provide management reports and other data that laboratories need to run their operations and to support continuous improvement and quality initiatives (Henricks, 2015).

The clinical laboratory’s primary business purpose is to provide testing results requested by physicians and other healthcare professionals. In a broad sense, this testing is used to help solve diagnostic problems (Verboeket-van de Venne et al., 2012). To continue adding value to the laboratory’s business purpose, laboratory professionals can add value beyond just running the provided tests. Laboratory professionals can add value through both reflective and reflex testing. Automated analyzers add most tests based on rules (algorithms) established by laboratory professionals; this is defined as ‘reflex testing.’ Clinical biochemists add the remainder of tests after considering a more comprehensive range of information that can readily be incorporated into reflex testing algorithms; this is defined as ‘reflective testing’ (Srivastava et al., 2010). Both reflex and reflective testing became possible with the advent of laboratory information systems (LIS) that were sufficiently flexible to permit modification of existing test requests at various stages of the analytical process (Srivastava et al., 2010). This research study will focus specifically on reflex testing, those tests added automatically by a set of rules established in each laboratory. In most current clinical laboratories, reflex testing is performed with a ‘hard’ cutoff, using a specifically established range with no means of flexibility (Murphy, 2021). This study will examine the use of Machine learning to develop algorithms to allow flexibility for automatic reflex testing in clinical chemistry. The goal is to fill the gap between hard-coded reflex testing and fully manual reflective testing using machine learning algorithms.

## Purpose and Research Statement

Develop and test a machine learning algorithm to establish if said algorithm can perform better then current hard coded rules to reduced unnecessary patient testing.

## Significance

Health spending in the U.S. increased by 4.6% in 2019 to $3.8 trillion or $11,582 per capita. This growth rate is in line with 2018 (4.7 percent) and slightly faster than what was observed in 2017 (4.3 percent) (American Medical Association, 2021). Although laboratory costs comprise only about 5% of the healthcare budget in the United States, it is estimated that laboratory services drive up to 70% of all downstream medical decisions, which encompass a substantial portion of the budget (Ma et al., 2019). As healthcare budgets increase, payers, including Medicare, commercial insurers, and employers, will demand accountability and eliminate the abuse and misuse of ineffective testing strategies (Hernandez, 2003). Increasingly, payers demand to know the value of the tests, with value equaling quality per unit of cost. Payers want laboratories to prove that tests are cost-effective; as reimbursement rates decline for many standard laboratory tests, the incentives for automated reflex testing rise for many clinical laboratories (Hernandez, 2003). Unnecessary laboratory tests are a significant source of waste in the United States healthcare system. Prior studies suggest that 20% of labs performed are unnecessary, wasting 200 billion dollars annually [Li et al. (2022)].

A typical example of reflex testing is thyrotropin (TSH), relaxing to free thyroxine (Free T4 or FT4). TSH measurement is a sensitive screening test for thyroid dysfunction. Guidelines from the American Thyroid Association, the American Association of Clinical Endocrinologists, and the National Academy of Clinical Biochemistry have endorsed TSH measurement as the best first-line strategy for detecting thyroid dysfunction in most clinical settings (Plebani & Giovanella, 2020). Traditionally the cutoff for reflex testing was simply the reference range for a patient’s sex and race. However, recent studies have suggested that widening these ranges reduces reflex testing by up to 34% (Plebani & Giovanella, 2020). In an additional study, the authors concluded that the TSH reference range leading to reflex Free T4 testing could likely be widened to decrease the number of unnecessary Free T4 measurements performed. This reduction would reduce overall costs to the medical system without likely causing negative consequences of missing the detection of people with thyroid hormone abnormalities [Woodmansee (2018)]. Even with the potential reduction in testing, the hard-coded reflex rule still exists.

