CME/SAM

Using Machine Learning to Predict Laboratory Test Results

Yuan Luo, MS,[1] Peter Szolovits, PhD,[1] Anand S. Dighe, MD, PhD,[2,3] and Jason M. Baron, MD[2,3]

From the [1]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge; [2]Department of Pathology, Massachusetts General Hospital, Boston; and [3]Harvard Medical School, Boston, MA.

Key Words: Machine learning; Ferritin; Clinical decision support; Statistical diagnosis; Imputation; Computational pathology

Am J Clin Pathol June 2016;145:778-788

DOI: 10.1093/AJCP/AQW064

ABSTRACT

Objectives: *While clinical laboratories report most test results as individual numbers, findings, or observations, clinical diagnosis usually relies on the results of multiple tests. Clinical decision support that integrates multiple elements of laboratory data could be highly useful in enhancing laboratory diagnosis.*

Methods: *Using the analyte ferritin in a proof of concept, we extracted clinical laboratory data from patient testing and applied a variety of machine-learning algorithms to predict ferritin test results using the results from other tests. We compared predicted with measured results and reviewed selected cases to assess the clinical value of predicted ferritin.*

Results: *We show that patient demographics and results of other laboratory tests can discriminate normal from abnormal ferritin results with a high degree of accuracy (area under the curve as high as 0.97, held-out test data). Case review indicated that predicted ferritin results may sometimes better reflect underlying iron status than measured ferritin.*

Conclusions: *These findings highlight the substantial informational redundancy present in patient test results and offer a potential foundation for a novel type of clinical decision support aimed at integrating, interpreting, and enhancing the diagnostic value of multianalyte sets of clinical laboratory test results.*

Upon completion of this activity you will be able to:
- describe the value of machine learning in integrating and mining clinical laboratory data.
- examine the information redundancy present in a set of common laboratory test results.

The ASCP is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The ASCP designates this journal-based CME activity for a maximum of 1 *AMA PRA Category 1 Credit*™ per article. Physicians should claim only the credit commensurate with the extent of their participation in the activity. This activity qualifies as an American Board of Pathology Maintenance of Certification Part II Self-Assessment Module.

The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Exam is located at www.ascp.org/ajcpcme.

Clinical laboratories report most test results as individual numerical or categorical values. However, individual tests results, viewed in isolation, are typically of limited diagnostic value. To adequately use test results for patient diagnosis and management, clinicians usually must integrate many individual test results from a patient and interpret them in the context of clinical data and medical knowledge, judgment, and experience. While this manual approach to test result interpretation is the current standard in most cases, computational approaches to laboratory data integration and analysis offer tremendous potential to enhance diagnostic value.[1] In particular, many patients will have hundreds or thousands of these individual test results, often spanning years. As a consequence, busy clinicians can easily overlook key results or important patterns

and trends within sets of laboratory data. Furthermore, important diagnostic information may sometimes be contained within patterns across numerous data elements that may be too subtle or complex to identify without the aid of computational approaches.[2] In addition, because the human brain faces great challenges in simultaneously considering a large number of data points, even the most experienced clinicians may be unable to extract all the useful information from existing clinical and laboratory data.[2]

Electronic clinical decision support represents an important tool to improve test result interpretation and the efficiency with which diagnostic data can be converted into useful information. For example, we have developed and implemented an algorithm within our laboratory information system to identify and append an alert to creatinine results that are trending upward and suggestive of acute kidney injury.[3] Likewise, we have demonstrated that we can identify spurious elevations in plasma glucose results using the results of other analytes measured on the same sample.[4]

While rule-based and statistically based algorithms can both provide a foundation for clinical decision support, most currently used decision support relies on rule-based approaches.[5] Rule-based algorithms tend to be easier than statistical algorithms to develop, validate, implement, and explain and can often be adapted directly from guidelines or literature.[5] However, most rule-based algorithms applied in clinical practice provide decision support based on previously established knowledge. In contrast, statistically based approaches offer an opportunity to combine knowledge discovery with knowledge application to provide decision support based on previously unknown patterns.[5]

Here, we describe a novel framework for statistical integration of test results, and as a proof of concept, we apply this framework to patients receiving ferritin testing. Ferritin is a marker of iron stores and is used in the diagnosis of iron deficiency[6] but by itself can be misleading. For example, ferritin is increased in inflammation,[6] and iron-deficient patients undergoing inflammatory responses may have normal ferritin values. In this article, we first posit and test the hypothesis that ferritin test results can be predicted from the results of other tests ordered on the same patient. We confirm this hypothesis, suggesting that the information provided by ferritin is often substantially *redundant* given other tests performed alongside it. We then consider cases where the ferritin and predicted ferritin are discrepant and show that predicted ferritin may have diagnostic value in these cases. Finally, we propose three strategies in which this statistically based approach could be applied to clinical decision support.

Materials and Methods

Data

This study used data from testing performed at the Massachusetts General Hospital (MGH), a 989-bed tertiary care hospital in Boston, Massachusetts, collected with approval from our hospital's institutional review board. The data set included all outpatient ferritin results collected over a 3-month period in 2013. Each ferritin result was linked to the patient's age, sex, and results for "predictor tests" performed on the same collection (patient, date, time combination). In rare cases, in which more than one result for a test was reported on a patient at the same date and time, the mean result was used. Due to limitations of our data extraction strategy, laboratory results reported with a value of 0 (with the exception of nucleated RBCs) were excluded from analysis and treated as missing. The complete list of predictor tests can be viewed in **Table 1**. Results were included only for testing performed within the main hospital laboratories; point-of-care test results and results of testing performed in satellite laboratories at affiliated health centers were excluded. In the final data sets, collections were excluded if they did not contain at least two predictor tests. The cases were split randomly into training and test partitions in a 7:3 ratio for a final training set of 3,590 cases and a test set of 1,538 cases. Because the data set was quite large, we expected that random selection alone would ensure an acceptable degree of similarity between the training and test data sets, and we did not perform any specific stratification. Supplemental Figure 1 (all supplemental materials can be found at *American Journal of Clinical Pathology* online) shows Q-Q plots for ferritin values in the training and test data sets and helps confirm that the distribution of ferritin values in the training data set is quite similar to that in the test data.

