AMIA

INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

Use of machine learning to predict clinical decision support compliance, reduce alert burden, and evaluate duplicate laboratory test ordering alerts

Jason M. Baron[1,2], Richard Huang[1,2], Dustin McEvoy[3], and Anand S. Dighe[1,2,3]

[1]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA, [2]Havard Medical School, Boston, Massachusetts, USA and [3]Partners eCare, Partners HealthCare System, Somerville, Massachusetts, USA

Corresponding Author: Anand S. Dighe, Gray 5 Chemistry, Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA; asdighe@partners.org

ABSTRACT

Objectives: While well-designed clinical decision support (CDS) alerts can improve patient care, utilization management, and population health, excessive alerting may be counterproductive, leading to clinician burden and alert fatigue. We sought to develop machine learning models to predict whether a clinician will accept the advice provided by a CDS alert. Such models could reduce alert burden by targeting CDS alerts to specific cases where they are most likely to be effective.

Materials and Methods: We focused on a set of laboratory test ordering alerts, deployed at 8 hospitals within the Partners Healthcare System. The alerts notified clinicians of duplicate laboratory test orders and advised discontinuation. We captured key attributes surrounding 60 399 alert firings, including clinician and patient variables, and whether the clinician complied with the alert. Using these data, we developed logistic regression models to predict alert compliance.

Results: We identified key factors that predicted alert compliance; for example, clinicians were less likely to comply with duplicate test alerts triggered in patients with a prior abnormal result for the test or in the context of a nonvisit-based encounter (eg, phone call). Likewise, differences in practice patterns between clinicians appeared to impact alert compliance. Our best-performing predictive model achieved an area under the receiver operating characteristic curve (AUC) of 0.82. Incorporating this model into the alerting logic could have averted more than 1900 alerts at a cost of fewer than 200 additional duplicate tests.

Conclusions: Deploying predictive models to target CDS alerts may substantially reduce clinician alert burden while maintaining most or all the CDS benefit.

Key words: machine learning, clinical decision support (CDS), alert fatigue, alert burden, duplicate laboratory test, practice variation

LAY SUMMARY

Health systems often build clinical decision support (CDS) alerts into their electronic health records systems to help clinicians with decisions such as which laboratory tests to order. A common problem with CDS is that clinicians often see so many alerts that they begin to reflexively overlook them, a phenomenon known as alert fatigue. In this article, we looked at alerts designed to notify clinicians when they attempt to order a laboratory test on a patient who already had the test recently. The so-called "duplicate" test orders subject to these alerts were often, but not always, placed in error. We developed a machine learning model to predict for each alert that fired whether the clinician seeing it would accept the alert (ie, remove the duplicate test order) or conversely override it. We postulated that this model could be used to determine when to fire an alert. Using our proposed approach, the electronic health record would only display alerts likely to be accepted, thereby reducing the number of alerts that clinicians see ("alert burden") with minimal impact on the number of duplicate laboratory tests ordered. We simulated the impact of our model and proposed implementation strategy and show that it has substantial potential utility.

INTRODUCTION

Clinical decision support (CDS) is among the most powerful tools health systems can deploy to optimize clinical decision-making, utilization management, and population health.[1–14] Hospitals will often build CDS within their electronic health record (EHR) systems to support laboratory and radiology test ordering and medication prescribing.[2–4,7–12,15] For example, common pharmacy-related CDS targets include alerts around critical drug-drug interactions and proper dosing of medications in the setting of renal insufficiency.[5,16] Likewise, CDS may provide warnings when a clinician is attempting to order a laboratory test that is inappropriate for the patient[2,3,8,11–13,17–21] or may assist clinicians in selecting the most appropriate imaging study for their patient.[4] Other common CDS targets include reminders regarding preventative care and notifications regarding potentially overlooked diagnoses.[22,23]

While CDS has enormous value, health systems must deploy CDS judiciously; every CDS intervention has costs in terms of clinician alert burden.[2,3,9,24–26] Excessive CDS alerts, and especially irrelevant ones, tend to lead to clinician annoyance and distract precious time and attention away from critical patient-care activities.[2,3,9,24–26] Eventually, clinicians, when faced with an endless stream of alerts, will simply start to ignore alerts altogether, even those related to critical patient-safety issues in a phenomenon described as "alert fatigue."[24–26] A common approach to combating alert fatigue is to minimize use of CDS and not build alerts around many important, yet not top priority, clinical situations. However, simply avoiding use of CDS, while perhaps effective in combatting alert fatigue, has the obvious drawback that it limits the important benefits of CDS. In this regard, informaticians and health systems often view CDS implementation decisions in terms of a delicate tradeoff between alert burden and CDS benefit.

