

Review

Machine Learning for Patient-Based Real-Time Quality Control (PBRTQC), Analytical and Preanalytical Error Detection in Clinical Laboratory

Nathan Lorde , Shivani Mahapatra and Tejas Kalaria 

Blood Sciences, Black Country Pathology Services, The Royal Wolverhampton NHS Trust,
Wolverhampton WV10 0QP, UK

* Correspondence: nathan.lorde@nhs.net

Abstract: The rapidly evolving field of machine learning (ML), along with artificial intelligence in a broad sense, is revolutionising many areas of healthcare, including laboratory medicine. The amalgamation of the fields of ML and patient-based real-time quality control (PBRTQC) processes could improve the traditional PBRTQC and error detection algorithms in the laboratory. This narrative review discusses published studies on using ML for the detection of systematic errors, non-systematic errors, and combinations of different types of errors in clinical laboratories. The studies discussed used ML for detecting bias, the requirement for re-calibration, samples contaminated with intravenous fluid or EDTA, delayed sample analysis, wrong-blood-in-tube errors, interference or a combination of different types of errors, by comparing the performance of ML models with human validators or traditional PBRTQC algorithms. Advantages, limitations, the creation of standardised ML models, ethical and regulatory aspects and potential future developments have also been discussed in brief.

Keywords: machine learning; artificial intelligence; PBRTQC; quality control; laboratory error; bias



Citation: Lorde, N.; Mahapatra, S.; Kalaria, T. Machine Learning for Patient-Based Real-Time Quality Control (PBRTQC), Analytical and Preanalytical Error Detection in Clinical Laboratory. *Diagnostics* **2024**, *14*, 1808. <https://doi.org/10.3390/diagnostics14161808>

Academic Editor: Jae-Ho Han

Received: 5 July 2024

Revised: 14 August 2024

Accepted: 16 August 2024

Published: 20 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

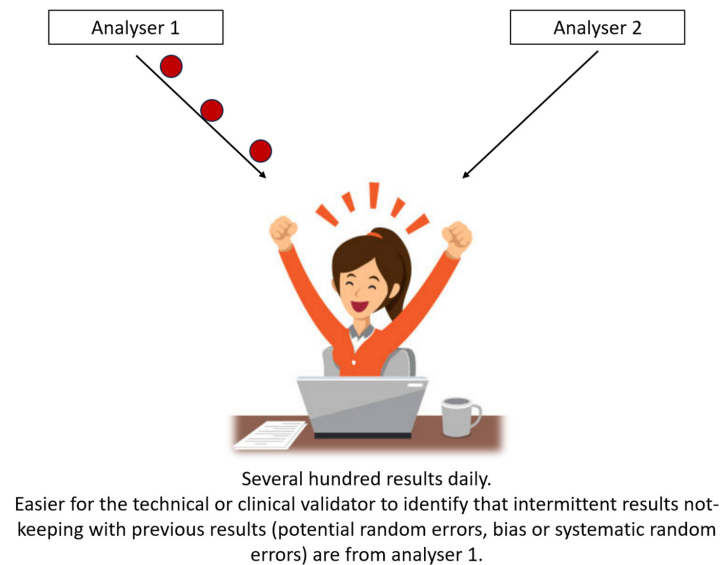
1. Introduction

Clinical laboratories face ever-increasing test demand, with an annual 8% to 10% increase in the workload [1]. In addition, some healthcare systems have opted for consolidated pathology service models to improve efficiency [2]. These changes in demand and operating model lead to large laboratories housing numerous analysers for the same set of tests, operating in parallel. One of the challenges of such transformation, compared to smaller laboratories, is the risk of human technical and clinical validators becoming overwhelmed by the volume and turnover of the data. Though the human brain can hold vast amounts of information, limitations in the amount of information it can process and the speed at which information is processed are well known. With the increasing complexity of input (for example, more parallel data streams, wider test repertoire, wider patient pool), information processing time lengthens (Hick's Law) [3]. The manual result authorisation, as traditionally employed, of large volumes of such complex laboratory data is challenging and could lead to decreased chances of error detection and increased risk of releasing erroneous results, or hinder progress towards improvements in error detection (Figure 1).

All clinical laboratories have internal quality control (IQC) programmes using the measurement of quality control (QC) samples at specified times. However, this approach has several limitations. To list a few: IQC is retrospective rather than real-time, and therefore, a significant change may be detected several hours later. The assumption inherent to the IQC programme that errors are sustained until the next QC run is known not to be the case for all errors. In addition, the cost of obtaining QC material for each test and the occasional lack of commutability of QC materials pose additional challenges [4–6]. The retrospective nature of traditional IQC particularly can result in the need to repeat many or all analyses up to the point of the last satisfactory IQC result, leading to increased

burden on laboratory staff, including the possibility of having to directly contact clinicians if a repeat analysis gives a significantly different result from the one released before. In addition, traditional forms of IQC do not assess other aspects of the total testing process, such as sample transportation issues in the preanalytical phase.

(a)



(b)

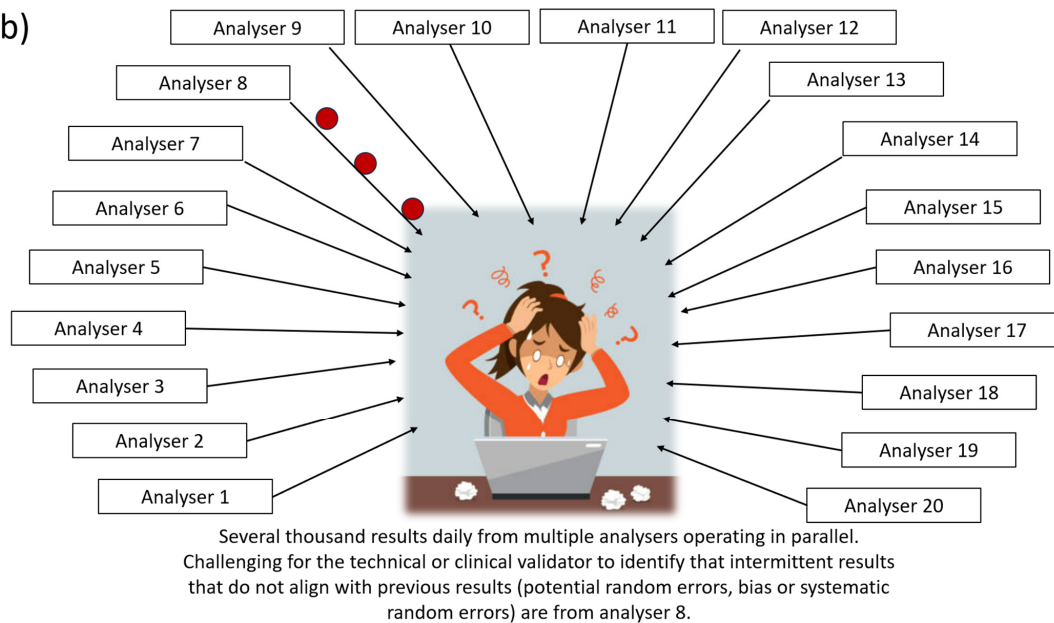


Figure 1. (a) Fewer analysers and fewer samples in a smaller laboratory, making pattern recognition and error detection by human technical or clinical validator easier. (b) Multiple data streams from numerous analysers and large amounts of data present challenges for the human validator in pattern recognition or error detection.