Data Transformations

Many laboratory tests have results that closely follow a lognormal distribution. In regression analysis, minimizing the root mean square error is equivalent to the maximum likelihood estimation, only under the assumption that the target variable adopts a normal distribution.[7] Thus, we transformed ferritin values using a natural log transformation: $y = \ln(1 + x)$, where y is the transformed ferritin value and x is the original ferritin value. Supplemental Figure 2 shows Q-Q plots and demonstrates approximate lognormal distributions for ferritin. We inverted this transformation on predicted values of log ferritin (predictions as described below) to calculate predicted values of ferritin in untransformed units.

DOI: 10.1093/ajcp/aqw064

Table 1
Description of Laboratory and Demographic Parameters

Abbreviation	Description	Missing, %	Median (IQR)	Reporting Units	Adult Reference Range
%Baso	Percent basophils	65	0.4 (0.3-0.7)	%	0-3
%Lymph	Percent lymphocytes	63	27.4 (20.5-35.15)	%	22-44
%Mono	Percent monocytes	63	7.9 (6.3-9.6)	%	4-11
%Neut	Percent neutrophils	63	60.0 (52.6-67.7)	%	40-70
%NucRBC	Percent nucleated RBCs	29	0 (0-0)	/100 WBCs	0
AbsBaso	Absolute basophil count	65	0.3 (0.2-0.4)	$\times 10^3/\mu L$	0.0-0.3
AbsEos	Absolute eosinophil count	64	0.14 (0.08-0.23)	$\times 10^3/\mu L$	0.0-0.9
AbsLymph	Absolute lymphocyte count	63	1.77 (1.30-2.38)	$\times 10^3/\mu L$	1.0-4.8
AbsMono	Absolute monocyte count	63	0.52 (0.4-0.68)	$\times 10^3/\mu L$	0.2-1.2
AbsNeut	Absolute neutrophil count	63	3.98 (2.92-5.33)	$\times 10^3/\mu L$	1.8-7.7
Age	Age (y)	NA	52 (36-67)	NA	NA
Albumin	Albumin	51	4.5 (4.2-4.7)	g/dL	3.3-5.0
AlkPhos	Alkaline phosphatase	56	77 (61-99)	U/L	Female: 30-100; male: 45-115
ALT	Alanine transaminase	54	18 (13-28)	U/L	Female: 7-33; male: 10-55
Anion	Anion gap	47	13 (11-15)	mEq/L	3-15
AST	Aspartate transaminase	54	23 (18-30)	U/L	Female: 9-32; male: 10-40
B12	B_{12}	59	604 (449-870)	pg/mL	>250
Bicarb	Bicarbonate	47	25.0 (23.3-26.6)	mmol/L	23.0-31.9
BUN	Blood urea nitrogen	45	15 (11-20)	mg/dL	8-25
Ca	Calcium	46	9.4 (9.1-9.7)	mg/dL	8.5-10.5
Cl	Chloride	47	101 (99-103)	mmol/L	100-108
Cr	Creatinine	45	0.83 (0.7-1.0525)	mg/dL	0.60-1.50
Eos	Percent eosinophils	64	2.1 (1.2-3.4)	%	0-8
Fe	Iron	14	74 (50-100)	µg/dL	Female: 30-160; male: 45-160
FER	Ferritin	0	67 (27-163)	ng/mL	Female: 10-200; male: 30-300
Gender	Gender	NA	NA	NA	NA
Glob	Globulin	56	2.6 (2.3-2.9)	g/dL	2.3-4.1
Glu	Glucose	46	90 (82-103)	mg/dL	70-110
Hct	Hematocrit	29	38.4 (35.1-41.4)	%	Female: 36.0-46.0; male: 41.0-53.0
Hb	Hemoglobin	29	12.8 (11.5-13.9)	g/dL	Female: 12.0-16.0; male: 13.5-17.5
K	Potassium	46	4.0 (3.7-4.3)	mmol/L	3.4-4.8
MCH	Mean cell hemoglobin	29	29.7 (27.7-31.2)	pg/RBC	26.0-34.0
MCHC	Mean cell hemoglobin concentration	29	33.2 (32.3-34.1)	g/dL	31.0-37.0
MCV	Mean cell volume	29	89 (84-93)	fL	Female: 80-100; male: 80-100
Na	Sodium	46	139 (138-141)	mmol/L	135-145
NucRBC	Absolute nucleated RBCs	29	0 (0-0)	$\times 10^3/\mu L$	0
Plt	Platelets	29	250 (202-301)	$\times 10^3/\mu L$	150-400
Prot	Total protein	56	7.1 (6.8-7.4)	g/dL	6.0-8.3
RBC	RBC count	29	4.38 (3.98-4.73)	$\times 10^6/\mu L$	Female: 4.00-5.20; male: 4.50-5.90
RDW	RBC distribution width	29	13.8 (13-15.3)	%	11.5-14.5
TBILI	Total bilirubin	56	0.4 (0.3-0.6)	mg/dL	0.0-1.0
TIBC	Total iron-binding capacity	14	303.5 (264-347)	µg/dL	230-404
WBC	WBC count	29	6.7 (5.4-8.3)	$\times 10^3/\mu L$	4.5-11.0

IQR, interquartile range; NA, not applicable.
[a]Shown are demographic and laboratory test parameters used in this study. The abbreviations in the leftmost column correspond to the row labels in Figure 1. Median and IQR values represent the median and IQR for each analyte (when available) in the study data set.

Overview of Imputation and Prediction

Most of our analysis relies on a two-stage procedure. In the first stage of this procedure, we imputed missing test results (tests not performed) for tests other than ferritin ("predictor tests"). Then, in the second stage, we used the measured and imputed values for these predictor tests plus age and sex to predict ferritin test results. In this second stage, we predicted both numerical results for ferritin ("regression") and whether ferritin results would be normal or abnormal ("classification"). Although no ferritin test results were actually missing from our data set per the inclusion criteria, we assessed model performance and ferritin predictability by masking ferritin results from a held-out test partition of our data and then comparing predicted ferritin results with the masked (measured) values. The masked-measured values were treated as the "ground truth" in assessing model performance.