In the initiative described in this manuscript, we considered that a CDS alert that fires, but the clinician ignores, is essentially "wasted"; such an alert produces costs in terms of alert burden without a corresponding benefit in terms of a change in patient management. Eliminating these wasted alerts would reduce alert burden without trading off benefit. We hypothesized that we could develop machine learning models to predict whether a clinician is likely to accept or override a given alert. Such models might then be applied to suppress alerts likely to be ignored (and only fire on those likely to be accepted) thereby reducing alert burden while maintaining most of the alert benefit.

To develop such an approach, we considered alerts surrounding duplicate laboratory test orders.[11,12,17,18] As the name implies, a "duplicate" test occurs when the same test is repeated in the same patient across a time interval that is too short to provide meaningful new information.[17] The acceptable time interval for repeating a test varies by clinical circumstance and by analyte and may range from a matter of minutes in select circumstances for blood gases to never for germline genetic tests. Duplicate testing is common, particularly when patients are seen by multiple clinicians, who may each order the same test, not knowing about their colleagues' orders.[27] Collection, performance, and interpretation of duplicate laboratory tests wastes laboratory, nursing, and physician resources.[28] Duplicate testing can also lead to false-positive findings and necessitate additional unneeded downstream testing.[28] Most computerized provider order entry (CPOE) systems include the capacity for duplicate alerting but these systems may not be used due to the aforementioned risk of alert burden and alert fatigue.[25,29–32]

In this report, we specifically considered duplicate laboratory test alerts implemented across 8 hospitals and 115 different tests. Using performance data from these alerts, we developed a series of machine learning models to predict, based on clinical context, whether a clinician will comply with the alert. Finally, we simulated how predictive models might be used to drive "smart" alerting protocols intended to reduce alert burden while maintaining most of the alerting benefit.

MATERIALS AND METHODS

Setting

We implemented duplicate test alerts (as described in detail below) within the EHR at Partners HealthCare and monitored alert performance during a study period consisting of 8 consecutive months in 2019. Partners is a multi-hospital, not-for-profit healthcare system. We specifically evaluated data from 8 Partners HealthCare hospitals (2 academic medical centers: Brigham and Women's Hospital and Massachusetts General Hospital; and 6 regional/community-based hospitals: Brigham and Women's Faulkner, North Shore Medical Center, Newton Wellesley Hospital, Cooley Dickinson Hospital, Martha's Vineyard Hospital, and Nantucket Cottage Hospital). The Partners HealthCare system used Epic version 2018 (Epic Clinical Systems, Verona, WI) as its EHR system during the study period.

This project was completed as a quality improvement initiative, and, as such, formal review by the institutional review board (IRB) was not required, per IRB protocol.

Table 1. Predictor variables

Predictor variable	Definition/derivation
Hospital type (academic vs community)	Extracted the hospital associated with each alert and grouped hospitals as academic (2 hospitals) or community (6 hospitals).
Department specialty (eg, neurology)	The specialty of the clinician seeing the alert as defined in the EHR.
Specialty age (pediatric vs adult)	Applied a custom grouper to the department specialties. Based on the clinician's primary specialty, not necessarily the patient age.
Provider type	The type of provider seeing the alert. Used definitions from our EHR and grouped into the categories attending, NP/PA, trainee, and "other."
Prior result abnormal (yes or no)	Based on whether the prior results for the test on which the alert fired was within the laboratory's normal reference range.

EHR: electronic health record; NP: nurse practitioner; PA: physician assistant.

Duplicate test alerts

Within the EHR, we implemented interruptive alerts that prompted providers to discontinue orders for testing considered to be duplicative using standard Epic functionality. We defined test-specific "lookback" periods (Supplementary Table S1) for 115 different tests based on laboratory and clinician input. Alerts were set to fire when a clinician selected a laboratory order for a test on a patient that had a prior result for the same test within the look period. The trigger for alert firing was the selection of a test order, either from the results of a search or from selection from a predefined menu or order set item. For example, the hemoglobin A1c duplicate check would fire if a provider selected a hemoglobin A1c order on a patient with a prior hemoglobin A1c within 30 days (30 days is the hemoglobin A1c lookback period). These alerts were implemented as "soft stops" as the alerts gave users the option to accept the alert's advice and discontinue the order or to bypass the alert and continue with the order. A representative screenshot of an alert is provided as Figure 1.

Key characteristics surrounding the order, clinician, and patient that triggered each alert (Table 1) and whether the alert was ultimately accepted or overridden were derived from data extracted from Epic using Epic Clarity reporting functionality. Although we initially reviewed performance of alerts implemented across 115 different tests, we focused most of our analysis on 5 key tests: hemoglobin A1c, ferritin, vitamin B12, 25 hydroxy vitamin D, and TSH. We selected these 5 tests based on alert frequency, acceptance rate, and a desire to include tests ordered across a range of settings. We analyzed outpatient alerting separately from inpatient/emergency department (ED) alerting (but considered inpatient/ED alerts together) since outpatient care pathways and workflows tend to be quite different from those in the inpatient and ED settings.

