Steef Kurstjens*, Thomas de Bel, Armando van der Horst, Ron Kusters, Johannes Krabbe and Jasmijn van Balveren

Automated prediction of low ferritin concentrations using a machine learning algorithm

Abstract

Objectives: Computational algorithms for the interpretation of laboratory test results can support physicians and specialists in laboratory medicine. The aim of this study was to develop, implement and evaluate a machine learning algorithm that automatically assesses the risk of low body iron storage, reflected by low ferritin plasma levels, in anemic primary care patients using a minimal set of basic laboratory tests, namely complete blood count and C-reactive protein (CRP).

Methods: Laboratory measurements of anemic primary care patients were used to develop and validate a machine learning algorithm. The performance of the algorithm was compared to twelve specialists in laboratory medicine from three large teaching hospitals, who predicted if patients with anemia have low ferritin levels based on laboratory test reports (complete blood count and CRP). In a second round of assessments the algorithm outcome was provided to the specialists in laboratory medicine as a decision support tool.

Results: Two separate algorithms to predict low ferritin concentrations were developed based on two different chemistry analyzers, with an area under the curve of the ROC of 0.92 (Siemens) and 0.90 (Roche). The specialists in laboratory medicine were less accurate in predicting low ferritin concentrations compared to the algorithms, even when knowing the output of the algorithms as support tool. Implementation of the algorithm in the laboratory system resulted in one new iron deficiency diagnosis on average per day.

Conclusions: Low ferritin levels in anemic patients can be accurately predicted using a machine learning algorithm based on routine laboratory test results. Moreover, implementation of the algorithm in the laboratory system reduces the number of otherwise unrecognized iron deficiencies.

Keywords: artificial intelligence; case-finding; diagnostics; iron deficiency; laboratory information system; reflective testing.

*Corresponding author: Steef Kurstjens, PhD, Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Hospital, Henri Dunantstraat 1, 5223 GZ, P.O. Box 90153, 5200, 's Hertogenbosch, the Netherlands, Phone: +031(0) 0626281521, E-mail: s.kurstjens@jbz.nl
Thomas de Bel, Diagnostic Image Analysis Group, Radboudumc, Nijmegen, the Netherlands
Armando van der Horst, Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Hospital, 's Hertogenbosch, the Netherlands
Ron Kusters, Laboratory of Clinical Chemistry and Hematology, Jeroen Bosch Hospital, 's Hertogenbosch, the Netherlands; and Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands
Johannes Krabbe, Laboratory of Clinical Chemistry and Laboratory Medicine, Medisch Spectrum Twente, Enschede, the Netherlands; and Laboratory of Clinical Chemistry and Laboratory Medicine, Medlon BV, Enschede, the Netherlands
Jasmijn van Balveren, Department of Health Technology and Services Research, Technical Medical Centre, University of Twente, Enschede, the Netherlands; and Laboratory of Clinical Chemistry and Hematology, St Jansdal, Harderwijk, the Netherlands

Introduction

Iron-deficiency is the most common cause of anemia, with over one billion cases worldwide in 2016 [1]. Unfortunately, iron deficiency still remains underdiagnosed [2]. Bone marrow iron staining (Perl's stain) is considered the gold standard in diagnosis of iron-deficient anemia, but is invasive, expensive and time-consuming [3]. In practice, the plasma ferritin concentration reflects the body's iron status, and is therefore most frequently used to diagnose iron deficiency [4]. However, there is no consensus on the optimal reference limit for ferritin, resulting in many different lower reference limits [5]. Moreover, the ferritin assay is poorly harmonized, and considerable inter-analyzer differences have to be taken into account when setting reference intervals and interpreting concentrations [6].

In a state of iron depletion erythropoiesis is disrupted, and newly produced red blood cells are usually microcytic (resulting in a mean corpuscular volume (MCV) <80 fL) and hypochromic (a mean corpuscular hemoglobin (MCH) of <1.7 fmol). Importantly, the information contained in these changing patterns of laboratory parameters is used to identify possible iron deficiencies. However, subtle changes in patterns across a multitude of laboratory results can be overlooked by physicians. In many anemic primary care patients ferritin is not included in the laboratory order, resulting in underdiagnosis of iron deficiency. Importantly, specialists in laboratory medicine can provide assistance to physicians in diagnostics of anemia in the form of a 'laboratory anemia workup' or 'reflective testing' [7]. Computational assistance in the interpretation of laboratory results can provide a key automated diagnostic support tool for physicians as well as specialists in laboratory medicine, reducing the number of overlooked and/or undiagnosed iron deficiencies.