Thoughtfully harnessed patient data could contribute to quality control and pattern recognition. Patient-based real-time quality control (PBRTQC) is a technique that uses the long-term monitoring of patient data through calculations such as moving average (MA), moving standard deviation (MovSD) or moving median (MM) for each analyte. PBRTQC is often superior to periodic IQC in the detection of newly introduced bias or imprecision, and therefore it is particularly useful for analytes that are difficult to keep in

control for prolonged periods of time [4,7,8]. PBRTQC requires more complex modelling than simple IQC charting, as changes in requesting practices over the course of a day or week may impact any real-time metric. PBRTQC is not suitable for every test, such as low throughput, qualitative or semi-quantitative results, and IQC runs will remain important in scenarios such as post-maintenance or post-calibration checks [4,8]. In addition, the simple monitoring of summary statistics, such as MA or MM must allow for sufficiently wide limits, so that the noise from random variation does not lead to unnecessary stoppages and investigations. However, these limits must still be tight enough to detect significant systematic errors [4,8,9]. Another key limitation is that random point errors are usually not detected by most current PBRTQC methods [10].

There has been a big leap forward in the integration of machine learning (ML), a subfield of artificial intelligence (AI), in a variety of industries, including the healthcare sector. ML encompasses a range of algorithms that allow computer systems to perform processes to sort data or use data for prediction in ways not directly programmed by humans [11–14]. The computer programme will “learn” to handle and analyse data and then create an output. There are a vast number of different ML models that can broadly be divided into supervised and unsupervised learning techniques [12,13,15]. Supervised learning models are trained using input data with labelled outputs and allowed to “learn” how to sort the inputs into these pre-defined outputs. The labelled output data are said to “supervise” learning [11–13,15]. Supervised learning algorithms include regression analyses, decision trees (DTs), random forest (RF), support vector machine (SVM), gradient boosting methods, k-nearest neighbours (k-NN), Bayesian networks (BN) and neural networks (NN) [15]. Unsupervised ML models are those that use unlabelled data and sort data into inherent groups using patterns that a human may not be able to see [15]. Examples include principal components analysis (PCA) and k-means clustering [15]. There are also semi-supervised learning models, which lie between supervised and unsupervised learning models. ML models are ever growing and the above list is by no means exhaustive [15,16].

ML models have been used in clinical laboratories for prediction of sample turnaround times [17], creation of reference intervals [18], predicting ferritin concentrations [19,20], predicting test number and machine workload [21], managing reagent stocks [21], predicting a diagnosis of COVID-19 [22–26] and for predicting patients’ most likely diagnoses [27]. ML has also been trialled by many laboratories for PBRTQC and for error detection in recent years. For ease of understanding, in this narrative review we have classified studies employing ML in the field into those identifying systematic, non-systematic or a combination of different types of errors.

2. ML for Detection of Systematic Errors

A relatively early example from 2016 is by Sampson et al. using a logistic regression algorithm to create error scores and then monitoring the cumulative sum of the error score [28]. It was called CUSUM Logistic Regression (CSLR). In this model 14 chemistry analytes were measured, namely sodium, potassium, chloride, urea, creatinine, bicarbonate, alkaline phosphatase (ALP), alanine transaminase (ALT), aspartate transaminase (AST), glucose, albumin, calcium, total protein and total bilirubin, with log-transformation if not normally distributed. Then the result of each analyte was predicted using a stepwise linear multiple regression that incorporated the results of the other 13 analytes in the sample as the independent variables, using 179,280 results collected over three years. The predicted and actual results for each analyte, day of the week and the time of the day were combined in a logistic regression model to generate a probability of error in the assay for that analyte. The logistic regression model was trained using 53,607 results captured from one year. They randomly transformed half of these to simulate biases. The probabilities, minus the mean probability score of “good” data, were added cumulatively over the day (score started at 0 each midnight) and when the score reached a specified threshold, different for each analyte, an investigation of significant bias was triggered. The thresholds were set at levels to create a specificity of 90% for detection of errors. The inclusion of time of day and day of

week in the regression model was to allow for the significant changes in request trends in some analytes. The full CSLR had a lower number of samples tested once a systematic error was intentionally incorporated into the data until detection of the error, compared with the simpler regression which included only the analyte results as independent variables. The full model detected errors in 7 to 80 samples and found 98% of all simulated errors in albumin as an example, compared with 87 to 172 sample run length for the simpler model, which could only detect 61% of the albumin biases. The authors proposed either using the CSLR to alert to the possibility of a bias that can then be investigated with standard IQC testing or of using it to help decide when to run QC material, instead of having to perform IQC at fixed times.

Duan et al., in a study published in 2021, collected patient data for four chemistry analytes: sodium, chloride, ALT and creatinine, initially passed through winsorization, transformed by Box-Cox and then calculated three different PBRTQC algorithms: moving average (MA), exponentially weighted moving average (EWMA) and moving standard deviation (MovSD) [29]. The authors then created what they called Regression-Adjusted Real-Time Quality Control (RARTQC). For this, the respective PBRTQC model had incorporated into it a multiple regression that included patient sex, outpatient versus inpatient status, requesting department information and diagnosis information. Thus, for each analyte, there were three PBRTQC models and three RARTQC models. All six models were tested for the detection of constant and proportional bias, as well as random error. For random error, the simulated error was added to individual results, while for the biases, the entire data streams were altered to simulate an assay drift. Over the course of a year, 79,587 sodium results, 79,588 chloride results, 328,883 ALT results and 418,494 results for creatinine were obtained, with the data from the first nine months being used for training, while the data from the last three months were used for testing. The RARTQC model based on EWMA was the quickest to identify intentionally introduced systematic bias. The authors identified that no model was good at detecting random error. The models were compared using the “trimmed average of number of patients affected before detection” (tNAPed) at different bias levels and plotted against the bias introduced. The total allowable error was 2% for sodium, 5% for chloride, 16% for ALT and 12% for creatinine. The best constant error average tNAPed values were 56.5 by RARTQC EWMA for sodium, 7.5 for chloride by RARTQC MA, 51.5 for ALT by RARTQC EWMA and 56.2 for creatinine by RARTQC EWMA at the total allowable error for each analyte. The tNAPed for the best performing non-RARTQC models were 83.4 by EWMA for sodium, 11.0 by EWMA for chloride, 126.0 for ALT by EWMA and 199.8 for creatinine by EWMA. The RARTQC models similarly outperformed the non-RARTQC models in proportional bias detection. The study demonstrated that ML PBRTQC algorithms require the data to be optimised first in order to achieve good results. The authors trialled different block sizes, which are the numbers of preceding results that are included in the calculation of the moving parameter at any one time, and the truncation limits, which are the limits outside of which the outlier results are excluded. These are called hyperparameters. Most ML and PBRTQC algorithms will have some hyperparameters that require “tuning” to optimise performance.

Zhou et al. developed what they called machine learning internal quality control (MLiQC) [30]. The authors used data from five chemistry analytes and obtained 1,195,000 data points from a Siemens ADVIA® 2400. The analytes were albumin, AST, ALT, glucose and total protein. Of the total dataset, 956,000 data points were used in training and 239,000 were used to test the model for the detection of intentionally introduced biases. The random forest (RF) model was able to detect systematic biases in fewer samples than the four PBRTQC models tested—EWMA, MA, moving median (MM) and the Harrell–Davis median (HD50). The number of results with bias missed before the bias was detected and the accuracy of the MLiQC and PBRTQC models at different bias thresholds were assessed, including the “critical bias” (elsewhere known as the desirable bias limit [31]), calculated as $0.25 \times \sqrt{((CV_i)^2 + (CV_g)^2)}$, where CV_i is the intra-individual coefficient of variation and CV_g is the between-individual coefficient of variation. The critical biases were 1.3% for

albumin, 5.4% for AST, 12% for ALT, 2.5% for glucose and 1.2% for total protein. MLiQC achieved an area under the receiver operating characteristics curve (AUROC) at these biases of 0.9848, 0.9927, 0.9889, 0.9946 and 0.9816, respectively. The average number of patient samples from the point a bias was introduced until it was detected were between 8 and 22 using MLiQC at critical errors, whereas they were from 50 to >1000 for traditional PBRTQC.