The imputation stage was required because the prediction algorithms used in the second stage of our procedure could not directly accommodate missing data in predictors. Our data set, like most clinical and laboratory data sets obtained in clinical practice, contained many missing values, which in our particular analysis represent tests not performed. The imputation step allowed us to infer missing

DOI: 10.1093/ajcp/aqw064

values by tapping into observed associations between results from various tests (plus age and sex) and in turn to apply the prediction methods used in the second stage to our data set.

More specifically, we applied four different imputation methods and five different prediction (regression or classification) methods. In addition, we trained and tested the performance of each regression and classification model using two different sets of predictors. The first predictor set consisted of just patient demographics (age and sex) and laboratory test results and did not distinguish between measured test results (tests performed) and imputed test results (tests not performed). Like the first predictor set, the second predictor set also used demographics and test results but also included a set of dichotomous variables describing whether each test result was measured (test performed) or missing (not preformed and thus requiring that the result be imputed). We refer to models trained and tested with just demographics and actual or imputed test results as using the "predictor set without missingness." We refer to models trained and tested with demographics, actual or imputed test results, and the dichotomous variables denoting whether results were measured or missing (imputed) as using the "predictor set with missingness."

We paired each of the four imputation methods with each of the five prediction methods across each of the two predictor set types to generate a total of 40 sets (4*5*2) of predicted ferritin test results. We describe the four imputation and five prediction techniques below. When not otherwise specified, we use the term *impute* (and *imputation*) to refer to prediction of missing predictor test results, *regression* to refer to prediction of numerical ferritin results, and *classification* to refer to prediction of whether ferritin results would be normal or abnormal.

Imputation

The four imputation techniques used were the following: mean, multiple imputation with chained equations–full (MICE-full), multiple imputation with chained equations–select (MICE-sel), and missForest. Mean imputation imputes missing values as the mean of the available values for each variable. MICE-full, MICE-sel, and missForest were performed using the MICE[8] and missForest[9] R-packages (https://cran.r-project.org/). These techniques are described in greater detail in the supplemental methods. Training and test data were combined into a single data set for imputation; however, outcome ferritin results were excluded from the test cases prior to imputation. This way, the imputation step could not "leak" ferritin information from the test data set. Thus, overfitting should not bias classification or regression performance on the test data, and test data set evaluation should provide an unbiased estimate of generalizable performance. MICE-full, MICE-sel, and missForest are designed to provide nondeterministic outputs intended to model the uncertainty in missing values. To capture this uncertainty, we ran each of the imputation algorithms 100 times with random initialization to generate 100 imputed data sets. The reported values for correlation and area under the curve (AUC) and the plotted values for sensitivity and specificity represent the mean of each statistic across the 100 imputation runs. The numerical values of predicted ferritin as reported or plotted represent the median across the 100 imputation runs.

Regression

The four regression techniques used were linear regression, Bayesian linear regression, random forest regression (RFR), and lasso regression (lasso). We used the Scikit-learn[10] Python package to implement them. (http://scikit-learn.org) Additional technical detail regarding the imputation and regression methods is provided in the supplemental methods.

Classification

Ferritin results were classified as normal (within the normal reference limits set by the MGH Core laboratory) or abnormal. The lower limit of normal for ferritin at MGH is 10 ng/mL in females and 30 ng/mL in males. Although ferritin also has an upper reference limit (200 ng/mL in females and 300 ng/mL in males), only low ferritin results were classified as abnormal since the goal was to identify iron deficiency, which is indicated by a low ferritin. Classification was performed using logistic regression as implemented in the Python Scikit-learn package.[10] Receiver operator characteristic (ROC) curves were generated by varying the probability threshold at which the logistic regression would classify results as abnormal.

Univariate Analysis

To provide an assessment of the contribution of each predictor laboratory test in predicting ferritin, we calculated the correlation expressed as Pearson's r between each analyte and log ferritin. We also calculated the percentage of variance in log ferritin explained by each analyte (R^2) and the discriminative power of each analyte to distinguish normal from abnormal values of ferritin, expressed as the AUC (c-statistic). The univariate analysis was based on pairwise complete cases and was performed in R.

Case Review

We selected cases (as described in the Results section) for detailed chart review. Each selected case was independently reviewed by two pathologists (J.M.B. and A.S.D.) to assess the patient's underlying iron status based on each patient's electronic medical record, including physician notes

DOI: 10.1093/ajcp/aqw064

and laboratory test results performed before, subsequent to, and at the same time as the ferritin test under consideration. Note that the pathologist review used information (including future progression) not available to the computational algorithms and was treated as a reference to which predicted ferritin results could be compared.

Results

Ninety-seven percent of patient collections, including a measured, outpatient ferritin result, were accompanied by at least two other predictor tests and included in the final data sets. This final data set, consisting of 5,128 test collections, was randomly split into a training set of 3,590 results and a test set of 1,538 results. Besides ferritin, each collection had a median of 23 of the 40 other tests measured (interquartile range [IQR], 11-31). Table 1 lists the predictor tests. For each predictor test, Table 1 also lists the percentage of values that were missing (and thus required imputation), the median and interquartile range for the study population, and the adult reference range. **Figure 1** describes the raw data and reveals the relative correlations between ferritin and other tests.

We first present the extent to which ferritin can be predicted by examining the correlation statistics (all prediction methods) and scatterplots (selected method) denoting the relationship between measured and predicted ferritin values. We then present the ability of classification algorithms to predict whether ferritin will be abnormal, with the quality of the prediction measured by AUC. We next present data regarding the univariate association between each predictor analyte and ferritin and finally present results from chart review of select cases.