Factors impacting alert acceptance

We used logistic regression to interrogate the impact of key clinical and contextual predictor variables on alert compliance. For this analysis, we trained separate models for each test.

Predictive models

For each of the 5 key tests, we randomly split the outpatient data (excluding cases where a future expected date were set as discussed further in the results) into training and testing partitions in an approximately 80:20 ratio. Using the training partition, we trained logistic regression models and evaluated model performance using the testing partition. We did not build predictive models for the inpatient/ED dataset since the compliance was high at baseline.

Impact of individual provider practice patterns

To evaluate whether differences in practice patterns between clinicians impacted alert compliance, we pooled outpatient data (excluding orders with an expected date set) across the 5 key tests. We then trained 2 additional logistic regression models using the training partition of this pooled data. One model, the "provider-independent model" used only the predictors from Table 1. The other model, the "provider-specific" model, included the specific provider as a feature, in addition to the predictors from the provider-independent model (including the provider specialty). Thus, other factors being equal, the provider-specific model would predict a lower likelihood of alert compliance for a clinician who is generally less compliant with alerts (and vice versa). The model would assess whether a clinician is "generally" more or less likely to comply with alerts based on the clinicians' compliance in the training data, after adjusting for other variables.

More specifically, the provider-specific model treated providers as a fixed effect and thus the model assigned a unique intercept to each clinician. The models treated each providers' individual impact as constant across all 5 tests. The impact of all other predictors (eg, those in Table 1) was allowed to vary by test (ie, a separate coefficient was included for each test-predictor combination; each test was also allowed a test-specific intercept, but providers were only allowed a single intercept adjustment across all tests). The provider-independent model was structurally identical, except for the absence of provider-specific effects. The provider-specific model can be described symbolically in Equation 1 as provided in the Supplementary information.

Providers who saw fewer than 20 alerts within the training data were assigned to a generic category of "other" and were not in-
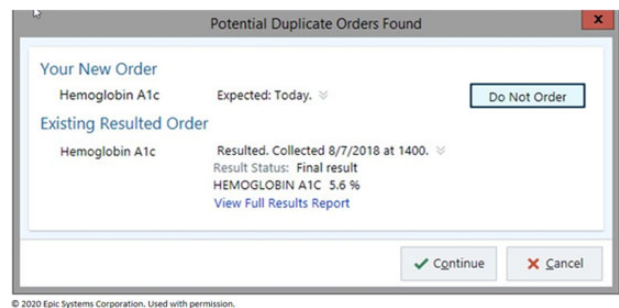

Figure 1. Screenshot of duplicate test alert "soft stop." The prior test result value and date are shown, and the provider can decide to continue with the duplicate order or cancel the order. Permission to use this image was obtained from Epic Systems Corporation.

cluded individually in the model. We chose to pool data across tests for this analysis to obtain meaningful data regarding providers who may not have seen the alert very many times for any given test but did see the alert across tests a sufficient number of times.

Alert burden reduction

We simulated what would happen if we applied trained models to determine whether to fire the alert. We assumed a model application in which the alert would fire only if (1) the order was a duplicate (per the criteria used in this study) and (2) the model predicted that the clinician would accept the alert with a likelihood greater than a predetermined probability threshold. By varying this probability threshold, we considered both the additional duplicate tests that would likely be ordered and the reduction in alert burden compared to alerting criteria implemented during this study. We assumed that the duplicate test would be performed if the alert did not fire (ie, the model had predicted noncompliance was likely) and considered actual alert acceptance data to evaluate which duplicate tests were prevented by the alert.

Predictor category groupings

Within categorical predictor variables, specific categories occurring in fewer than 2.5% of the cases in the training data (or all data in the analysis of predictor variable importance) were grouped into a generic "other" category. For example, specialties seeing an alert for a given test very few times were grouped together as "other".

Data analysis

Data analyses and statistical evaluations were performed using the R statistical scripting language[33] with initial data preprocessing performed using Microsoft Access and Excel. Logistic regression mod-els were generated using functionally in the R stats package. We did not apply any L1 or L2 penalties to the logistic regression because we wanted our coefficients and models to be maximally interpretable (and we further found that regularization would not have been particularly useful given the general concordance between training and testing performance). Alert compliance rates were compared using the Fisher exact test. Models were evaluated using area under the ROC curve (AUroC) and area under the precision-recall curve (AUprC). Binomial confidence intervals were calculated using the Wilson method as implemented in the R "binom" package. Plots were generated using the R ggplot2 package.[34]

RESULTS

Alert characteristics

With input from laboratory leadership and following clinician review, we defined health system-wide duplicate alerts for 115 individual laboratory orders as discussed in the Materials and Methods section and in Supplementary Table S1. A screenshot from a representative alert is provided as Figure 1.