Zhou et al. also developed other ML algorithms to augment the PBRTQC procedures. An ML algorithm was created combining RF, a support vector machine (SVM) and an NN to detect deliberately introduced biases in prostate specific antigen (PSA) results [32]. PSA results from 5½ years from one centre were collected, excluding the outlying 1%. The training dataset comprised 43,699 results and the testing dataset 39,700 results. The 43,699 results in the training dataset were augmented by randomly adding a number in the range of -0.3 to $+0.3$ $\mu\text{g/L}$ to each result, doubling the pool of results. This process was repeated for a total of 80 rounds creating a dataset of 3,495,920 results, which served as non-biased data. To each of these 3,495,920 datapoints, either 0.02 $\mu\text{g/L}$ or 0.15 $\mu\text{g/L}$ was added to create two different streams of biased data for training. Using the combined error-free and biased training data points, each of the ML models (SVM, RF and NN) were trained, and five-fold cross-validation was performed to determine the average accuracy of each. The error from the prediction of each individual model was used to determine the individual model's proportion of the sum of the error of all three. This proportion of the error of each model was used to calculate the weight given to the model's score to calculate the overall probability score for the fusion of the models. SVM, RF and NN were weighted 0.34, 0.27 and 0.39, respectively. In the test dataset, they added biases of 0.01, 0.03, 0.05, 0.08, 0.10 and 0.20 $\mu\text{g/L}$. This fusion ML model performed very well against the best performing PBRTQC algorithm, which, in this case, was a moving sum of the outputs (MovSO), at all levels of bias. Taking -0.1 and $+0.1$ $\mu\text{g/L}$ biases, the optimal MovSO algorithm had a median number of patients missed of 157 and 245, versus 9.5 for both biases by the fusion ML model.

In another study by the group, Liang et al. employed an RF model, called machine learning quality control (MLQC), to detect performance shift in seven full blood count parameters, namely lymphocyte count, lymphocyte ratio, haemoglobin, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, red blood cell distribution width and platelet count [33]. Total 423,290 results were obtained over one year, with the first ten months used for training and the last two months for validation. A further 22,460 results from another centre were used for testing. The results below the 25th centile and above the 75th centile were removed, and so were the results from patients with only one sample during the study period, after which intentional biases were introduced. The delta data for the individual patients were recorded. The Isolation Forest (IF) algorithm was used to pre-process the data before passing them to the RF model, an action the authors showed improved the discriminating ability of the MLQC model. MLQC, using delta data, was able to detect biases faster and had superior accuracy compared to the two PBRTQC models (MA and MovSD), which analysed both the numerical results and the delta data. Lymphocyte count, for example, had a median number of patients missed of five before the error was detected by MLQC at the critical bias, versus 72 for the best performing PBRTQC, which was MovSD using delta data.

Liang et al., in another study by the previously described group, used a classification and regression tree (CART) model to detect intentionally introduced biases in 10 measurands, including white cell count, red cell count, haematocrit, haemoglobin, platelet count, AST, ALT, glucose, total protein and albumin [34]. The results from 11 hospitals were collected, with eight used for training and three for testing. In addition to the test results, laboratory identification, department, out- versus in-patient, diagnosis, sex, age, the date of the sample, the time of the sample, the brand of the analyser, the unique identification of the analyser and the units the result was issued in were collected. The training dataset contained 3,097,661 chemistry datapoints and 6,786,110 haematology points. Fifty biases were introduced in 50 steps, from -50% to $+50\%$. This model was called machine-learning

nonlinear regression-adjusted patient-based real-time quality-control (mNL-PBRTQC). The authors compared the model's performance against a basic PBRTQC (EWMA), as well as a linear regression-adjusted real-time quality control (L-RARTQC). The mNL-PBRTQC mostly outperformed the other two in terms of number of samples missed and accurate delineation of biased samples. The CART's outputs served as the input of an EWMA in this model. The best performing L-RARTQC used four non-result variables and the best mNL-PBRTQC used eight. The red blood cell count was the only result where mNL-PBRTQC performed worse compared to the other models. For this parameter, the median number of samples missed was 2460 for mNL-PBRTQC versus 2298 for the plain EWMA and 2415 for L-RARTQC across the range of biases. The median number of patients missed was lower for the mNL-PBRTQC for all other analytes, though at some specific bias levels in some analytes, it had a slightly higher number of missed patient samples compared to the L-RARTQC model.

Regulski et al. recently described different applications for ML in their network of medical laboratories [21]. One of the applications tested involved using a multilayer perceptron (MLP), a class of NN, to detect significant assay drift and therefore the need for re-calibration. The best performing MLP in this study was configured with nine input neurons, one hidden layer with 13 neurons and an output layer with two neurons. It was developed using 16,500 measurement results and 13 calibration points. The MLP performed better than a DT model that was also tested. The nine input neurons used in the MLP were MA, natural logarithms of moving median ($\ln(\text{MME})$), square-root of MA ($\text{Sqrt}(\text{MA})$), average of normals (AoN), natural logarithm of MA ($\ln(\text{MA})$), square root of MME ($\text{sqrt}(\text{MME})$), natural log of the result ($\ln(x).r$) and square-root of the result ($\text{Sqrt}(x).r$). The best performing MLP missed only 1.5% of out-of-control assay drifts that required re-calibration. The authors discussed that this ML model would allow less frequent re-calibrations and thus reduce wastage compared with re-calibrations after a failed internal quality control result. They showed that IQC tests could fail despite there being no error in the patient results being output at the same time. Thus, reliance on IQC alone would result in re-calibration events when they are not needed. This MLP was part of a prototype suite of ML models to optimise laboratory performance, including predicting future request patterns, predicting the need for ordering reagents and predicting the load on various analysers within the laboratories.

In all the above examples of using ML to augment PBRTQC for systematic error detection, except in the studies by Sampson et al. [28] and Regulski et al. [21], the ML augmented algorithms were compared with standard PBRTQC dataflows. In each of these cases, the ML-augmented algorithms were able to detect systematic errors faster than traditional PBRTQC, which, if these results are generalizable, would result in fewer patients' samples being affected by an assay drift before the drift is detected.

3. ML for Detection of Non-Systematic Errors

3.1. Intravenous Fluid Contamination

Baron et al. used DTs to create an algorithm to detect spuriously high glucose from contamination with glucose-containing intravenous fluid [35]. They retrospectively collected 6 months of reported glucose concentrations of >28 mmol/L, as well as the corresponding records for glucose results before and after the high result, the presence of a diagnosis of diabetes mellitus and any record of intravenous fluid or parenteral nutrition around the time of the high result. Expert review of all these factors was used to label samples as contaminated versus not contaminated. One DT was trained using the coexisting analytes collected from the same sample (sodium, potassium, chloride, bicarbonate and anion gap) as well as the patients' glucose results in the preceding year as inputs. This DT revealed that the most significant factors in determining whether a high glucose level was due to intravenous fluid contamination was a glucose result >44 mmol/L and an anion gap ≤ 15 mmol/L. This information was used to create an algorithm for the laboratory staff. Though the DT was not promoted as something to be used "live", the authors stated this

efficient data mining for feature selection was greatly beneficial. Upon implementation, the algorithm identified only 14 out of 43 samples with spurious high glucose because of intravenous fluid contamination, when concentrations >28 mmol/L were evaluated (33% sensitivity, 100% specificity). However, this rose to 14 out of 19 samples with spurious high glucose >44 mmol/L (74% sensitivity).