Predictability of Numerical Ferritin Results (Regression)

We applied 16 different imputation-regression pairings to the training and held-out test data partitions to identify the best-performing methods. Supplemental Table 1 shows the correlation between measured and predicted ferritin (both log transformed) for each of the imputation-regression pairings. The best performance on held-out test data in the data set without missingness achieved a correlation of 0.732, using MICE-sel imputation and RFR. Lasso regression following missForest imputation performed almost as well (correlation = 0.729 in test data, predictor set without missingness). We selected the lasso-missForest pairing as the basis for additional evaluation, including subsequently described case studies.

Regression and bias plots for the lasso-missForest pairing are shown in **Figure 2** and demonstrate the substantial predictability of ferritin.

Predictability of Ferritin Classifications

In clinical decision making, a key consideration in interpreting numerical laboratory results is often just whether the results fall within the normal reference range. Accordingly, we sought to determine if we could accurately predict whether results would be within this normal range. **Figure 3** provides ROC curves showing the predictability of abnormal ferritin classifications on the independent test data, using logistic regression following imputation by each of the four methods. MissForest imputation performed best overall. This technique, when paired with logistic regression, achieved AUCs of 0.96 and 0.97, respectively, in classifying ferritin results using the predictor sets with and without missingness. As in the case of regression, the additional information provided in the predictor set with missingness added little or no value. As a negative control, classification was performed on both predictor set types with ferritin results randomly reshuffled between patient cases of predictor data. ROC curves for the reshuffled data sets are shown in Supplemental Figure 3 and, as expected, demonstrate AUC values of approximately 0.5.

Contributions of Individual Analytes

Regression coefficients for individual analytes from the imputation-regression analysis are difficult to interpret, because the regression used a combination of measured and imputed values, and thus the coefficients for each analyte will not be based just on the measured values for that analyte. Furthermore, colinearity and regularization are likely to further distort regression coefficients. Thus, to assess potential contributions of individual analytes in predicting ferritin, we performed a series of univariate analyses, as shown in **Table 2**. Here, we present the pairwise correlation between nonmissing values for each individual analyte and ferritin (log transformed). Likewise, to assess the ability of each individual analyte to distinguish normal from low ferritin results, Table 2 lists the univariate AUC (*c*-statistic) for each analyte. Not surprisingly, total iron-binding capacity, mean cell hemoglobin, and mean cell hemoglobin concentration were the most informative analytes with respect to both classification and regression.

Selected Case Review

Hypothesizing that in some cases, predicted ferritin may be more representative of patient iron status than measured ferritin, as shown in **Figure 4**, we identified 26 (1.7%) cases in our held-out test data set in which the predicted ferritin and actual ferritin were highly discrepant. For this purpose, we defined *highly discrepant* to mean that actual and predicted ferritin differed by a factor of 10 or more, when predictions were made using missForest imputation