The number of machine learning diagnostic tools is rapidly increasing, mainly in the fields of radiology and pathology. The diagnostic performance of such tools is impressive, and they often outperform trained specialists [8, 9]. Recently, during the COVID-19 pandemic, machine learning algorithms to predict COVID-19 disease based on clinical chemistry parameters have been developed and widely used, showing the importance and potential of this development [10–12]. However, broad implementation of machine learning tools in clinical practice is still limited [13]. In this article we aimed to develop and implement a machine learning algorithm that automatically assesses the risk of low ferritin levels using a minimal set of basic laboratory tests, namely complete blood count and C-reactive protein (CRP). Moreover, we set out to investigate the benefits of implementing this tool in our laboratories.

Materials and methods

Data collection

Laboratory reports requested by general practitioners for anemic adult patients (Hb <7.5 mmol/L for females and <8.5 mmol/L for males) were selected based on a complete blood count, CRP and ferritin concentration. Data were received from three hospital laboratories (Jeroen Bosch Hospital, 's-Hertogenbosch (n=3,797), Medlon BV Enschede (n=8,021) and St Jansdal Hospital, Harderwijk (n=191)). Exclusion criteria were missing data (invalid measurement of blood count or CRP, <1% of cases), duplicate patients and age under 18 years. At the Jeroen Bosch Hospital, vitamin B12 and folic acid concentrations were included when available.

Measurements

At the Jeroen Bosch Hospital, CRP was measured on the Advia Chemistry XPT (Siemens Healthineers, Erlangen, Germany), ferritin and vitamin B12 were measured on the Advia Centaur XPT (Siemens Healthineers), folic acid was measured on the Immulite 2000 XPi (Siemens Healthineers), and hematological parameters were measured on the Advia 2120i (Siemens Healthineers). At the St Jansdal Hospital, CRP and ferritin were measured on the Atellica Analyzer (Siemens Healthineers), and hematological parameters were measured on the Sysmex XN-9000 (Sysmex Corporation, Hyogo, Japan). At Medlon BV, CRP and ferritin were measured with the C702 and e801 module (COBAS 8000 routine chemistry analyzer, Roche Diagnostics, Mannheim, Germany) respectively, and hematological parameters were measured on the XN-9000 hematology analyzer (Sysmex Corporation).

Machine learning algorithm

Laboratory measurements of 2,657 patients (70% of the dataset), referred for blood analyses by their primary care physicians, were used to develop and test the machine learning algorithm of the Jeroen Bosch Hospital (using a Siemens clinical chemistry analyzer), referred to as the JBH-S algorithm. The model was validated on a separate dataset of 1,140 measurements (Table 1, 30% of the dataset). The JBH-S model was additionally validated using a dataset from the St Jansdal Hospital (191 patients, Table 1), using identical clinical chemistry platform but Sysmex for hematology. To develop the model for Medlon BV (using a Roche clinical chemistry analyzer), referred to as the Medlon-R model, 6,417 patients (80% of the dataset) were used to develop and test the model and 1,604 patients (20% of the dataset) were used for validation (Table 1).

A random forest classifier model was used to fit the data. We chose the random forest because it can handle combinations of binary, categorical and numerical features. Furthermore, random forest models require little hyperparameter tuning and can deal well with imbalanced datasets. The model was trained using the standard parameters of the scikit-learn package, using 100 trees. To account for the imbalance between positive and negative cases, balanced subsampling was used, which adjusts the weights inversely proportional to the frequencies of the input classes for each tree. Analyses were carried out using Python 3.6; with the packages NumPy (1.16.2), pandas (0.24.2), scikit-learn (0.20.3). The machine learning algorithm was trained to maximize the AUC of the ROC, using a lower reference limit for the ferritin concentration of <10 μg/L for women and <22 μg/L for men (Siemens), and <13 μg/L for women and <30 μg/L for men (Roche). The machine learning algorithms for vitamin B12 and folic acid were trained using a lower reference limit of <156 pmol/L for vitamin B12 and <6 nmol/L for folic acid.

Folic acid measurements >54 nmol/L were assigned the value 55 nmol/L, vitamin B12 values >1,476 pmol/L were assigned the value 1,477, and CRP values <3 mg/L were assigned the value 2 mg/L.

Comparison of the performance of machine learning algorithm vs. specialists in laboratory medicine

A dataset of 336 independent laboratory reports (n=145 Jeroen Bosch Hospital and n=191 from St Jansdal (from Table 1) was divided into six

Table 1: Patient characteristics of the validation populations.