Spies et al. used an unsupervised approach called Uniform Manifold Approximation and Projection (UMAP) for intravenous fluid contamination detection [36]. The results of sodium, potassium, chloride, blood urea nitrogen, total CO_2 , creatinine, glucose, calcium and calculated anion gap for 2,567,403 samples from 312,721 inpatients were collected. Simulated contaminated samples were created by theoretically adding various proportions of 10 commonly used intravenous fluids to 54,000 patient samples. A training dataset was created by adding these 54,000 simulated contaminated samples to 1,620,275 real-life samples. The remaining collected data were used as the testing dataset. If the UMAP prediction agreed with the interpretation given by the laboratory staff, then it was deemed accurate. For cases of non-concordance, a laboratory expert reviewed the results and the patient information to adjudicate. The area under the precision recall curve was 0.89 for the UMAP prediction. Unlike Baron et al. [35], Spies et al. proposed the real-time deployment of the model.

3.2. Wrong-Blood-in-Tube (WBIT) Errors

Stylewicz and Doctor used data from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial, in which separate samples for glucose and HbA1c were collected from participants [37]. WBIT errors were introduced in 50% of the 8-month samples in the dataset, changing some HbA1c results by $\pm 1.0\%$ and some glucose results by ± 3.89 mmol/L. A probabilistic error detection using BN was built to detect the WBIT events. The BN outperformed the laboratory experts in identifying the simulated errors, with an AUROC of 0.79 compared to an AUROC of up to 0.73 for the experts.

Doctor and Stylewicz used another BN to detect intentional errors in HbA1c and glucose results in data from the National Health and Nutrition Examination Survey (NHANES) and the Diabetes Prevention Program (DPP) studies [38]. The ability of the BN to detect mismatches was compared against a commercially available automated algorithm in the NHANES data and against 11 laboratory experts in the DPP data. The BN outperformed the commercial algorithm and 7 out of 11 experts, and in no case was it inferior.

Rosenbaum and Baron used ML multianalyte delta checks to identify WBIT [39]. The results of 11 analytes for patients, who had to have all 11 analytes tested twice in 36 h, were collected. Some of the more recent samples were switched among patients to simulate WBIT. A logistic regression model and a SVM that had as its inputs the absolute difference between the results, the actual values of the results, the velocity of change of the results and combinations of these were trained and tested. The SVM and logistic regression both had greater AUROC for error detection compared to a method employing only the delta checks of individual analytes. They then deployed the SVM model on biochemistry laboratory results in real time. Of the 217 samples flagged by the SVM, the authors reviewed the electronic health records of 23 patients and concluded that the majority were explainable by patient factors, such as dialysis. Even though this would suggest poor positive predictive value, the authors argued that use of their ML model will flag only those samples at high risk of WBIT, which would be fewer than those flagged by their standard delta checks and would reduce the number of samples that a laboratory medical professional would have to review.

Farrell tested eight different ML models against human performance for the detection of WBIT [40]. Patients with at least two sets of results of six biochemistry analytes in seven days were included, and half had their most recent samples exchanged to simulate WBIT. All eight of these models—an artificial neural network (ANN), extreme gradient boosting (XGB), SVM, RF, logistic regression, k-NN, simple DT and complex DT—outperformed the 50 laboratory professionals in accurately detecting the errors. The ANN performed the best,

with an accuracy of $92.1 \pm 0.44\%$ and the simple DT, which was the lowest performing ML model, had an accuracy of $86.5 \pm 0.25\%$. The laboratory professionals had an accuracy of $77.8 \pm 4.6\%$.

Zhou et al. tested six ML delta check classifiers with different algorithms for WBIT detection [41]. A deep belief network (DBN), RF, SVM, logistic regression, k-NN and a Naïve Bayesian Classifier were assessed in this study. Deidentified haematology results were divided into training and validation datasets, and the results from another laboratory were used as a test dataset. WBIT was simulated by switching the results of haematology tests. The models were compared with one another and with an empirical delta check, a revised weighted delta check and the use of reference change value (RCV)-based delta check. The data input as a group of 22 haematology tests led to a more accurate WBIT error detection compared to the input of individual test results. The DBN, using absolute delta values for 22 of the full blood count parameters, performed the best in the test dataset (accuracy 93.1%, AUROC 0.977).

Mitani et al. collected sample data in their laboratory, in which 11 to 15 of specified biochemistry and haematology analytes were measured, and for which the patient had at least three preceding samples with 11 to 15 of these specified analytes measured as well [42]. Random shuffling of the latest results was carried out to simulate WBIT. A gradient-boosted-decision-tree (GBDT) was trained and then compared with use of delta checks and with an index calculated from the delta data for each analyte called the weighted calculated difference index (wCDI). In the testing dataset, the AUROCs were 0.9984, 0.9378 and 0.9818 for the GBDT ML model, the delta checks and wCDI, respectively. The positive predictive values were 0.4353, 0.0037 and 0.0196, respectively.

3.3. Interference

In a study to assess serum quality, Yang et al. assessed a deep learning convolutional NN model trained on a dataset of centrifuged blood sample images with known serum haemolysis, icteric and lipaemic indices [43]. The image recognition ML model achieved AUROCs of 0.989, 0.996 and 0.993 for sample subclassification, with the analyser-measured haemolysis, icteric and lipaemic indices, respectively, being used as the underlying indicators. The argument for this use case was that the image recognition would subclassify the samples faster than the measurement of indices by the automated analyser, with only negligible reduction in accuracy.

4. ML for Detection of Combinations of Errors Simultaneously

Demirci et al. used an ANN to determine whether samples could be autoverified without human intervention in 2016 [44]. The consensus of seven laboratory specialists was obtained to create rules to verify 13 chemistry analytes: sodium, potassium, chloride, calcium, magnesium, glucose, urea, creatinine, AST, ALT, gamma-glutamyl transferase (GGT), ALP and uric acid. The input parameters were the results of these analytes, as well as delta sodium, delta potassium, delta calcium, delta magnesium, delta glucose, delta uric acid, delta chloride, patient age, the haemolysis index, the lipaemic index and the icteric index. The errors that were being detected were the following: interference by a serum index above threshold for the respective analyte, significant delta change that has no clinical explanation, intravenous fluid contamination, EDTA contamination, delayed sample analysis causing spurious results and results incompatible with life. They first created reference rules for accepting or rejecting results and then selected 1847 patient samples with at least 50% of the 24 inputs required from historically analysed samples to test against these rules. At the optimum input parameters, the ANN was compared initially with the real-time evaluation of 3829 samples by laboratory specialists. The ANN had a sensitivity of only 27.8%, but a specificity of 99.7% for the detection of invalid results. However, when the samples with discordance between the ANN and the human validator were re-analysed by laboratory specialists without time pressure, they found that 127 results that were accepted as valid by the human in real time but marked invalid by the ANN

were indeed invalid, and, similarly, nine samples rejected by the human but marked as valid by the ANN were in fact valid. When these alterations were made, the sensitivity of the ANN for detecting errors causing invalid results was 91%, and the specificity was 100%. A DT was also created in this study. However, the DT did not perform as well, with a sensitivity of 48.9% and a specificity of 99.9% on the same test dataset.