# Literature Review

The application of machine learning in medicine has garnered enormous attention over the past decade (Rabbani et al., 2022). Artificial intelligence (AI) and especially the subdiscipline of machine learning (ML) have become hot topics generating increasing interest among laboratory professionals. AI is a rather broad term and can be defined as the theory and development of computer systems to perform complex tasks typically requiring human intelligence, such as decision-making, visual perception, speech recognition, and translation between languages. ML is the science of programming, allowing computers to learn from data without being explicitly programmed (De Bruyne et al., 2021). The ever more extensive use of ML in clinical and basic medical research is reflected in the number of titles and abstracts of papers indexed on PubMed and published until 2006 as compared to 2007–2017, with a nearly 10-fold increase from 1000 to slightly more than 9000 articles in that time frame (Cabitza & Banfi, 2018). A literature review by Rabbani et al. found 39 articles about the field of clinical chemistry in laboratory medicine between 2011 and 2021 [-Rabbani et al. (2022)].

## A Primer on Machine Learning

While this literature review aims not to provide an extensive representation of the mathematics behind ML algorithms, some basic concepts will be introduced to allow a sufficient understanding of the topics discussed in the paper. ML models can be classified into broad categories based on several criteria. These categories include the type of supervision, whether are not the algorithm can learn incrementally from an incoming stream of data (batch and online learning), and how they generalize (instance-based versus model-based learning) (De Bruyne et al., 2021). Rabbani et al. further classified the specific clinical chemistry uses into five board categories, predicting laboratory test values, improving laboratory utilization, automating laboratory processes, promoting precision laboratory test interpretation, and improving laboratory medicine information systems [-Rabbani et al. (2022)].

### Supervised vs Unsupervised Learning

Four important categories can be distinguished based on the amount and type of supervision the models receive during training: supervised, unsupervised, semi-supervised, and reinforcement learning. In supervised learning, training data are labeled, and data samples are predicted with knowledge about the desired solutions (De Bruyne et al., 2021). They are typically used for classification and regression purposes. Some of the essential supervised algorithms are Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Decision Trees (DTs), Random Forests (RFs), and supervised neural networks. In unsupervised learning, training data are unlabeled. In other words, observations are classified without prior data sample knowledge (De Bruyne et al., 2021). Unsupervised algorithms can be used for clustering (e.g., k-means clustering, density-based spatial clustering of applications with noise, hierarchical cluster analysis), visualization, and dimensionality reduction (e.g., principal component analysis (PCA), kernel PCA, locally linear embedding, t-distributed stochastic neighbor embedding), anomaly detection and novelty detection (e.g., one-class SVM, isolation forest) and association rule learning (e.g. apriori, eclat). However, some models can deal with partially labeled training data (i.e., semi-supervised learning). At last, in reinforcement learning, an agent (i.e., the learning system) learns what actions to take to optimize the outcome of a strategy (i.e., a policy) or to get the maximum cumulative reward [De Bruyne et al. (2021)]. This system resembles humans learning to ride a bike and can typically be used in learning games, such as Go, chess, or even poker, or settings where the outcome is continuous rather than dichotomous (i.e., right or wrong)(De Bruyne et al., 2021). The proposed study will use supervised learning, as the data is labeled and a particular outcome is expected.

### Machine Learning Workflow

Since this study will focus on supervised learning, the review will focus on that. Machine learning can be broken into three board steps, data cleaning and processing, training and testing the model, and finally, the model is evaluated, deployed, and monitored (De Bruyne et al., 2021). In the first phase, data is collected, cleaned, and labeled. Data cleaning or pre-processing is one of the essential steps in designing a reliable model (De Bruyne et al., 2021). Some examples of common pre-processing steps are the handling of missing data, detection of outliers, and encoding of categorical data. Data at this stage is also split into training and testing data, typically following somewhere near a 70-30 split. These two data sets are used for different portions of the rest of the model building. The Training set data is used to develop feature sets, train our algorithms, tune hyperparameters, compare models, and all the other activities required to choose a final model (e.g., the model we want to put into production) (Boehmke & Greenwell, 2020). Once the final model is chosen, the test set data is used to estimate an unbiased assessment of the model’s performance, which we refer to as the generalization error (Boehmke & Greenwell, 2020). Most time (as much as 80%) is invested into the data processes stage. After feature engineering, an ML model is trained and tested on the collected data in the second phase. Feature engineering is performed on the training set to select a good set of features to train on. The ML model will only be able to learn efficiently if the training data contains enough relevant features and minimal irrelevant ones [Géron (2019)]. The data is then run through various models, Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs).