Across the 115 tests, the duplicate test alerts triggered 60 399 times during the 8-month study period across the 8 hospitals. Overall, 21.5% of the duplicate alerts were in the inpatient/ED setting (11 027 inpatient and 1935 ED alerts out of the 60 399 total alerts). 78.5% of the alerts (47 437 out of 60 399) were in the outpatient settings.

The overall compliance for the inpatient/ED duplicate order alerts was 66.9% (95% CI, 66.1%–67.7%; 8670 orders removed following 12 962 alerts). The overall compliance for the outpatient duplicate order alerts was significantly lower ($P < 10^{-6}$) at 12.7% (95% CI 12.4%–13.0%; 6038 orders removed following 47 437 alerts). The frequency of firing and the compliance rates for the top
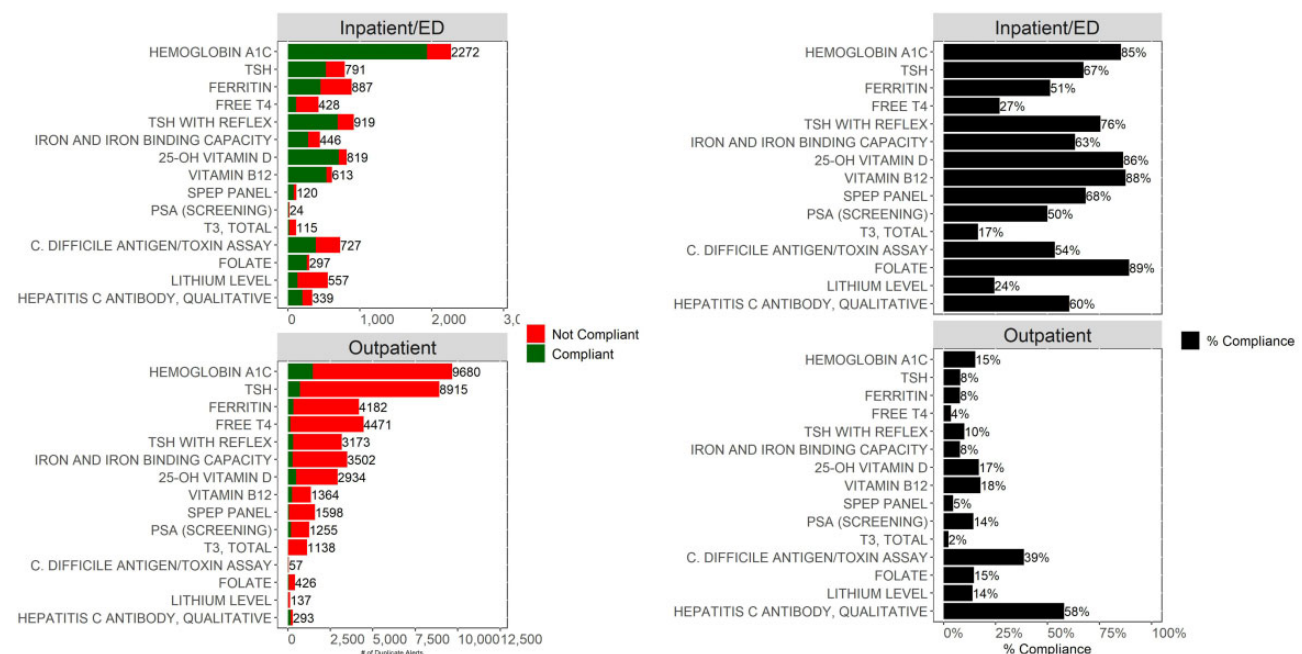