Wang et al. created an ensemble ML model for autoverification [45]. Four base models were trained and tested: a Naïve Bayes Classifier, k-NN, RF and XGBoost; and 52 biochemistry analytes were included. Given the large number of analytes, there was a high number of missing results. They used three different methods of handling this “missingness” with each of the four models: replacing the missing result with the population median of that analyte, replacing it with the mean of the reference interval and thirdly, performing PCA to condense the data prior to introducing them to the ML algorithms. PCA is an unsupervised ML model that summarises the variation in a set of observations to one or more principal components. A principal component will not be a particular analyte, but a kind of summary of the effect of combinations of analytes. This process in effect created 12 different models of ML combined with “missingness” correction. They trained and tested the models initially using a dataset of 36,500 real life samples and then performed external testing on a further dataset of 13,355 samples. The pre-labelling was performed by three laboratory medical practitioners by consensus. Invalid results were either those that needed review of the patient’s clinical scenario and possibly discussion with the clinician looking after the patient or those which were implausible and would require checking for analytical anomalies. In this first round of testing, the authors ranked the methods by their AUROCs and showed that the models based on XGBoost performed the best (AUC 0.972 to 0.98). However, the false negative rates were high (44.1% to 68.2%). The authors believed that the initial training set was too imbalanced, the valid samples vastly outnumbered the invalid ones, and therefore, the models were retrained using more data (61,617 samples) as well as by performing oversampling. Two methods were used for each ML model and “missingness” strategy combination. These were the Synthetic Minority Over-sampling Technique (SMOTE) and the Adaptive Synthetic Sampling Approach (ADASYN). SMOTE works by creating synthetic invalid results by choosing an invalid sample at random and then selecting its k-nearest neighbours and creating the synthetic cases so that their results lie between the randomly sampled point and its neighbours. ADASYN similarly creates synthetic invalid results but instead of selecting the invalid samples at random, it chooses them based on the ones that are the most represented in the invalid class, calculating a density function of the invalid class first in order to do so. The models trained with these modifications were tested on 21,063 samples, and this time, the models with the lowest false negative rates were ranked highest. The three top models this time around all had a false negative rate of 2.743%. These models were the following: XGBoost with the population median used to replace missing values and ADASYN as the oversampling method (AUC 0.982); XGBoost with the reference range mean as the missing imputation value and SMOTE for oversampling (AUC 0.981); and thirdly the RF model with the population median for missing data and ADASYN for oversampling (AUC 0.953). The final ensemble ML averaged the output scores of these three models. The samples with an average score > 50 and ≤ 50 were judged as invalid and valid, respectively. This final ensemble ML model had a false negative rate of 0.411% and an AUROC of 0.998. This ensemble ML model was assessed on real-life results by comparing the model’s efficiency with the laboratory’s current standard practice of using a rules-based autoverification algorithm. The ensemble ML marked fewer samples as potentially invalid. The rules-based system passed 50.2 to 65.1% of samples as valid in each 30 min window, while the ML programme passed 87 to 94% of samples. The authors stated that this ML ensemble would greatly improve laboratory efficiency, as fewer samples would need to be evaluated by a human than is currently the case with the rules-based algorithm. The low false negative rate of 0.411% indicated that this greater efficiency only came with a small risk of missing invalid samples.

5. Standardised ML Model Creation and Ethical Considerations

Many studies have emphasised the importance of data optimisation, including block size, as a preliminary step [29,30,33,34,41]. Significant work is required to identify and optimise the hyperparameters for the ML algorithms. An inherent risk is that of overfitting; that is, tuning the hyperparameters to create optimum performance in one dataset could make the ML model perform less optimally in another dataset [15]. ML models require substantial amounts of data in a suitable format to train. It is important that the data used are representative of the population/environment the model is to be used in. A large, diverse and representative training dataset is of utmost importance, and the use of best practice recommendations and checklists for the ethical design of the ML systems could help to maximise the fairness, equity and robustness of the models [46,47].

The selection of the model type, the construction of the data set, data sampling and data preprocessing are some of the factors that could have significant impacts on how the model performs. For example, Wang et al. demonstrated improvement in their models' performance when data imbalance was addressed by oversampling [45]. ML algorithms may sometimes be very sensitive to missing data, causing an inaccurate output [15,45]. The robustness of ML algorithms to "missingness" typically requires studies on a case-by-case basis, with different strategies to replace or remove observations with missing values tested [15]. Biases that exist when datasets are created may become embedded in the model, leading to the model giving incorrect outputs. A systematic review highlighted that a group of ML models created to detect COVID-19 by analysing chest X-rays and CT scans, which relied on a publicly available dataset of paediatric radiographs as negative comparators, led to models that became efficient at differentiating adult from paediatric images, instead of differentiating COVID-19 patients from those with other causes of pneumonia [48,49]. ML models, like any other device or tool, are not immune to errors or limitations of experimental design, data selection, data processing or model selection and execution. Therefore, several groups have produced recommendations for a reproducible and reliable ML prediction model development in healthcare, including for clinical laboratories [46,50–54].

The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) has created a working group to address potential pitfalls in the field and, in 2023, the group published recommendations for best practices for the use of ML in laboratory medicine [46]. These recommendations are aimed at improving the quality of published ML models and to ensure that the models are appropriately used, and their outputs are valid, reproducible and reliable. For results to be reproducible, studies should include detailed methodology on the ML model and the specific hardware and software used, as well as the training data sets. The recommendations, however, did not cover how to integrate the ML models into workflows or how the models should be regulated.

There exists a danger that creators view ML as being inherently better than current approaches and seek to incorporate its use even though it may not be accurate. The IFCC working group suggested that laboratories consider whether rules-based approaches or algorithms may be better suited for a given problem before attempting to introduce ML [6].

Additionally, some models function as black-box models, and it is not possible to decipher how the model outputs are determined. If such a model produces erroneous outputs, this lack of explainability could impede detection [55]. A companion concern is of liability for unsatisfactory outcomes because of an erroneous output from a model [56]. Laboratories will have to know the deficiencies in the model and implement measures to mitigate them. Laboratory professionals using the ML model as part of their role in QC or result validation may not have expertise in statistical modelling and thus may not easily identify issues with the model. Two recent studies have highlighted that while healthcare professionals understand the potential of AI to revolutionise the healthcare sector, many fear the potential negative impacts [57,58]. These studies have also shown that healthcare professionals have other ethical concerns, including data security [57,58]. There also exists the difficulty of diverting funding from more pressing needs in the healthcare budget.

6. Regulation

The development of AI has been largely unregulated so far. Many countries, including the UK [59] and the EU [60], are now establishing frameworks. The British Medicines and Healthcare products Regulatory Agency (MHRA) policy paper from April 2024 considers the use of AI in healthcare as being within the MHRA's remit for regulation and will be providing further guidance [61]. The Food and Drug Administration (FDA) of the USA has published a commitment to increase its capacity to understand and regulate the use of AI and ML in the delivery of patient care [62].

7. Future

ML models have shown higher accuracy at detecting errors and patterns compared to human validators in studies. ML and AI may free up time for laboratory scientists to work on other aspects of the role and achieve cost, workflow and productivity efficiencies. However, as can be seen from the studies discussed, so far, most ML models have not been tested on real-life, real-time patient data. Therefore, studies demonstrating reproducible real-life, real-time use are required before wider use in laboratory quality control and error detection can be recommended. The development of best practice recommendations and checklists for creating reliable and reproducible ML models is a welcome step, and so are the evolving regulatory aspects.