DOI: 10.1093/ajcp/aqw064

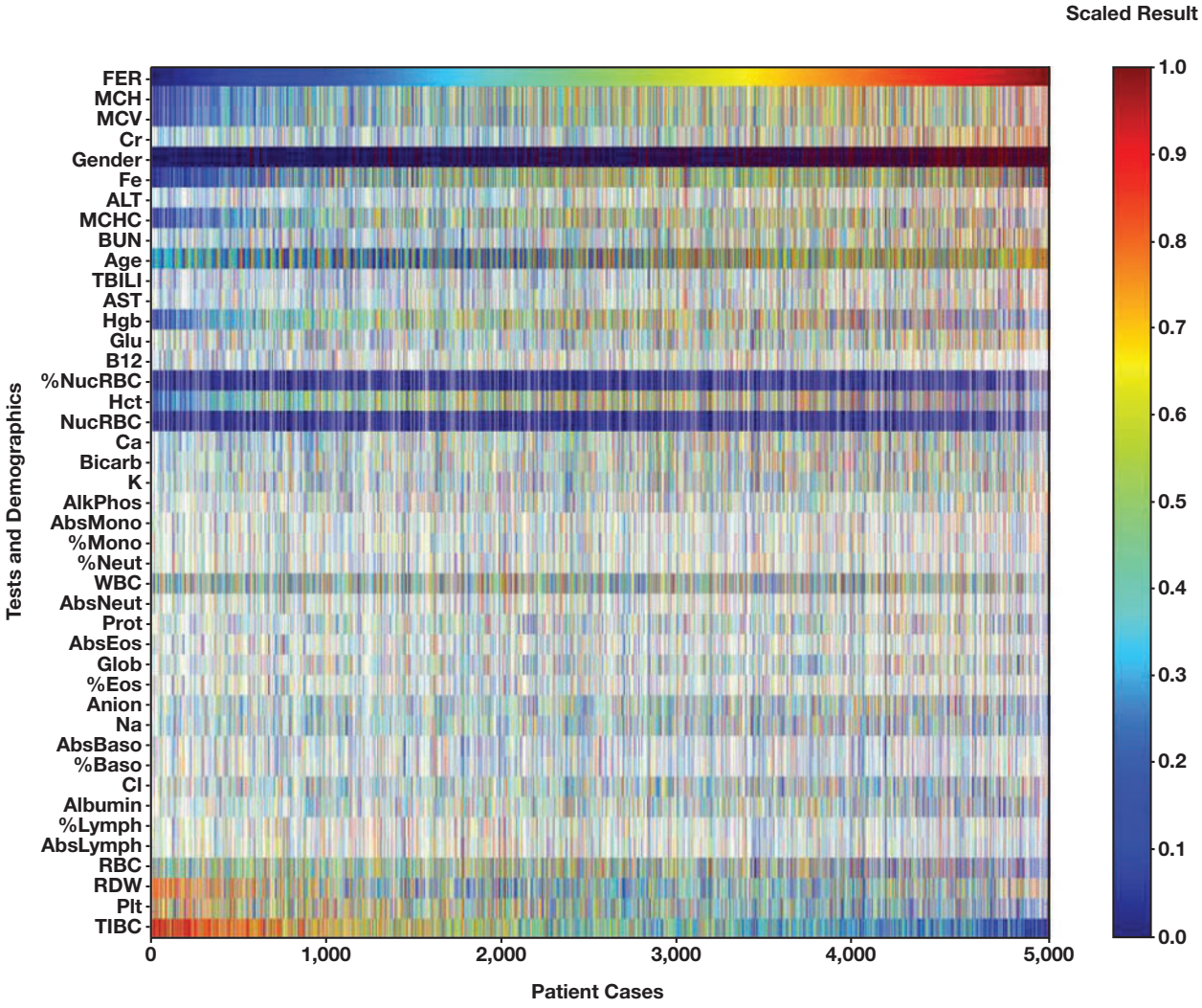


Figure 1 Data set overview and correlations. In this heat map, rows represent predictors (see Table 1 for translations) and columns represent patient cases. Cell colors are based on result quantiles (eg, 75th percentile = 0.75), and missing results are white. Parameters are ordered by increasing correlation with ferritin, and cases are ordered from lowest to highest ferritin.

followed by lasso regression (using the model coefficients trained on the training data as described previously). In four of these 26 cases, predicted ferritin was less than 30 ng/mL, and the measured ferritin was more than an order of magnitude higher. Thus, in these four cases, the predicted ferritin was concerning for iron deficiency while the actual ferritin was not, and we reviewed these cases in detail. In the other 22 of the 26 cases, the predicted ferritin was greater than 30 ng/mL, so the discrepancy was less likely to indicate a falsely reassuring measured ferritin result, and so we focused our review on the four cases in which the predicted ferritin was less than 30 ng/mL.

One of these four patients (patient 1; see supplemental case studies for additional case information for all four patients) almost certainly had iron deficiency based on review of the medical record by two pathologists, despite having a measured ferritin well within the normal range. In this case, the predicted ferritin was almost certainly more reflective of the patient's iron status. In one other case (patient 2), the patient had recently completed a course of intravenous iron infusions and was recovering from iron deficiency. In this case, the predicted ferritin may have better reflected the patient's irons stores or may have at least provided an indication that the patient's underlying iron status was not entirely normal. The third patient (patient 3) had only two predictor tests available, and thus the algorithms had limited data to use in prediction. Patients lacking sufficient predictive information, such as patient 3, would likely be excluded from future decision support algorithms. The fourth patient (patient 4) had a history of iron deficiency and, although most likely was not iron deficient at the time of the testing, had a complex hematologic picture (see case description in the supplement). Thus, even in this fourth case where the measured ferritin probably better reflected the patient's current

DOI: 10.1093/ajcp/aqw064

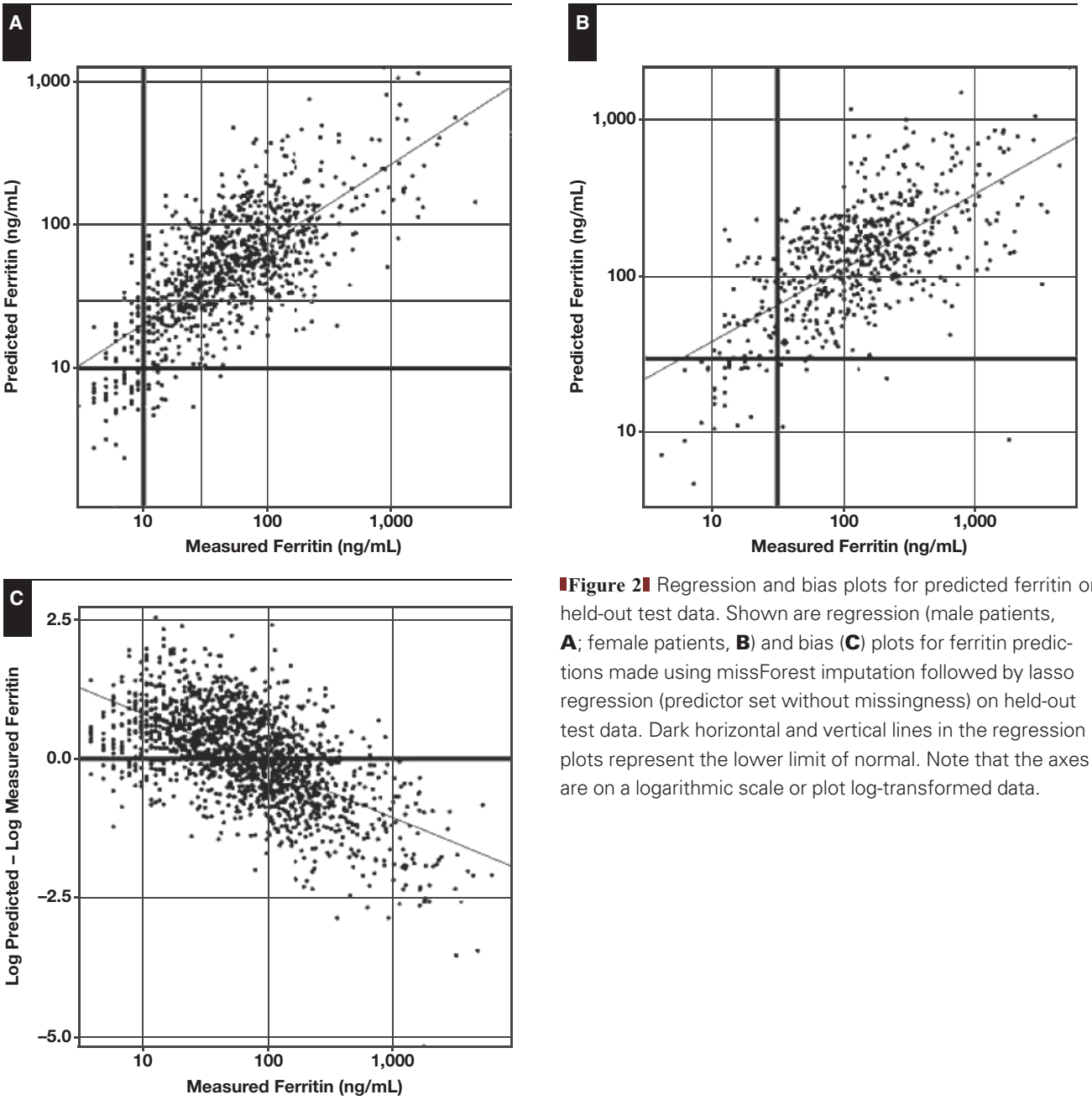


Figure 2 Regression and bias plots for predicted ferritin on held-out test data. Shown are regression (male patients, **A**; female patients, **B**) and bias (**C**) plots for ferritin predictions made using missForest imputation followed by lasso regression (predictor set without missingness) on held-out test data. Dark horizontal and vertical lines in the regression plots represent the lower limit of normal. Note that the axes are on a logarithmic scale or plot log-transformed data.

iron status, the predicted ferritin may have provided an important indication that the case was more complex and deserved more consideration than the measured ferritin alone might have suggested.