Figure 2. Shown are the total number of duplicate alerts by test for outpatient encounters (bottom left) and inpatient/ED encounters (top left). Bar colors indicate whether the clinician was compliant with the alert (did not proceed with the order) or was noncompliant. Percent compliance for outpatient (bottom right) and inpatient/ED encounter (top right) represents the proportion of alerts for which the clinician did not proceed with the order. This analysis demonstrates that most duplicate alerts are displayed on outpatient encounters; moreover, alerts displayed on outpatient encounters have a much lower compliance rate. The tests included in this figure represent the 15 tests with the most duplicate alerts across both inpatient/ED and outpatient encounters. ED: emergency department.
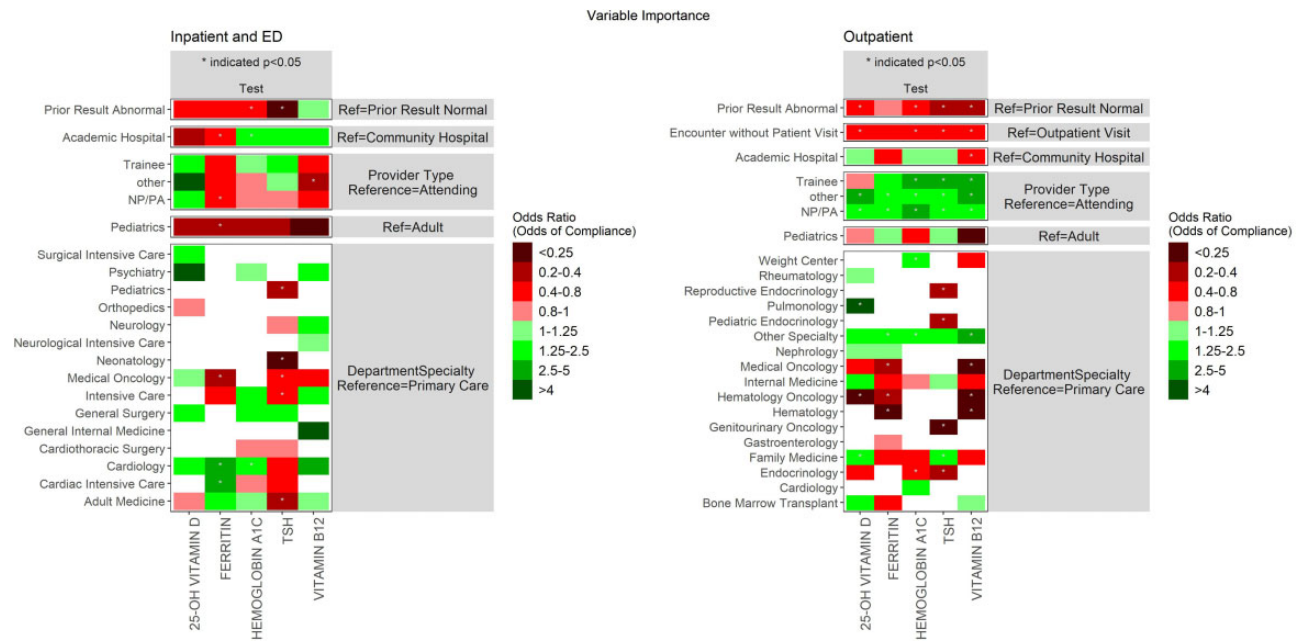
Figure 3. Shown are odds ratios for each predictor variable based on logistic regression models (separate models trained for each test). Cell colors indicate the odds ratio; shades of green indicate the predictor is associated with greater alert compliance (in comparison to the reference level shown to the right), while shades of red indicate associations with lower compliance. White cells indicate that the specified predictor category was not represented in enough alerts to be included in the model.

15 inpatient/ED alerts and the top 15 outpatient alerts by volume are shown in Figure 2. In addition, Supplementary Table S2 provides a comparison of inpatient/ED to outpatient alerting for 5 key tests. Given that outpatient alerts were generally ignored and given that the outpatient setting represented the majority of the total duplicate test alert burden, we decided to focus our subsequent analysis primarily on outpatient alerting.

Impact of future expected date

In some outpatient situations, providers may be ordering tests that are not intended to be drawn until a future date. For example, a provider might order a hemoglobin A1c during an office visit for the patient to have collected in 6 months, prior to the patient's next visit. To place an order for future collection within our EHR, the clinician must first select the order and then subsequently add the expected date. Since the duplicate alerts trigger on order selection, the alert would trigger before the clinician has an opportunity to set the expected date. Thus, clinicians intending to set an expected data will often see an alert that is of little use. Indeed, 40.8% of the outpatient alerts (19 335 out of 47 437) fired on orders where the clinician

went on to set an expected date. However, even if we eliminate the alerts from our analysis where the clinician went on to set an expected date, the overall compliance for the remaining outpatient alerts was still only 21.5% (6038 out of 28 102). The 21.5% compliance rate suggests that the alert was likely reducing duplicate testing only modestly, and at a high cost in terms of alert burden. We thus sought to develop a predictive model that could improve outpatient alerting as will be described subsequently.

Factors impacting alert compliance

As a first step toward understanding factors that impact alert acceptance, we trained a series of logistic regression models for 5 key tests (as discussed in the Materials and Methods section). We trained separate models for each test and encounter type. As shown (Figure 3), having a prior abnormal result for the test was uniformly associated with a lower likelihood of accepting the alert (though only statistically significant for some tests). Orders entered on outpatient encounters in which the patient did not have a visit (eg, phone calls) were uniformly associated with a trend toward lower alert compliance that was statistically significant for all tests except ferritin. Trainees, NPs/PAs, and other providers were in most cases more compliant in the outpatient setting compared to attendings. The impact of hospital type (academic vs community) varied by test as shown in Figure 3. Specialist compliance as compared to primary care physician (PCP) compliance varied by test and setting (Figure 3).