ML, as well as AI more broadly, is rapidly advancing, with evolving applications. With the increasing availability of higher computing power, sophisticated ML algorithms can be applied to the analysis of more complex data [63,64]. We anticipate increasingly advanced ML models incorporating many different error detection scenarios discussed in this review and perhaps some not even trialled or envisioned yet. Considering the developments so far and the pace of the progress in the field, it would not be surprising to envisage the ML models of the near future contributing to several aspects of a laboratory's analytical quality, such as the detection of bias, random errors, flagging potential interference and contaminated samples, samples analysed with a delay, wrong-blood-in-tube errors and perhaps a few more not yet thought or trialled aspects (Figure 2). Additionally, we anticipate the incorporation of ML capabilities into commercially available productivity packages, laboratory information management systems and middleware, which may facilitate widespread acceptance and adoption.

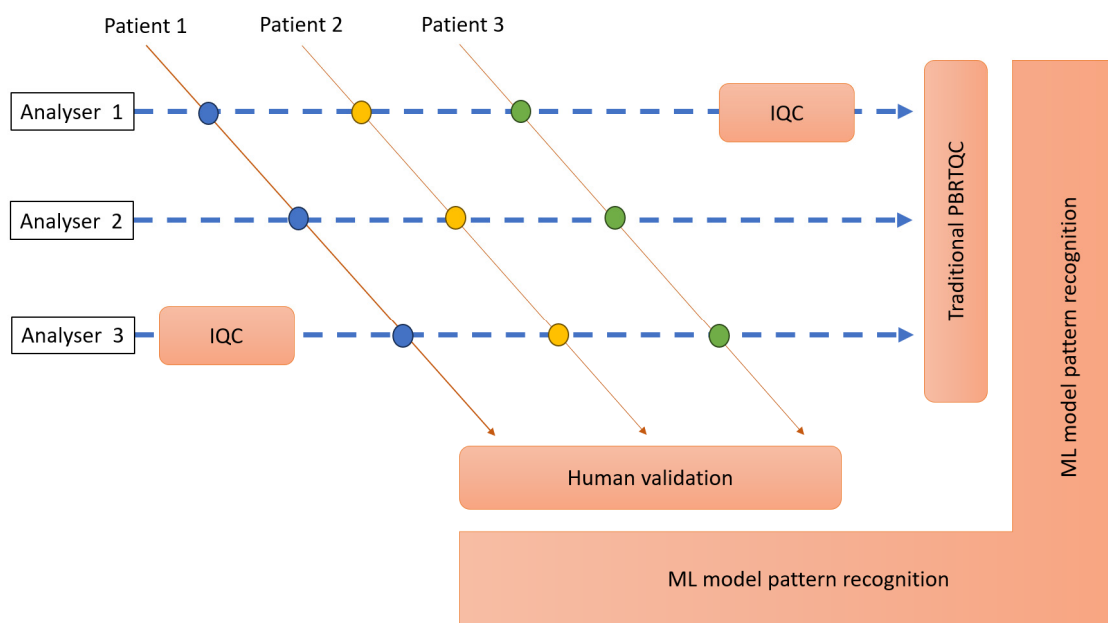


Figure 2. Contribution to analytical quality by traditional IQC methods, result validation by humans, traditional PBRTQC, and potential future ML model-based pattern recognition.

8. Conclusions

The examples from the literature demonstrate the capability of ML in error detection and QC in medical laboratories. Because the ML studies are addressing different questions and are carried out in different circumstances, using different methods, the direct comparison of models is not possible. We remain optimistic that laboratory professionals will be using such approaches in the future, hopefully with great benefit to the service offered to patients and clinicians. The creation of regulatory frameworks, working groups and best practice recommendations in the field are welcome moves. However, considering that we are in the early days of this rapidly evolving field, and many, if not most, ML models described have only been tested in studies rather than on real-time patient data, due diligence, focusing on safe, standardised, ethical and regulation-compliant use of ML will be essential.

Author Contributions: Conceptualization, T.K.; review of literature, N.L., S.M. and T.K.; writing—original draft preparation, N.L. and S.M.; writing—review and editing, N.L. and T.K.; visualization, T.K.; supervision, T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

ML	machine learning
PBRTQC	patient-based real-time quality control
EDTA	ethylenediaminetetraacetic acid
IQC	internal quality control
QC	quality control
MA	moving average
MovSD	moving standard deviation
MM	moving median
AI	artificial intelligence
DT	decision tree
RF	random forest
SVM	support vector machine
k-NN	k-nearest neighbour
BN	Bayesian network
NN	neural network
ANN	artificial neural network
PCA	principal components analysis
CUSUM	cumulative sum
CSLR	CUSUM logistic regression
ALP	alkaline phosphatase
ALT	alanine transaminase
AST	aspartate transaminase
EWMA	exponentially weighted moving average
RARTQC	regression-adjusted real-time quality control
tNAPed	trimmed average of number of patients affected before detection
MLiQC	machine learning internal quality control
HD50	Harell-Davis median
CVi	intra-individual coefficient of variation
CVg	between-individual coefficient of variation
AUC	area under curve
AUROC	area under the receiver operating characteristics curve

PSA	prostate specific antigen
MovSO	moving sum of the outputs
MLQC	machine learning quality control
IF	isolation forest
CART	classification and regression tree
mNL-PBRTQC	machine-learning non-linear regression-adjusted patient-based real-time quality control
L-RARTQC	linear regression-adjusted real-time quality control
MLP	multilayer perceptron
ln(MME)	natural logarithm of moving median
Sqrt(MA)	square-root of moving average
AoN	average of normals
ln(MA)	natural logarithm of moving average
Sqrt(MME)	square-root of moving median
ln(x).r	natural logarithm of result
Sqrt(x).r	square-root of result
UMAP	uniform manifold approximation and projection
WBIT	wrong blood in tube
ACCORD trial	Action to Control Cardiovascular Risk in Diabetes trial
NHANES	National Health and Nutrition Examination Survey
DPP	Diabetes Prevention Program
XGB/XGBoost	extreme gradient boosting
DBN	deep belief network
RCV	reference change value
GBDT	gradient boosted decision tree
wCDI	weighted calculated difference index
GGT	gamma-glutamyl transferase
SMOTE	synthetic minority over-sampling technique
ADASYN	adaptive synthetic sampling approach
IFCC	International Federation of Clinical Chemistry and Laboratory Medicine
MHRA	Medicines and Healthcare products Regulatory Agency
FDA	Food and Drug Administration