Variable	Jeroen Bosch Hospital (n=1,140)		Medlon BV (n=1,604)		St. Jansdal Hospital (n=191)	
	Median	Mean	Median	Mean	Median	Mean
Sex, %male, %female	48%m, 52%f		49%m, 51%f		57%m, 43%f	
Age, years	74	68	73	74	73	74
Hemoglobin, mmol/L	7.1	7.0	7.2	7.0	7.0	6.8
MCV, fL	88	87	91	87	92	91
MCH, fmol	1.77	1.73	1.79	1.73	1.80	1.78
Thrombocyte count, $\times 10^9$	279	299	274	291	286	297
Erythrocyte count, $\times 10^{12}$	4.1	4.1	4.1	4.1	3.9	3.9
Leukocyte count, $\times 10^9$	7.4	4.3	7.3	8.1	7.6	12.3
CRP, mg/L	5	37	6	25	14	33
Ferritin, µg/L	50	269	112	247	116	238
Vitamin B12, pmol/L[a]	326	408				
Folic acid, nmol/L[a]	14	18				

[a]n=808 for vitamin B12, n=656 for folic acid. CRP, C-reactive protein; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume.

batches of 56. Four specialists in laboratory medicine from Jeroen Bosch Hospital and two from St Jansdal were each asked to judge a unique batch of 56 patient's laboratory reports providing their sex, age, complete blood count and CRP, and estimate whether a ferritin would be 'low', 'not low' or 'unsure'. In a second round of assessments, each specialist in laboratory medicine was given a different batch of 56 reports which included the result from the machine learning algorithm, which could be used to aid their decision making (as decision support tool). Similarly, six specialists in laboratory medicine at Medlon BV were given the same task of judging 360 independent laboratory reports (in batches of 60 reports), and thereafter a different batch of 60 reports that included the result of the machine learning algorithm.

Implementation and case-finding

The Python script for the machine learning algorithm was directly implemented in our laboratory information system (LIS) using the middleware Gaston Lab®. For the entire month of October 2021 all orders from primary care adult anemic patients at the Jeroen Bosch Hospital were extracted (Figure 2). All values of ≥0.5 were considered predictive of 'low ferritin'. Four samples were no longer available, and no ferritin measurement could be performed.

Statistical analyses

Data were analyzed using Excel 2010 (Microsoft Corporation, USA) and SPSS statistics v25 (IBM, USA). Linear regression analyses were used to assess correlation, with ferritin as dependent variable. Categorical variables were compared by a chi-square test. Logistic regression analyses were used to assess significance between normal vs. low ferritin groups. A p-value <0.05 was considered statistically significant.

Results

Machine learning algorithms were developed to detect deficiencies in iron (ferritin), vitamin B12 and folic acid in anemic primary care patients, based on age, sex, a routine blood count and CRP concentration. We focused on identifying low ferritin values, as the diagnostic performance for folic acid and vitamin B12 deficiencies was limited, with an AUC of the ROC for the validation population of 0.52 for vitamin B12 and 0.57 for folic acid (Supplementary Figure 1A, B). Table 1 shows the demographics of the validation populations. Patients with low ferritin concentrations compared to patients with normal ferritin levels were significantly more often female, younger, and had significantly lower hemoglobin (Hb), MCV, MCH, leukocyte count, CRP, and significantly higher thrombocyte and erythrocyte counts, in the populations of the Jeroen Bosch Hospital and Medlon BV (Table 2).

Using these parameters two separate machine learning algorithms were developed; one based on data from the Jeroen Bosch Hospital (using Siemens, referred to as the JBH-S algorithm) and the other based on data from Medlon BV (using Roche, referred to as the Medlon-R algorithm), to account for poor harmonization and different reference intervals for ferritin. For the JBH-S algorithm an arbitrary algorithm value between 0.5 and 1.0 was considered predictive of low-ferritin, whereas 0.1 and lower was considered predictive for non-low ferritin. Similarly, for the Medlon-R algorithm a value between 0.4 and 1.0 was predictive of a low ferritin and 0.05 and below was considered

Table 2: Laboratory parameters of patients from the validation population with normal plasma ferritin concentrations compared to patients with low plasma ferritin concentrations.