Predictive modeling

We trained logistic regression and random forest models for each of the 5 key tests to predict whether outpatient alerts would be accepted or overridden. Our intention was to begin developing an algorithm that could be applied prospectively to determine whether to fire an alert. As described in the Materials and Methods section, we
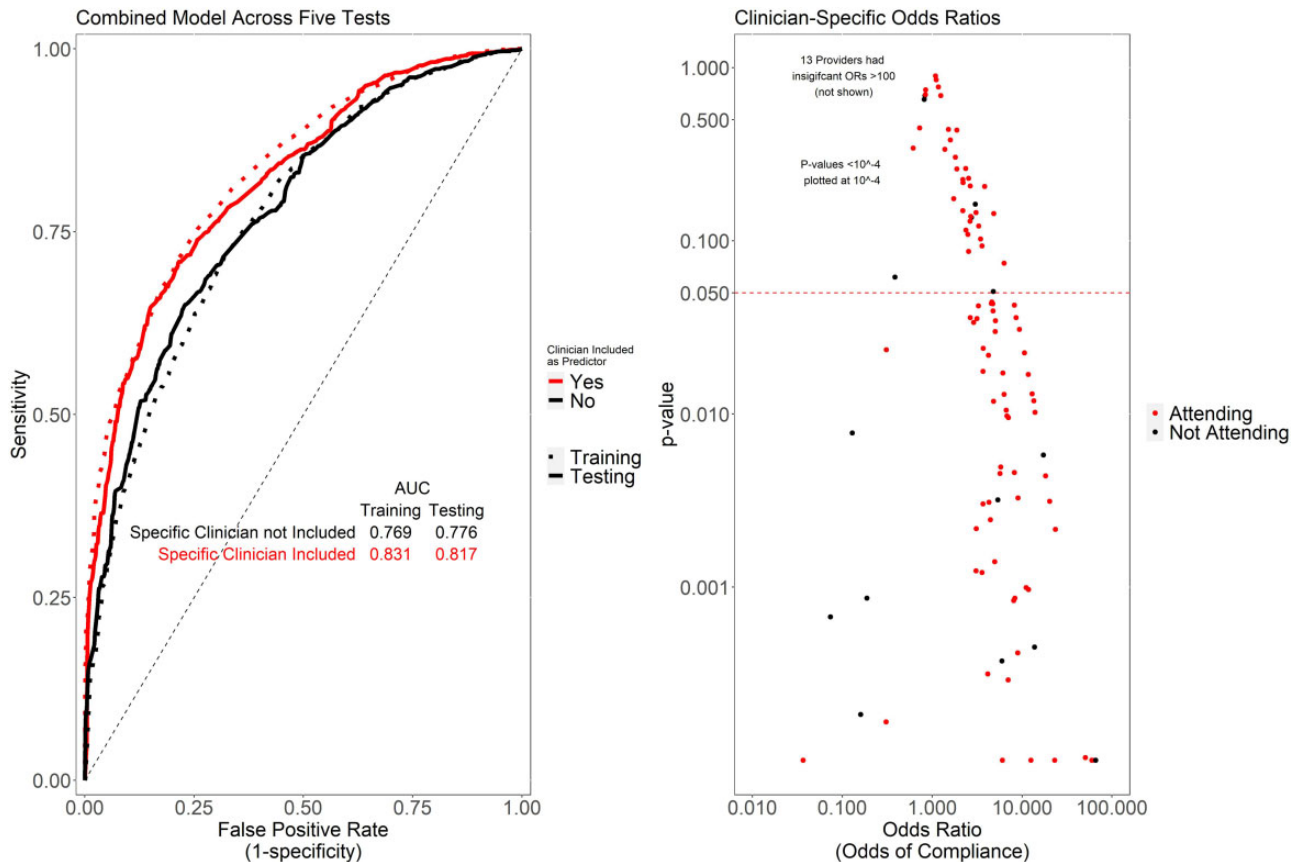
Table 2. Performance of initial outpatient alert compliance models

Test	AUroC		AUprC	
	Training	Testing	Training	Testing
25-OH Vitamin D	0.72	0.68	0.81	0.79
Ferritin	0.75	0.78	0.94	0.94
Hemoglobin A1C	0.70	0.69	0.81	0.81
TSH	0.78	0.78	0.95	0.92
Vitamin B12	0.79	0.76	0.88	0.84

AUroC, area under the receiver operating characteristic (ROC) curve; AUprC, area under the precision-recall curve.