Once a model is selected, the third phase begins to evaluate the model’s performance. Historically, the performance of statistical models was primarily based on goodness-of-fit tests and the assessment of residuals. Unfortunately, misleading conclusions may follow from predictive models that pass these assessments (Breiman, 2001). Today, it has become widely accepted that a more sound approach to assessing model performance is to assess the predictive accuracy via loss functions (Boehmke & Greenwell, 2020). *Loss functions* are metrics that compare the predicted values to the actual value (the output of a loss function is often referred to as the error or pseudo residual). When performing resampling methods, we assess the predicted values for a validation set compared to the actual target value. The overall validation error of the model is computed by aggregating the errors across the entire validation data set [Boehmke & Greenwell (2020)]

.

### Machine Learning in the Clinical Laboratory

Table 1. Summary of characteristics of machine learning algorithms (Rabbani et al., 2022).

| **Author and Year** | **Objective and Machine Learning Task** | **Best Model** | **Major Themes** |
| --- | --- | --- | --- |
| Azarkhish (2012) | Predict iron deficiency anemia and serum iron levels from CBC indices | Neural Network | Prediction |
| Cao (2012) | Triage manual review for urinalysis samples | Tree-based | Automation |
| Yang (2013) | Predict normal reference ranges of ESR for various laboratories based on geographic and other clinical features | Neural Network | Interpretation |

## Reflex Testing

The laboratory diagnosis of thyroid dysfunction relies on the measurement of circulating concentrations of thyrotropin (TSH), free thyroxine (fT4), and, in some cases, free triiodothyronine (fT3). TSH measurement is generally regarded as the most sensitive initial laboratory test for screening individuals for thyroid hormone abnormalities (Woodmansee, 2018). TSH and fT4 have a complex, nonlinear relationship, such that small changes in fT4 result in relatively large changes in TSH (Plebani & Giovanella, 2020). Many clinicians and laboratories check TSH alone as the initial test for thyroid problems and then only add a Free T4 measurement if the TSH is abnormal (outside the laboratory normal reference range), this is known as reflex testing (Woodmansee, 2018). Reflex testing became possible with the advent of laboratory information systems (LIS) that were sufficiently flexible to permit modification of existing test requests at various stages of the analytical process (Srivastava et al., 2010). Reflex testing is widely used, the major aim being to optimize the use of laboratory tests. However the common practice of reflex testing relies simply on hard coded rules that allow no flexibility. For instance in the case of TSH, free T4 will be added to the patient order whenever the value falls outside of the established laboratory reference range. This bring into the fold the issue that the thresholds used to trigger reflex addition of tests vary widely. In a study by Murphy he found the hypocalcaemic threshold to trigger magnesium measurement varied from 1.50 mmol/L up to 2.20 mmol/L (2021). Even allowing for differences in the nature, size and staffing of hospital laboratories, and populations served, the extent of the observed variation invites scrutiny (Murphy, 2021).

# Chapter 3

## Proposed Study Set Up

Using the Medical Information Mart for Intensive Care (MIMIC) IV Database develop and test a machine learning algorithm to determine if TSH reflex testing can be further reduced.

The MIMIC-IV database contains patient records from 2008 to 2019 for patients admitted to the critical care units of Beth Israel Deaconess Medical Center. It is a common database used for various studies. The data will be cleaned and tided to contain various patient demographics, and all available laboratory testing for each patient. The exact structure of the cleaned data will be determined later. Once cleaned the data will be split into a training and testing data set. The training data will be used to develop various machine learning algorithms to attempt to develop an algorithm that can perform better then the hard coded rules in place today. The study will primarily focus on TSH reflex testing as this is the most common reflex test used in most laboratories. The hypothesis however is that this model could be used for many different types of reflex testing in the lab.