Jeroen Bosch Hospital (n=1,140)

Variable	Normal ferritin (75%) mean ± SD	Low ferritin (25%) mean ± SD	Significance
Sex, %	51% male, 49% female	40% male, 60% female	0.001
Age, years	71 ± 18	57 ± 21	<0.001
Hemoglobin, mmol/L	7.1 ± 0.8	6.6 ± 1.0	<0.001
MCV, fL	89 ± 7	79 ± 8	<0.001
MCH, fmol	1.79 ± 0.18	1.53 ± 0.21	<0.001
Thrombocyte count, ×10^9 cells	294 ± 115	302 ± 107	0.009
Erythrocyte count, ×10^{12} cells	3.99 ± 0.55	4.31 ± 0.52	<0.001
Leukocyte count, ×10^9 cells	8.4 ± 4.6	7.6 ± 2.6	<0.001
CRP, mg/L	24 ± 41	5 ± 7	<0.001
Ferritin, µg/L	189 ± 297	8 ± 5	<0.001

Medlon BV (n=1,604)

Variable	Normal ferritin (69%) mean ± SD	Low ferritin (31%) mean ± SD	Significance
Sex, %	52% male, 48% female	33% male, 67% female	<0.001
Age, years	71 ± 18	55 ± 20	<0.001
Hemoglobin, mmol/L	7.2 ± 0.8	6.6 ± 1.1	<0.001
MCV, fL	91.2 ± 6.4	81.9 ± 8.6	<0.001
MCH, fmol	1.79 ± 0.16	1.53 ± 0.22	<0.001
Thrombocyte count, ×10^9 cells	287 ± 105	318 ± 97	<0.001
Erythrocyte count, ×10^{12} cells	4.1 ± 0.5	4.3 ± 0.5	<0.001
Leukocyte count, ×10^9 cells	8.4 ± 7.4	6.8 ± 2.1	<0.001
CRP, mg/L	29 ± 51	5 ± 14	<0.001
Ferritin, µg/L	288 ± 803	11 ± 6	<0.001

St Jansdal hospital (n=191)

Variable	Normal ferritin (90%) mean ± SD	Low ferritin (10%) mean ± SD	Significance
Sex, %	56% male, 44% female	60% male, 40% female	0.74
Age, years	72 ± 18	55 ± 21	<0.001
Hemoglobin, mmol/L	6.9 ± 1.0	6.1 ± 1.7	0.001
MCV, fL	93.2 ± 7.3	77 ± 13	<0.001
MCH, fmol	1.82 ± 0.18	1.40 ± 0.31	<0.001
Thrombocyte count, ×10^9 cells	296 ± 146	304 ± 68	0.82
Erythrocyte count, ×10^{12} cells	3.8 ± 6.2	4.3 ± 0.9	0.001

Table 2: (continued)

St Jansdal hospital (n=191)

| Variable	Normal ferritin (90%) mean ± SD	Low ferritin (10%) mean ± SD	Significance
Leukocyte count, ×10^9 cells	12.7 ± 55.1	8.3 ± 2.6	0.72
CRP, mg/L	34 ± 50	18 ± 27	0.15
Ferritin, µg/L	191 ± 188	8 ± 5	<0.001

CRP, C-reactive protein; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume.

predictive for non-low ferritin. However, based on local preferences, different cut-off values can be chosen based on the preferred sensitivity and specificity as presented in Supplementary Table 1.

The validation dataset of the JBH-S algorithm (n=1,140, Jeroen Bosch Hospital) reached an AUC of the ROC of 0.92 (Figure 1A). The two most important parameters predicting a ferritin value under the reference limit were MCH and MCV (Figure 1B). The JBH-S algorithm was also validated by an external dataset of 191 patients of the St Jansdal Hospital (Siemens equipment for chemistry but Sysmex for hematology) reaching an AUC of the ROC of 0.92 (Figure 1C). The validation population (n=1,604) of the Medlon-R algorithm reached an AUC of the ROC of 0.90 (Figure 1D), with MCH and MCV as the two most important predictive parameters (Figure 1E). Precision-recall curves are presented in Supplementary Figures 1C, D.