Discussion

These findings support our hypothesis that in most cases, ferritin provides information that is substantially redundant given other available test results. In particular, we find that coexisting data can discriminate normal from abnormal ferritin results with a high degree of accuracy with AUCs as high as 0.97. Predictions of numerical ferritin results were moderately accurate. We show that in at least certain cases, predicted ferritin may better represent a patient's underlying iron deficiency status. Further clinical validation of predicted ferritin with a larger case review will also be an important consideration for future research.

While our approach provides the potential to discover previously unsuspected or unknown associations between disparate elements of laboratory data, the predictability of ferritin was largely unsurprising. As an acute phase reactant, ferritin is known to increase with other inflammatory markers.[6] Likewise, ferritin will decrease with other markers of iron deficiency such as mean cell volume. More

DOI: 10.1093/ajcp/aqw064

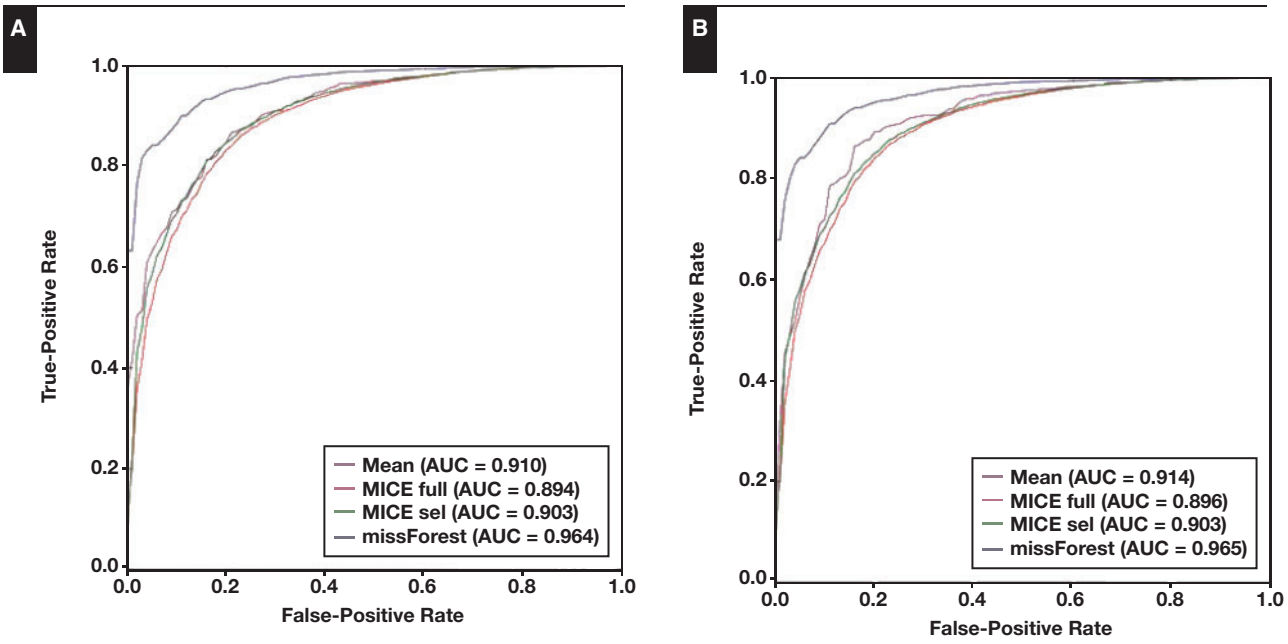


Figure 3 Receiver operator characteristic (ROC) curves for ferritin classification on held-out test data, plotting sensitivity of imputation followed by logistic regression in identifying abnormal results as a function of false-positive rate (1 – specificity). Curves are based on performance in the held-out test data set and are provided for the predictor set without missingness (**A**) and the predictor set with missingness (**B**). Curves are denoted by the imputation method used prior to logistic regression.

generally, our results most likely reflect that fact that various physiologic and pathologic states tend to affect (and thus be reflected by) multiple analytes. However, traditional "manual" approaches to test result interpretation lack a method for quantitatively applying this knowledge. Thus, much of the value of our approach lies in the framework it offers for the automated application of such knowledge to test result integration. Nonetheless, our statistical techniques also offer value in identifying subtle and previously unknown associations between analytes and incorporating these into test result integration models.

We anticipate several potential applications of these findings to novel types of clinical decision support. The first application would be to use predicted ferritin to flag patient ferritin results that are anomalous or otherwise misrepresentative of a patient's underlying iron status. In particular, our case analysis suggests that at least some ferritin results that are much higher than predicted may not accurately reflect patient iron status. We are currently planning work to evaluate and validate options for decision support flags that would append a comment to certain ferritin results that are substantially higher than predicted. In cases such as patient 1, this type of flag may lead to more timely identification of iron deficiency. Since iron deficiency can be a sign of serious underlying illness such as gastrointestinal cancer[11] or, if untreated in pediatric patients, can lead to developmental delay,[12] timely identification is of key importance.