Figure 4. Shown (left) are ROC curves for logistic regression models. The base model ("provider-independent model") includes as predictors general characteristics of the order setting, prior result, department and specialty, paralleling the predictors included in the previously described models. The provider-specific also includes as predictors the specific clinician seeing the alert, for all providers seeing at least 20 alerts. The provider-specific model outperformed the base model on independent test data, suggesting the likelihood of difference between individual providers in alert compliance patterns, even after adjusting for other factors. Shown (right) are the provider-specific odds ratios and corresponding *P*-values for individual providers.

trained these models on a training partition and evaluated using a randomly selected testing partition. The performance of each model for each test is shown in Table 2. The models overall provided moderate discriminative power (AUC 0.69–0.78, testing data) in predicting alert acceptance.

Individual provider impact

Across the training and testing data, 1912 unique clinicians saw the alerts; most of these clinicians saw the alert very few times (medians alerts per clinician = 3; 1–6.25 interquartile range). However, some clinicians frequently encountered this alert; for example, within the training partition, 119 clinicians saw at least 20 alerts. We hypothesized that these clinicians seeing the alert frequently may respond to it differently than other clinicians. We further hypothesized that these clinicians seeing the alert frequently may likewise demonstrate interprovider variation in compliance rates.

To test these hypotheses, we trained 2 additional logistic regression models as discussed in the Materials and Methods section: a "provider-independent model" using only the predictors considered previously and a "provider-specific model" that considered who the specific ordering clinician seeing the alert was for the 119 clinicians who had seen the alert 20 or more times within the training data. (The remaining clinicians were grouped into a reference class of "other".) We trained the models using data from the training data pooled across the 5 tests (but allowed test-specific slopes and inter-

cepts for all predictors besides the specific clinician) and tested it on testing data pooled across the 5 tests.

As shown in Figure 4, the provider-specific model outperformed the provider-independent model (baseline model), suggesting that knowledge of the specific clinician is useful in understanding alert acceptance, even after adjusting for other factors. While we suspect that this is due to differences in practice patterns between clinicians, it is possible that in some cases, the clinician may be serving as a confounder for clinical or contextual variables not otherwise captured in our predictive models. Figure 4 also plots the odds ratio of a given provider against the significance (*P*-value) of that odds ratio being different than 1. This analysis demonstrated that a substantial number of providers are significantly more or less likely to comply with alerts, again suggesting that provider practice patterns exert a significant influence on provider alert acceptance rates.

Supplementary Table S3 provides specific characteristics of the provider-specific model (built using pooled data across the 5 tests) when applied at various probability cutoffs. Overall, this model achieved an AUroC of 0.82 and an AUprC of 0.94 on the testing data.

Achievable alert burden reduction

We analyzed the impact of using our provider-specific model to decide whether to fire the alert. In particular, we considered a strategy in which we only fire the alert if (1) the tests order is a duplicate
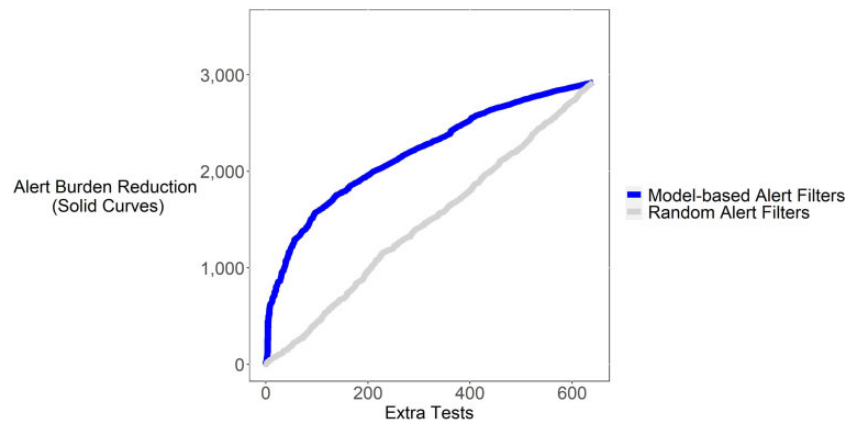
Figure 5. Shown is a plot of the alert burden reduction that could have been achieved by using the provider-specific model to determine whether to fire alerts in relation to additional tests that would have been performed. Paralleling the ROC curve, this analysis assumes alerts would only be fired if the predicted likelihood of compliance exceeded a specified cutoff threshold; we constructed the curve shown by varying the cutoff threshold. For example, if we were to set the sensitivity/specificity of the model to eliminate 1900 alerts (y-axis), this would be expected to result in few than 200 (x-axis) additional tests being ordered. For comparison, we show a similar curve for a model that uses a random number generator to decide whether to fire an alert. The random number model (negative control) as expected achieves poor performance in comparison to the trained model.

according to current duplicate flagging logic and (2) the provider-specific model predicts a substantial likelihood that the clinician will accept the alert. Compared to current state, such an approach would reduce the number of alerts (alert burden) at the expense of some extra duplicate tests getting performed. However, as shown in Figure 5 by applying the model at a suitable sensitivity/specificity threshold, such an approach could substantially reduce the alert firing rate (alert burden) while only modestly increasing the number of tests. As an example, at a "cost" of fewer than 200 extra tests, we could avert more than 1900 alerts within the testing data. (The rate of saved alerts and extra tests would presumably scale 5-fold if implemented prospectively, since it could apply to all alerts, not just the testing partition and would also scale over time.) This suggests that a key to optimizing alerting and combating alert fatigue may not simply be to reduce the number of alerts, but to use "intelligent" models to decide when to fire them.