References

1. Fryer, A.A.; Smellie, W.S.A. Managing demand for laboratory tests: A laboratory toolkit. *J. Clin. Pathol.* **2013**, *66*, 62–72. [\[CrossRef\]](#)
2. Satta, G.; Edmonstone, J. Consolidation of pathology services in England: Have savings been achieved? *BMC Health Serv. Res.* **2018**, *18*, 862. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Trazzi, M.; Yampolskiy, R.V. Artificial Stupidity: Data We Need to Make Machines Our Equals. *Patterns* **2020**, *1*, 100021. [\[CrossRef\]](#) [\[PubMed\]](#)
4. van Rossum, H.H.; Bietenbeck, A.; Cervinski, M.A.; Katayev, A.; Loh, T.P.; Badrick, T.C. Benefits, limitations and controversies on patient-based real-time quality control (PBRTQC) and the evidence behind the practice. *Clin. Chem. Lab. Med.* **2021**, *59*, 1213–1220. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Loh, T.P.; Cervinski, M.A.; Katayev, A.; Bietenbeck, A.; van Rossum, H.; Badrick, T. Recommendations for laboratory informatics specifications needed for the application of patient-based real time quality control. *Clin. Chim. Acta* **2019**, *495*, 625–629. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Bietenbeck, A.; Cervinski, M.A.; Katayev, A.; Loh, T.P.; van Rossum, H.H. Understanding Patient-Based Real-Time Quality Control Using Simulation Modeling. *Clin. Chem.* **2020**, *66*, 1072–1083. [\[CrossRef\]](#)
7. Hoffmann, R.G.; Waid, M.E. The “Average of Normals” Method of Quality Control. *Am. J. Clin. Pathol.* **1965**, *43*, 134–141. [\[CrossRef\]](#)
8. Van Rossum, H.H.; Van Den Broek, D. Design and implementation of quality control plans that integrate moving average and internal quality control: Incorporating the best of both worlds. *Clin. Chem. Lab. Med.* **2019**, *57*, 1329–1338. [\[CrossRef\]](#)
9. Loh, T.P.; Bietenbeck, A.; Cervinski, M.A.; van Rossum, H.H.; Katayev, A.; Badrick, T. Recommendation for performance verification of patient-based real-time quality control. *Clin. Chem. Lab. Med.* **2020**, *58*, 1205–1213. [\[CrossRef\]](#)
10. Duan, X.; Wang, B.; Zhu, J.; Shao, W.; Wang, H.; Shen, J.; Wu, W.; Jiang, W.; Yiu, K.L.; Pan, B.; et al. Assessment of patient-based real-time quality control algorithm performance on different types of analytical error. *Clin. Chim. Acta* **2020**, *511*, 329–335. [\[CrossRef\]](#)

11. Punchoo, R.; Bhoora, S.; Pillay, N. Applications of machine learning in the chemical pathology laboratory. *J. Clin. Pathol.* **2021**, *74*, 435–442. [\[CrossRef\]](#)
12. Bunch, D.R.; Durant, T.J.; Rudolf, J.W. Artificial Intelligence Applications in Clinical Chemistry. *Clin. Lab. Med.* **2023**, *43*, 47–69. [\[CrossRef\]](#)
13. De Bruyne, S.; Speeckaert, M.M.; Van Biesen, W.; Delanghe, J.R. Recent evolutions of machine learning applications in clinical laboratory medicine. *Crit. Rev. Clin. Lab. Sci.* **2021**, *58*, 131–152. [\[CrossRef\]](#)
14. Rabbani, N.; Kim, G.Y.E.; Suarez, C.J.; Chen, J.H. Applications of Machine Learning in Routine Laboratory Medicine: Current State and Future Directions. *Clin. Biochem.* **2022**, *103*, 1–7. [\[CrossRef\]](#) [\[PubMed\]](#)
15. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2022.
16. Kubat, M. *An Introduction to Machine Learning*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2017. [\[CrossRef\]](#)
17. Tsai, E.R.; Demirtas, D.; Hoogendijk, N.; Tintu, A.N.; Boucherie, R.J. Turnaround time prediction for clinical chemistry samples using machine learning. *Clin. Chem. Lab. Med.* **2022**, *60*, 1902–1910. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Poole, S.; Schroeder, L.F.; Shah, N. An unsupervised learning method to identify reference intervals from a clinical database. *J. Biomed. Inf. Inform.* **2016**, *59*, 276–284. [\[CrossRef\]](#)
19. Luo, Y.; Szolovits, P.; Dighe, A.S.; Baron, J.M. Using Machine Learning to Predict Laboratory Test Results. *Am. J. Clin. Pathol.* **2016**, *145*, 778–788. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Kurstjens, S.; De Bel, T.; Van Der Horst, A.; Kusters, R.; Krabbe, J.; van Balveren, J. Automated prediction of low ferritin concentrations using a machine learning algorithm. *Clin. Chem. Lab. Med.* **2022**, *60*, 1921–1928. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Regulski, K.; Opaliński, A.; Swadźba, J.; Sitkowski, P.; Wąsowicz, P.; Kwietniewska-Śmietana, A. Machine Learning Prediction Techniques in the Optimization of Diagnostic Laboratories' Network Operations. *Appl. Sci.* **2024**, *14*, 2429. [\[CrossRef\]](#)
22. Baik, S.M.; Hong, K.S.; Park, D.J. Application and utility of boosting machine learning model based on laboratory test in the differential diagnosis of non-COVID-19 pneumonia and COVID-19. *Clin. Biochem.* **2023**, *118*, 110584. [\[CrossRef\]](#)
23. Cabitza, F.; Campagner, A.; Ferrari, D.; Di Resta, C.; Ceriotti, D.; Sabetta, E.; Colombini, A.; De Vecchi, E.; Banfi, G.; Locatelli, M.; et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clin. Chem. Lab. Med.* **2021**, *59*, 421–431. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Çubukçu, H.C.; Topcu, D.I.; Bayraktar, N.; Gülşen, M.; Sarı, N.; Arslan, A.H. Detection of COVID-19 by Machine Learning Using Routine Laboratory Tests. *Am. J. Clin. Pathol.* **2022**, *157*, 758–766. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Tschoellitsch, T.; Dünser, M.; Böck, C.; Schwarzbauer, K.; Meier, J. Machine Learning Prediction of SARS-CoV-2 Polymerase Chain Reaction Results with Routine Blood Tests. *Lab. Med.* **2021**, *52*, 146–149. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Schwab, P.; Schütte, A.D.M.; Dietz, B.; Bauer, S. Clinical Predictive Models for COVID-19: Systematic Study. *J. Med. Internet Res.* **2020**, *22*, e21439. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Park, D.J.; Park, M.W.; Lee, H.; Kim, Y.; Kim, Y.; Park, Y.H. Development of machine learning model for diagnostic disease prediction based on laboratory tests. *Sci. Rep.* **2021**, *11*, 7567. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Sampson, M.L.; Gounden, V.; van Deventer, H.E.; Remaley, A.T. CUSUM-Logistic Regression analysis for the rapid detection of errors in clinical laboratory test results. *Clin. Biochem.* **2016**, *49*, 201–207. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Duan, X.; Wang, B.; Zhu, J.; Zhang, C.; Jiang, W.; Zhou, J.; Shao, W.; Zhao, Y.; Yu, Q.; Lei, L.; et al. Regression-Adjusted Real-Time Quality Control. *Clin. Chem.* **2021**, *67*, 1342–1350. [\[CrossRef\]](#)
30. Zhou, R.; Wang, W.; Padoan, A.; Wang, Z.; Feng, X.; Han, Z.; Chen, C.; Liang, Y.; Wang, T.; Cui, W.; et al. Traceable machine learning real-time quality control based on patient data. *Clin. Chem. Lab. Med.* **2022**, *60*, 1998–2004. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Fraser, C.G.; Petersen, P.H.; Libeer, J.; Ricos, C. Proposals for setting generally applicable quality goals based on biology. *Ann. Clin. Biochem.* **1997**, *34*, 8–12. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Zhou, R.; Liang, Y.; Cheng, H.; Padoan, A.; Wang, Z.; Feng, X.; Han, Z.; Song, B.; Wang, W.; Plebani, M.; et al. A multi-model fusion algorithm as a real-time quality control tool for small shift detection. *Comput. Biol. Med.* **2022**, *148*, 105866. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Liang, Y.; Wang, Z.; Huang, D.; Wang, W.; Feng, X.; Han, Z.; Song, B.; Wang, Q.; Zhou, R. A study on quality control using delta data with machine learning technique. *Heliyon* **2022**, *8*, e09935. [\[CrossRef\]](#)
34. Liang, Y.F.; Padoan, A.; Wang, Z.; Chen, C.; Wang, Q.; Plebani, M.; Zhou, R. Machine learning-based nonlinear regression-adjusted real-time quality control modeling: A multi-center study. *Clin. Chem. Lab. Med.* **2024**, *62*, 635–645. [\[CrossRef\]](#)
35. Baron, J.M.; Mermel, C.H.; Lewandrowski, K.B.; Dighe, A.S. Detection of Preanalytic Laboratory Testing Errors Using a Statistically Guided Protocol. *Am. J. Clin. Pathol.* **2012**, *138*, 406–413. [\[CrossRef\]](#)
36. Spies, N.C.; Hubler, Z.; Azimi, V.; Zhang, R.; Jackups Jr, R.; Gronowski, A.R.; Farnsworth, C.W.; Zaydman, M.A. Automating the Detection of IV Fluid Contamination Using Unsupervised Machine Learning. *Clin. Chem.* **2024**, *70*, 444–452. [\[CrossRef\]](#)
37. Strylewicz, G.; Doctor, J. Evaluation of an automated method to assist with error detection in the ACCORD central laboratory. *Clin. Trials* **2010**, *7*, 380–389. [\[CrossRef\]](#)
38. Doctor, J.N.; Strylewicz, G. Detecting 'wrong blood in tube' errors: Evaluation of a Bayesian network approach. *Artif. Intell. Med.* **2010**, *50*, 75–82. [\[CrossRef\]](#) [\[PubMed\]](#)