Another potential application to decision support would be to alert clinicians to patients in whom ferritin is predicted to be low based on available tests but where a measured ferritin had not been ordered and iron deficiency may not be suspected. These patients may benefit from an iron deficiency evaluation. We plan additional work to validate predicted ferritin for this purpose, noting that potential applications to clinical decision support remain hypothetical pending further validation. Finally, we will need to understand better the optimal strategy to fit models like this for use at other institutions. A future clinical application of the model at other sites might involve other institutions applying the same methods and approach described here but training the models on their own institution's data. Likewise, we could envision the possibility of training models on pooled data across multiple institutions.

Finally, the idea that the diagnostic information offered by ferritin often duplicates that provided by other diagnostic tests suggests a notion of "informationally" redundant testing. In many patients, it is likely that a diagnosis of iron deficiency could be confirmed or excluded without some of the tests often ordered in current practice. We speculate that informationally redundant testing occurs in a variety of diagnostic settings and diagnostic workups and is much more frequent than the more traditionally defined and narrowly framed notion of redundant testing, which most often just includes unintended duplications of the same or similar

DOI: 10.1093/ajcp/aqw064

Table 2
Univariate Associations Between Ferritin and Predictor Tests[a]

Analyte	Correlation With Ferritin	R^2	AUC
AbsBaso	0.05	0.00	0.54
AbsEos	0.03	0.00	0.52
Albumin	–0.11	0.01	0.51
AlkPhos	0.08	0.01	0.51
AbsLymph	0.02	0.00	0.52
AbsMono	0.10	0.01	0.51
AbsNeut	0.05	0.00	0.52
B12	0.12	0.02	0.59
%Baso	0.02	0.00	0.57
Ca	0.03	0.00	0.58
Eos	0.03	0.00	0.53
Fe	0.33	0.11	0.79
Glob	0.05	0.00	0.56
Hct	0.07	0.00	0.68
Hb	0.15	0.02	0.72
%Lymph	–0.05	0.00	0.51
MCH	0.43	0.19	0.84
MCHC	0.32	0.11	0.75
MCV	0.39	0.15	0.83
%Mono	0.06	0.00	0.54
%Neut	0.01	0.00	0.51
%NucRBC	0.12	0.02	0.51
NucRBC	0.12	0.02	0.51
Anion	–0.01	0.00	0.53
BUN	0.26	0.07	0.55
Cl	–0.07	0.01	0.55
Cr	0.26	0.07	0.56
Glu	0.13	0.02	0.50
K	0.10	0.01	0.52
Bicarb	0.05	0.00	0.55
Plt	–0.18	0.03	0.61
Na	–0.05	0.00	0.50
RBC	–0.17	0.03	0.56
RDW	–0.09	0.01	0.73
AST	0.21	0.04	0.59
ALT	0.20	0.04	0.63
TBILI	0.22	0.05	0.60
TIBC	–0.62	0.38	0.85
Prot	–0.01	0.00	0.55
WBC	0.06	0.00	0.52

For definitions of all analytes, see Table 1.
[a]Shown are the correlation and percentage of variance explained (R^2) between each analyte and ferritin (log transformed). Also shown is the area under the curve (AUC) describing the discriminative power of each individual analyte to predict whether ferritin would be abnormal. Pairwise complete cases were used; missing results were excluded from analyses.

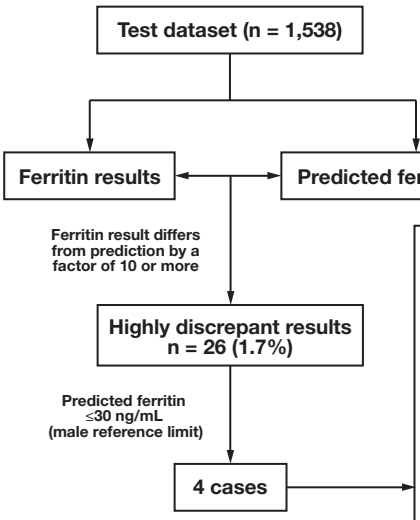
tests.[13,14] Redundant laboratory testing under this narrow definition is estimated to waste more than $5 billion annually in the United States,[14] an amount potentially dwarfed by the waste from informationally redundant testing. Nonetheless, since ferritin and most of the other predictor tests used in this study are generally performed on automated instruments with minimal analytic-phase variable costs, eliminating a ferritin or a small number of other tests from the workup without eliminating entire specimen tubes or patient collections may only lead to a small reduction in laboratory cost on any given patient.[15]

We plan to build on this proof of concept to apply similar approaches to a wide range of other pathology data, including anatomic pathology results and other clinical laboratory analytes. For example, we hypothesize that some immunohistochemistry results may be predictable based on specimen morphology, clinical characteristics, and other immunohistochemistry results. Future decision support systems could potentially provide clinicians a list of tests predicted to be normal and predicted to be abnormal at some specified confidence level; the clinician would then primarily select tests to order from a list with highly uncertain predictions. Likewise, we suspect that temporal trends may be highly informative in predicting certain test results.

Selecting predictor tests ("feature selection") for this type of analysis requires certain trade-offs. For example, at one extreme, we could have used as predictors only those tests that are most frequently performed alongside ferritin, minimizing missing data but at the expense of having a less complete feature set. The other extreme would be to include an expansive set of predictor tests, leading to a sparser data set with more missing elements. We sought a balance between these extremes. For many applications, optimal predictor test selection will represent an empirical question, guided largely by model performance. Practical considerations related to model deployment (intended application) may factor into feature selection. Another consideration is that including too many predictors relative to the number of cases in the training data set will make models more prone to overfitting. Nonetheless, including an expansive feature set with a high rate of missing data should not be problematic for many applications so long as the models perform acceptably on held-out test data. In the current study, the performance of the selected imputation-regression combination (misForest-Lasso) was only slightly better in the training data compared with the test data (see Supplemental Table 1), indicating that overfitting, while present, was minimal. Although we included a large number of predictors in the lasso regression (the model was permitted to fit nonzero coefficients for all 40 predictors, and most coefficients had absolute values $\geq$0.001), overfitting was likely controlled by the comparably large number of training cases and the fact that lasso has a built-in regularization procedure intended to control overfitting (see supplemental methods). Likewise, given that missForest imputation paired with logistic regression was able to classify test data with a high AUC of 0.97, overfitting in this classification model was likely well controlled. Assessing the ability of various subsets of predictor tests to predict ferritin represents an important topic for future work.