Performance of the ML algorithm vs. specialists in laboratory medicine

Twelve specialists in laboratory medicine assessed independent batches of 56–60 laboratory results (age, sex, blood count and CRP) of anemic primary care patients and predicted whether they would have a low or ferritin or not. Laboratory specialists could also choose the option 'unsure', if they felt a good prediction was not possible based on the test results. The specialists of the Jeroen Bosch Hospital and St Jansdal (Siemens analyzers) reached a sensitivity of 83% and a specificity of 92%, and made a judgement (low or non-low) in 79% of the patients (Table 3). The JBH-S algorithm outperformed the specialists with a sensitivity of 93% and specificity of 92% (Table 3). When the specialists were requested to assess a different batch of 56 anemic patients, but with the output of the JBH-S algorithm visible (as support tool), the specialists
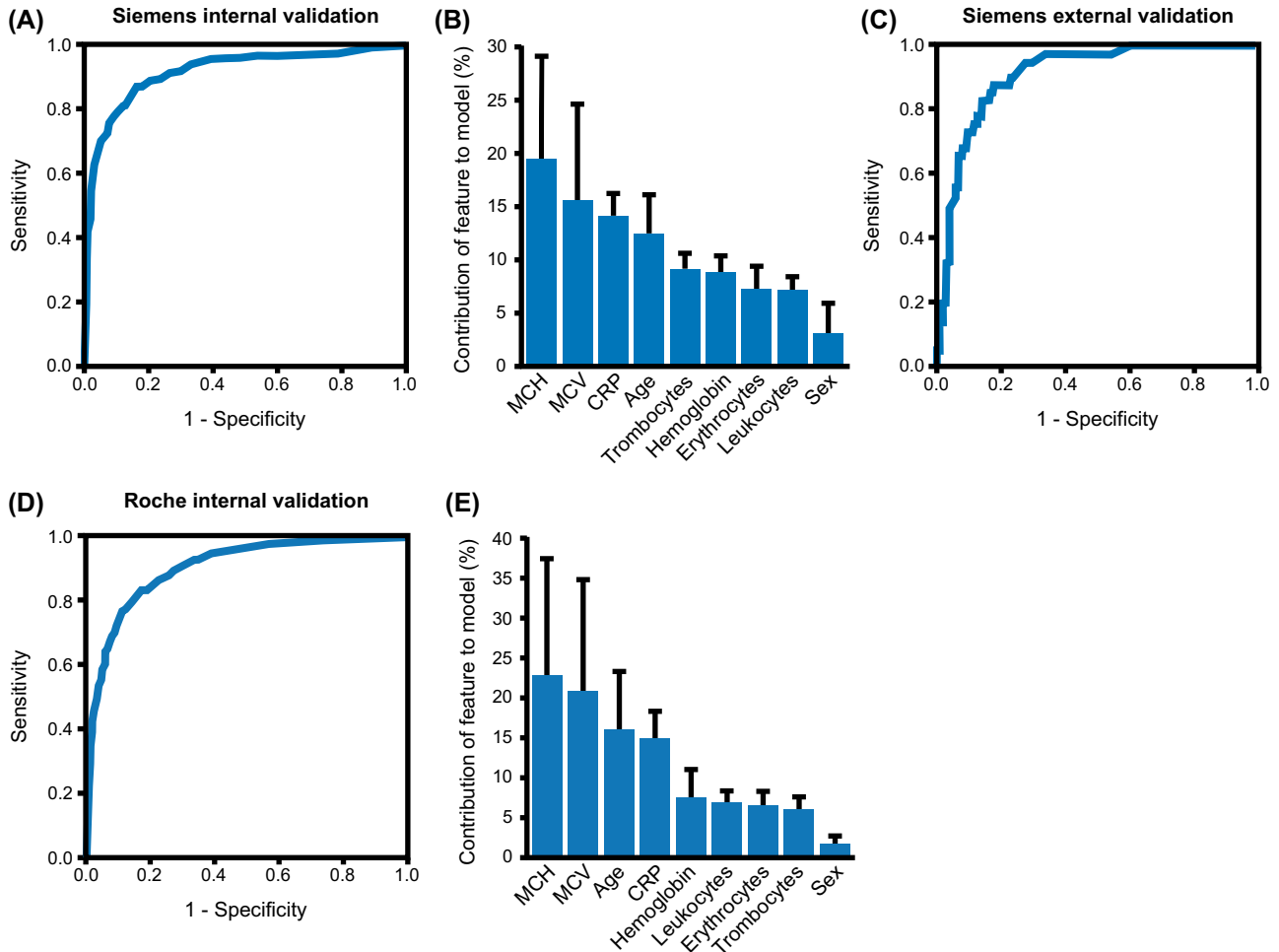
Figure 1: Performance of two machine learning algorithms trained to predict low ferritin concentrations.
(A) ROC plot for the prediction of low ferrin values (<10 µg/L in females and <22 µg/L in males) in the validation population of the Jeroen Bosch Hospital (Siemens, the JBH-S model), with an area under the curve (AUC) of 0.92. (B) Contribution of each individual parameter to the machine learning algorithm of Figure A, plotted as mean ± SD (C) ROC plot of the external validation of the JBH-S model, using data from the St Jansdal Hospital, reaching an AUC of the ROC 0.92. (D) ROC plot for the prediction of low ferritin values (<13 µg/L in females and <30 µg/L in males) in the validation population of Medlon BV (Roche, the Medlon-R model), with an AUC of 0.90. (E) Contribution of each individual parameter to the machine learning algorithm of Figure D, plotted as mean ± SD