# References

American Medical Association. (2021). *Trends in health care spending*. <https://www.ama-assn.org/about/research/trends-health-care-spending>

Berger, D. (1999). A brief history of medical diagnosis and the birth of the clinical laboratory: Part 2–laboratory science and professional certification in the 20th century. *Medical Laboratory Observer: MLO*, *31*(8), 32–34, 36, 38. <https://www.proquest.com/docview/223382876/citation/57E18BE0383F41C4PQ/1>

Boehmke, B., & Greenwell, B. (2020). *Hands-on machine learning with r*. <https://bradleyboehmke.github.io/HOML/>

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). <https://doi.org/10.1214/ss/1009213726>

Cabitza, F., & Banfi, G. (2018). Machine learning in laboratory medicine: waiting for the flood? *Clinical Chemistry and Laboratory Medicine (CCLM)*, *56*(4), 516–524. <https://doi.org/10.1515/cclm-2017-0287>

De Bruyne, S., Speeckaert, M. M., Van Biesen, W., & Delanghe, J. R. (2021). Recent evolutions of machine learning applications in clinical laboratory medicine. *Critical Reviews in Clinical Laboratory Sciences*, *58*(2), 131–152. <https://doi.org/10.1080/10408363.2020.1828811>

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Second edition). O’Reilly Media, Inc.

Henricks, W. H. (2015). Laboratory Information Systems. *Surgical Pathology Clinics*, *8*(2), 101–108. <https://doi.org/10.1016/j.path.2015.02.016>

Hernandez, J. S. (2003). Cost-effectiveness of laboratory testing. *Archives of Pathology & Laboratory Medicine*, *127*(4), 440–445. <https://doi.org/10.5858/2003-127-0440-COLT>

Li, L. T., Huang, T., Bernstam, E. V., & Jiang, X. (2022). External Validation of a Laboratory Prediction Algorithm for the Reduction of Unnecessary Labs in the Critical Care Setting. *The American Journal of Medicine*. <https://doi.org/10.1016/j.amjmed.2021.12.020>

Ma, I., Lau, C. K., Ramdas, Z., Jackson, R., & Naugler, C. (2019). Estimated costs of 51 commonly ordered laboratory tests in Canada. *Clinical Biochemistry*, *65*, 58–60. <https://doi.org/10.1016/j.clinbiochem.2018.12.013>

Murphy, M. J. (2021). Reflex and reflective testing: Progress, but much still to be done. *Annals of Clinical Biochemistry*, *58*(2), 75–77. <https://doi.org/10.1177/0004563221993153>

Park, J. Y., & Kricka, L. J. (2017). One hundred years of clinical laboratory automation: 19672067. *Clinical Biochemistry*, *50*(12), 639–644. <https://doi.org/10.1016/j.clinbiochem.2017.03.004>

Plebani, M., & Giovanella, L. (2020). Reflex TSH strategy: the good, the bad and the ugly. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *58*(1), 1–2. <https://doi.org/10.1515/cclm-2019-0625>

Rabbani, N., Kim, G. Y. E., Suarez, C. J., & Chen, J. H. (2022). Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clinical Biochemistry*, *103*, 1–7. <https://doi.org/10.1016/j.clinbiochem.2022.02.011>

Srivastava, R., Bartlett, W. A., Kennedy, I. M., Hiney, A., Fletcher, C., & Murphy, M. J. (2010). Reflex and reflective testing: efficiency and effectiveness of adding on laboratory tests. *Annals of Clinical Biochemistry*, *47*(3), 223–227. <https://doi.org/10.1258/acb.2010.009282>

Verboeket-van de Venne, W. P. H. G., Aakre, K. M., Watine, J., & Oosterhuis, W. P. (2012). Reflective testing: adding value to laboratory testing. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *50*(7), 1249–1252. <https://doi.org/10.1515/cclm-2011-0611>

Woodmansee, W. W. (2018). Determination of optimal TSH ranges for reflex free T4 testing. *Clinical Thyroidology for the Public*, *11*(2), 3–4. <https://www.thyroid.org/patient-thyroid-information/ct-for-patients/february-2018/vol-11-issue-2-p-3-4/>