In this study, MICE and missForest imputation performed better than mean imputation. Our results support findings by Waljee et al[16] that missForest tends to perform well in imputing missing laboratory results. We speculate

DOI: 10.1093/ajcp/aqw064



Case	Ferritin, ng/mL	Predicted Ferritin, ng/mL	Impression	Comment
1	230	21	Iron deficiency, not clinically identified	Ferritin increased secondary to inflammation
2	197	19	Recovering iron deficiency	Receiving IV iron therapy
3	1,768	9	Limited predictive data	Only two predictor tests available; decision support will likely require a minimum number of predictor tests
4	364	20	Complex hematologic picture	Referral to hematology would have likely been useful had the resting been ordered by a nonspecialist

Figure 4 Clinical review of selected cases. Shown is a summary of the case review. As shown, predicted ferritin may sometimes be a better indicator of a patient's underlying iron status than measured ferritin. See the text and supplemental results for additional clinical detail regarding each case.

that missForest performs well because it best accommodates nonlinear relationships and interactions between the predictor data and the ferritin results.

Although it is difficult to directly compare regression performance with classification performance, the high degree of ferritin classification accuracy may appear somewhat discordant with the only moderate regression performance. One explanation for this apparent discrepancy may be that predictions are most accurate toward the middle of the dynamic range of measured ferritin results (Figure 2C), presumably due to "regression toward the mean." Some of the numerical prediction error may thus be attributed to measured ferritin values toward the high end of the range with predicted values of ferritin that are considerably lower but that are still within the reference range. Furthermore, the regression algorithms must train to minimize error throughout the dynamic range, while classification algorithms must only minimize error at the specific classification threshold.

A potential limitation of this approach is that the imputation techniques are only unbiased under the assumption that the data are missing at random. Real clinical practice surely violates this assumption since clinicians usually order tests given some expectations about the likely results. Nonetheless, including in our ferritin prediction models a set of dichotomous variables to indicate whether each predictor test result was imputed (missing) or measured (ie, using the "predictor set with missingness") added little benefit in terms of improved ferritin prediction performance. The lack of predictive information gained by including these additional dichotomous variables suggests that imputation bias may have only a minimal impact on ferritin prediction. Another caveat is that this approach provides only a lower limit for the level of information redundancy in that different machine-learning algorithms could provide different results. Thus, true informational redundancy may be higher than that demonstrated here. Finally, our approach measures predictability of ferritin given those tests ordered in real clinical practice rather than potential information redundancy given a large complete set of test results. Nonetheless, since many decision support strategies must rely on the data available, prediction performance on real data is likely to be most relevant. While a limitation in our data extraction technique treated test results with a numerical value of zero as missing (other than nucleated RBCs), we expect that the impact on our overall analysis should be minimal. This is because most analytes and all key analytes have a minimum reportable and/or physiologic limit greater than zero, and thus this limitation should have affected few results. Furthermore, by slightly increasing the rate of missing data, this limitation would if anything lead our analysis to underestimate the predictability of ferritin.

In conclusion, we show that ferritin results are predictable given other concurrent test results. This suggests that common sets of laboratory results may contain substantial information redundancy. More broadly, this work provides a framework for a novel type of clinical decision support, which, with additional validation and refinement, we hope to implement for a variety of analytes.

DOI: 10.1093/ajcp/aqw064

Corresponding author: Jason Baron, MD, GRJ-239C, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114; jmbaron@partners.org.

This work was supported by a grant from the MGH-MIT Strategic Partnership.

References

1. Louis DN, Gerber GK, Baron JM, et al. Computational pathology: an emerging definition. *Arch Pathol Lab Med*. 2014;138:1133-1138.
2. Baron JM, Dighe AS, Arnaout R, et al. The 2013 symposium on pathology data integration and clinical decision support and the current state of field. *J Pathol Inform*. 2014;5:2.
3. Baron JM, Cheng XS, Bazari H, et al. Enhanced creatinine and estimated glomerular filtration rate reporting to facilitate detection of acute kidney injury. *Am J Clin Pathol*. 2015;143:42-49.
4. Baron JM, Mermel CH, Lewandrowski KB, et al. Detection of preanalytic laboratory testing errors using a statistically guided protocol. *Am J Clin Pathol*. 2012;138:406-413.
5. Matheny ME, Ohno-Machado L. Generation of knowledge for clinical decision support: statistical and machine learning techniques in clinical decision support. In: Greenes R, ed. *The Road Ahead*. Boston, MA: Elsevier Academic Press; 2007:227-248.
6. Guyatt GH, Oxman AD, Ali M, et al. Laboratory diagnosis of iron-deficiency anemia: an overview. *J Gen Intern Med*. 1992;7:145-153.
7. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006.
8. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1-67.
9. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28:112-118.
10. Pedregosa F, Varoquaux G. Gramfort GA, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
11. Goddard AF, James MW, McIntyre AS, et al. Guidelines for the management of iron deficiency anaemia. *Gut*. 2011;60:1309-1316.
12. Yager JY, Hartfield DS. Neurologic manifestations of iron deficiency in childhood. *Pediatr Neurol*. 2002;27:85-92.
13. Bates DW, Kuperman GJ, Rittenberg E, et al. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. *Am J Med*. 1999;106:144-150.
14. Jha AK, Chan DC, Ridgway AB, et al. Improving safety and eliminating redundant tests: cutting costs in U.S. hospitals. *Health Aff*. 2009;28:1475-1484.
15. Huck A, Lewandrowski K. Utilization management in the clinical laboratory: an introduction and overview of the literature. *Clin Chim Acta*. 2014;427:111-117.
16. Waljee AK, Mukherjee A, Singal AG, et al. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*. 2013;3:e002847.

DOI: 10.1093/ajcp/aqw064