still performed slightly poorer than the algorithm itself, with 87% sensitivity and 90% specificity. However, compared to judgement without algorithm they made a decision more often and required less time per patient (Table 3). Six specialists from Medlon BV (Roche), each assessing a batch of 60 anemic patients, reached a sensitivity of 88% and specificity of 91%, whereas the Medlon-R algorithm reached a sensitivity of 98% and specificity of 92% (Table 3). When using the Medlon-R algorithm output as a support tool, the specialists from Medlon BV reached a sensitivity of 84% and a specificity of 93%.

As patients with an iron deficiency can still have a ferritin concentration marginally above the lower reference limit, we took a closer look at the false positives (e.g. patient assessed as having low ferritin, but not having a low ferritin concentration when measured). The median difference from the lower limit of normal (LLN) of the specialists was 19 µg/L vs. 5 µg/L for the JBH-S algorithm, and 23 µg/L for the specialists vs. 11 µg/L for the Medlon-R algorithm (Table 3). Moreover, for the JBH-S algorithm, 50% of the false positives were within 5 µg/L of LLN vs. 28% of the specialists' false positives (Table 3). These data indicate that the magnitude of the errors of the false positives is much smaller for the ML-algorithms compared to the incorrect assessments made by the specialists.

Table 3: Diagnostic performance of specialists vs. the machine learning algorithm.

Jeroen Bosch Hospital/St Jansdal (n=336)

	Specialists	JBH-S algorithm	Specialists with ML as support tool
Sensitivity, %	83	93	87
Specificity, %	92	92	90
Decision, %	79	73	84
Low ferritins identified, %	63	75	77
Average time required per patient, seconds	19	<1	13
Median error of false positives, µg/L	19	5	21
False positives within 5 µg/L of the LLN,%	28	50	32

Medlon BV (n=360)

	Specialists	Medlon-R algorithm	Specialists with ML as support tool
Sensitivity, %	88	98	84
Specificity, %	91	92	93
Decision, %	68	67	70
Low ferritins identified, %	67	73	58
Average time required per patient, seconds	20	<1	16
Median error of false positives, µg/L	23	11	17
False positives within 5 µg/L of the LLN, %	10	29	13

JBH-S, Jeroen Bosch Hospital Siemens model; Medlon-R, Medlon BV Roche model; ML, machine learning; LLN, lower limit of normal. 'Specialists' are specialists in laboratory medicine. 'Low ferritins identified' indicates the percentages of patients with a low ferritin level that were correctly identified as having a low plasma ferritin concentration (true positives).

Implementation and case-finding

The JBH-S algorithm was directly implemented in the LIS of the Jeroen Bosch Hospital. In October 2021 all laboratory orders of anemic adult primary care patients were prospectively analyzed. Ferritin was measured in all adult anemic primary care patients with a machine learning value of ≥0.5 (Figure 2). Using this automated machine learning algorithm 18 new iron deficiencies were identified in 21 work days. Four samples were no longer available for measurement.
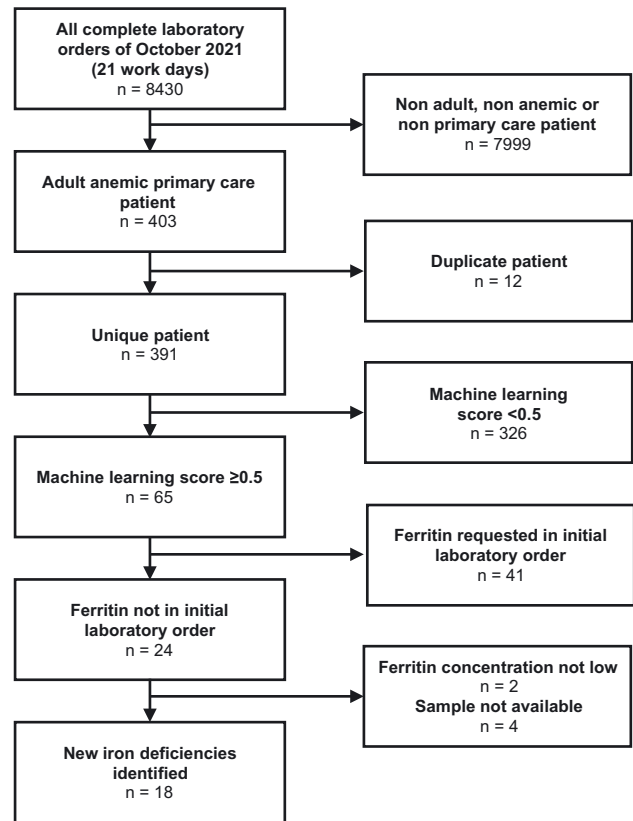
Figure 2: Prospective case-finding.
After implementation of the algorithm in the laboratory information system of the Jeroen Bosch Hospital prospective analysis was performed for all laboratory orders containing a complete blood count and CRP during the month of October 2021. Ferritin was measured in all adult anemic primary care patients with a machine learning value of ≥0.5. In total 18 new iron deficiencies were identified in 21 workdays. Four samples were no longer available for measurement.