39. Rosenbaum, M.W.; Baron, J.M. Using Machine Learning-Based Multianalyte Delta Checks to Detect Wrong Blood in Tube Errors. *Am. J. Clin. Pathol.* **2018**, *150*, 555–566. [\[CrossRef\]](#)
40. Farrell, C.J. Identifying mislabelled samples: Machine learning models exceed human performance. *Ann. Clin. Biochem.* **2021**, *58*, 650–652. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Zhou, R.; Liang, Y.F.; Cheng, H.L.; Wang, W.; Huang, D.W.; Wang, Z.; Feng, X.; Han, Z.W.; Song, B.; Padoan, A.; et al. A highly accurate delta check method using deep learning for detection of sample mix-up in the clinical laboratory. *Clin. Chem. Lab. Med.* **2022**, *60*, 1984–1992. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Mitani, T.; Doi, S.; Yokota, S.; Imai, T.; Ohe, K. Highly accurate and explainable detection of specimen mix-up using a machine learning model. *Clin. Chem. Lab. Med.* **2020**, *58*, 375–383. [\[CrossRef\]](#)
43. Yang, C.; Li, D.; Sun, D.; Zhang, S.; Zhang, P.; Xiong, Y.; Zhao, M.; Qi, T.; Situ, B.; Zheng, L. A deep learning-based system for assessment of serum quality using sample images. *Clin. Chim. Acta* **2022**, *531*, 254–260. [\[CrossRef\]](#)
44. Demirci, F.; Akan, P.; Kume, T.; Sisman, A.R.; Erbayraktar, Z.; Sevinc, S. Artificial Neural Network Approach in Laboratory Test Reporting: Learning Algorithms. *Am. J. Clin. Pathol.* **2016**, *146*, 227–237. [\[CrossRef\]](#)
45. Wang, H.; Wang, H.; Zhang, J.; Li, X.; Sun, C.; Zhang, Y. Using machine learning to develop an autoverification system in a clinical biochemistry laboratory. *Clin. Chem. Lab. Med.* **2020**, *59*, 883–891. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Master, S.R.; Badrick, T.C.; Bietenbeck, A.; Haymond, S. Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group. *Clin. Chem.* **2023**, *69*, 690–698. [\[CrossRef\]](#)
47. Chen, I.Y.; Pierson, E.; Rose, S.; Joshi, S.; Ferryman, K.; Ghassemi, M. Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 123–144. [\[CrossRef\]](#)
48. Roberts, M.; Driggs, D.; Thorpe, M.; Gilbey, J.; Yeung, M.; Ursprung, S.; Aviles-Rivero, A.I.; Etmann, C.; McCague, C.; Beer, L.; et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **2021**, *3*, 199–217. [\[CrossRef\]](#)
49. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Haymond, S.; Master, S.R. How Can We Ensure Reproducibility and Clinical Translation of Machine Learning Applications in Laboratory Medicine? *Clin. Chem.* **2022**, *68*, 392–395. [\[CrossRef\]](#)
51. Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E.; Larochelle, H. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *J. Mach. Learn. Res.* **2020**, *22*, 7459–7478. [\[CrossRef\]](#)
52. Norgéot, B.; Quer, G.; Beaulieu-Jones, B.K.; Torkamani, A.; Dias, R.; Gianfrancesco, M.; Arnaout, R.; Kohane, I.S.; Saria, S.; Topol, E.; et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nat. Med.* **2020**, *26*, 1320–1324. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Hernandez-Boussard, T.; Bozkurt, S.; Ioannidis, J.P.A.; Shah, N.H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *J. Am. Med. Inf. Assoc.* **2020**, *27*, 2011–2015. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Stevens, L.M.; Mortazavi, B.J.; Deo, R.C.; Curtis, L.; Kao, D.P. Recommendations for reporting machine learning analyses in clinical research. *Circ. Cardiovasc. Qual. Outcomes* **2020**, *13*, e006556. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Samek, W.; Müller, K.-R. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.B., Eds.; Springer: Cham, Switzerland, 2019; pp. 5–22.
56. Maliha, G.; Gerke, S.; Cohen, I.G.; Parikh, R.B. Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *Milbank Q.* **2021**, *99*, 629–647. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Castagno, S.; Khalifa, M. Perceptions of Artificial Intelligence Among Healthcare Staff: A Qualitative Survey Study. *Front. Artif. Intell.* **2020**, *3*, 578983. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Lai, M.C.; Brian, M.; Mamzer, M.F. Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France. *J. Transl. Med.* **2020**, *18*, 14. [\[CrossRef\]](#) [\[PubMed\]](#)
59. UK Department for Science, Innovation and Technology. A Pro-Innovation Approach to AI Regulation: Government Response. 2024. Available online: <https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response> (accessed on 16 May 2024).
60. European Union. Artificial Intelligence Act. 2021/0106(COD). Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html (accessed on 16 May 2024).
61. Medicines & Healthcare Products Regulatory Agency. Impact of AI on the Regulation of Medical Products. Implementing the AI White Paper Principles. 2024. Available online: <https://www.gov.uk/government/publications/impact-of-ai-on-the-regulation-of-medical-products> (accessed on 5 June 2024).
62. Food and Drug Administration (FDA). Artificial Intelligence and Medical Products: How CBER, CDER, CDRH, and OCP Are Working Together. 2024. Available online: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> (accessed on 15 May 2024).

63. Lohn, A.; Musser, M. *AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress?* Center for Security and Emerging Technology: Washington, DC, USA, 2022; pp. 1–11. [[CrossRef](#)]
64. Ghahramani, Z.; Daley, S.; Sayed, S.; Spittle, G.; Trefethen, A. Independent Review of the Future of Compute: Final Report and Recommendations. 2023. Available online: <https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts> (accessed on 15 May 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.