Discussion

Iron-deficient anemia is one of the most common diseases worldwide, but still remains underdiagnosed. In this study, test results from a blood count and CRP in anemic patients from primary care were used to develop machine learning algorithms which could accurately predict low ferritin concentrations. It was not possible to predict deficiencies in folic acid or vitamin B12 based on these parameters. Both algorithms outperformed the specialists in laboratory medicine in predicting low ferritin levels. Implementation of the algorithm in our LIS resulted in one new iron deficiency diagnosis on average per day.

In a time with more structured healthcare data in an electronic patient record, possibilities for machine learning

and decision support tools are expanding. Especially in the diagnostic field, algorithms to support medical decision making are rapidly growing, as they are fast, accurate and cheap. Decision support tools provide a welcome opportunity to lower the work pressure of physicians in an aging population with multimorbidity [14, 15].

In several regions of the Netherlands, general practitioners are supported by specialists in laboratory medicine with a service called 'reflective testing' [7, 16, 17]. This service provides diagnostic guidance by the addition of interpretative comments, and moreover additional tests can be added by the specialist in laboratory medicine to complete the diagnostic workup. This saves time and prevents additional phlebotomy for the patient, and has been proven to be highly appreciated by general practitioners [7]. As the evaluation of changes in patterns across numerous laboratory parameters can be complex or subtle the algorithm developed in this study can assist laboratory specialists to predict iron deficiency more accurately. Furthermore, assessment of test reports by specialists in laboratory medicine of patients suffering from anemia is faster when using the algorithm as a support tool. The tool developed in this study may be used as a decision support tool and as a case-finding tool. Moreover, when the algorithm evaluates the risk of a patient having a low ferritin level to be either extremely low or extremely high, the measurement of the ferritin concentration might be omitted, thereby saving time and costs.

However, several limitations have to be taken into consideration. Firstly, the algorithms may only be used for the specific population (e.g. anemic primary care adult patients) for which it has been validated. Therefore, before implementation a validation with local data is essential. Secondly, when implementing a machine learning algorithm the new European '*In Vitro* Diagnostics Regulation' (IVDR) should be taken into consideration, as this regulation can also apply to certain software tools. Lastly, ferritin measurements are poorly harmonized and different lower reference limits are used per laboratory. In this study, the models were trained using the reference limits provided by the product insert leaflet of the providers (Siemens and Roche). Therefore, the performance of the algorithms could be diminished when using different analyzers or different lower reference limits.

On the other hand, this study has several key strengths. The machine learning algorithm was developed using large amounts of data, representing laboratory results from thousands of patients. Therefore, laboratory data of diseases that have similarities with iron deficiencies (chronic disease, hemoglobinopathy and thalassemia) are included in the training and validation set. Secondly, the JBH-S

algorithm was validated using an internal validation and an external validation set, from a laboratory using different hematology analyzers (Siemens vs. Sysmex). Moreover, separate algorithms were developed for Siemens and Roche chemistry analyzers. Lastly, using Supplementary Table 1 it is possible to adapt the cut-off value of the algorithm in order to choose the optimal sensitivity and specificity based on local requirements and preferences.

Multiple studies have used machine learning on laboratory parameters to predict disease or the outcome of other laboratory parameters [18–20]. However, many of the algorithms developed in recent studies appear to find limited application in clinical practice. In this study we have focused on extensive validation of our models (internal and external), comparing the performance of the algorithm to experts, implementing the algorithm directly in our LIS and, finally, analyzing the benefit of implementation for patient care. Moreover, we have used a minimal amount of laboratory parameters (complete blood count + CRP), while retaining a high diagnostic value, to maximize the clinical applicability.

For the implementation of a machine learning algorithm in daily clinical care, software is needed. So called clinical decision support systems are commercially available, but need to be filled with content (algorithms or 'clinical rules') and need to be connected to clinical data from the electronic health record and other data systems, such as the laboratory information system [21]. Once a decision support system is implemented with clinical rules, it brings many opportunities to improve test result interpretation and the efficiency with which diagnostic data can be converted into useful information [22]. In this study we have shown that the integration of a machine learning algorithm to predict low ferritin levels in primary care patients is a valuable diagnostic tool which can support physicians and specialists in laboratory medicine, and can automatically identify unrecognized iron deficiencies.

Competing interests: Authors state no conflict of interest.
Informed consent: The need for informed consent was waived by the Medical Research Ethics Committee Brabant.
Ethical approval: The study was conducted according to the declaration of Helsinki, Guidelines for Good Clinical Practice. The execution of this retrospective observational study of patient records was approved by the local review board of the Jeroen Bosch Hospital. The execution of this retrospective observational study of patient records was judged by the Medical Research Ethics Committee Brabant (METC Brabant), which waived this study to be subject to the regulations of the WMO (Dutch Medical Research Involving Human Subjects Act), including a waiver of informed written consent.
Data availability: Information and data are available from the corresponding author upon reasonable request. The code for the algorithms is freely available on https://github.com/Tdebel/ferritin-ai.

References

1. Vos T, Abajobir AA, Abbafati C, Abbas KM, Abate KH, Abd-Allah F, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of disease study 2016. Lancet 2017;390:1211–59.
2. Thachil J. Iron deficiency: still under-diagnosed? Br J Hosp Med 2015;76:528–32.
3. Pasricha SR, Tye-Din J, Muckenthaler MU, Swinkels DW. Iron deficiency. Lancet 2021;397:233–48.
4. Peyrin-Biroulet L, Williet N, Cacoub P. Guidelines on the diagnosis and treatment of iron deficiency across indications: a systematic review. Am J Clin Nutr 2015;102:1585–94.
5. Daru J, Colman K, Stanworth SJ, De La Salle B, Wood EM, Pasricha SR. Serum ferritin as an indicator of iron status: what do we need to know? Am J Clin Nutr 2017;106:1634S.
6. Hoofnagle AN. Harmonization of blood-based indicators of iron status: making the hard work matter. Am J Clin Nutr 2017;106: 1615S–1619S.
7. Oosterhuis WP, van de Venne WPV, van Deursen CT, Stoffers HEJH, van Acker BA, Bossuyt PMM. Reflective testing – a randomized controlled trial in primary care patients. Ann Clin Biochem 2021;58:78–85.
8. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21:233–41.
9. Lessmann N, Sánchez CI, Beenen L, Boulogne LH, Brink M, Calli E, et al. Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. Radiology 2021;298:E18–28.
10. Kurstjens S, van der Horst A, Herpers R, Geerits MWL, Kluiters-De Hingh YCM, Göttgens EL, et al. Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing. Clin Chem Lab Med 2020;58:1587–93.
11. Yang HS, Hou Y, Vasovic LV, Steel PAD, Chadburn A, Racine-Brzostek SE, et al. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. Clin Chem 2020; 66:1396–404.
12. Çallı E, Murphy K, Kurstjens S, Samson T, Herpers R, Smits H, et al. Deep learning with robustness to missing data: a novel approach to the detection of COVID-19. PLoS One 2021;16. https://doi.org/10.1371/journal.pone.0255301.
13. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6.
14. McPhail SM. Multimorbidity in chronic disease: impact on health care resources and costs. Risk Manag Healthc Pol 2016;9: 143–56.
15. Rudolf JW, Dighe AS. Decision support tools within the electronic health record. Clin Lab Med 2019;39:197–213.
16. Simpson WG, Twomey PJ. Reflective testing. J Clin Pathol 2004;57: 239–40.
17. Murphy MJ. Reflex and reflective testing: progress, but much still to be done. Ann Clin Biochem 2021;58:75.
18. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol 2016;145: 778–88.
19. Park DJ, Park MW, Lee H, Kim YJ, Kim Y, Park YH. Development of machine learning model for diagnostic disease prediction based on laboratory tests. Sci Rep 2021;11:7567.
20. Boerman AW, Schinkel M, Meijerink L, van den Ende ES, Pladet LC, Scholtemeijer MG, et al. Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study. BMJ Open 2022;12:e053332.
21. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020; 3:17.
22. Van Balveren JA, Verboeket-Van De Venne WPHG, Erdem-Eraslan L, De Graaf AJ, Loot AE, Musson REA, et al. Impact of interactions between drugs and laboratory test results on diagnostic test interpretation – a systematic review. Clin Chem Lab Med 2018;56:2004–9.