DISCUSSION

We provide an analysis of CDS alerts related to duplicate laboratory test orders across 8 hospitals and 115 different tests. In addition, we developed logistic regression and random forest-based predictive models to predict whether a clinician will accept or override an alert and illustrate how such models might be applied to a smart alerting strategy intended to reduce alert burden and minimize alert fatigue. We demonstrate overall high levels of alert compliance in the inpatient and ED settings. In the outpatient setting, in contrast, we observed a significant alert burden with variable, but overall poor compliance. Key factors that impacted outpatient alert acceptance varied by test.

The best-performing predictive model we developed (the provider-specific model derived from data pooled across the 5 key tests) achieved an AUroC of 0.82 and an AUprC of 0.94 in determining whether an outpatient alert would be accepted or overridden. Moreover, we show that this model, if applied to determine whether to fire the alert, could lead to a large reduction in alert burden with only a modest number of additional tests. The additional tests may be even fewer than predicted since reducing ineffective alerts could make remaining ones more potent.

Of note, the predictive modeling results described in this manuscript were all based on logistic regression models. Our primary motivation for focusing on logistic regression as opposed to other types of machine learning models was that we wanted models that were interpretable and would be feasible to implement. While we considered a variety of other common supervised machine learning approaches including random forests, support vector machines, and artificial neural networks, these more complex models in many cases appear as a "black box" and can be difficult for end-users to intuit. In contrast, logistic regression coefficients can be easily translated into odds ratios to identify the impact of various factors in predicting whether an alert will be accepted, and the overall logistic regression model can be understood by a person. Likewise, while implementation of more complex machine learning models within the EHR may be theoretically possible, implementing a logistic regression model would generally be much more straightforward since it can be encoded as arithmetic operations and if-then statements. In earlier iterations of our analysis, we had tested use of random forest models. However, the random forest models appeared to offer little if any benefit in terms of improved performance and would have had the limitations noted above.

The observation that a substantial number of individual clinicians are significantly more or less likely to comply with the alerts even after adjusting for key clinical factors is important. We postulate that this interclinician variation represents differences in practice, although we cannot exclude the possibility of confounding with the clinician serving as a surrogate for important clinical or contextual variables. To the extent that the clinician is serving as a surrogate for other unincluded variables, the interclinician variation appearing in our models may represent an overestimate of the true clinician impact. Interclinician variation in practice patterns would be consistent with our prior observations from our institution as well as reports from other hospitals regarding wide interclinician variation in lab test ordering and other clinical care practices.[3,4,35–38] We plan to further explore this variation in future work; if the clinicians who routinely ignore alerts do not have a sound clinical foundation for doing so, it may be that a targeted educational strategy could be more effective than simply exempting these clinicians from seeing the alerts in many cases (as the models would effectively

do). We may also explore as future work the extent to which inclusion of the clinician is simply serving as a surrogate for other variables that were not included in the model. In addition, future work exploring the development of models incorporating a wider range of features such as patient-diagnoses and problem lists may be useful in optimizing model performance.

More generally, we anticipate that this concept of using predictive models to determine alert firing may represent an important new paradigm in combating alert fatigue. Considering that most "online" experiences from internet search to advertisement display are customized to individual users, it may make sense to use similar techniques to optimize interactions between clinicians and their EHRs.

Szymanski et al[39] recently published an analysis of inpatient duplicate laboratory testing alerts with some findings similar to our findings here. For example, they too found that alert compliance varied by test, and provider specialty. They also identified significant predictors of alert compliance including month of the year and overall clinician alert burden, which we did not include in our analysis. However, our analysis expands on the work of Szymanski et al in 2 key ways. First, we focus on outpatient alerts, for which we found compliance to be much lower. Second, we simulate the impact that a prediction model might have on alert burden and test utilization. Likewise, a few prior studies have used machine learning and other analytics to examine alert compliance with the goal of alert suppression and reduction of alert future. For example, Chen et al[40] model compliance with a vaccine reminder alert and implement their model to suppress selected alerts; their study demonstrated promising results. Other prior studies (such as Wong et al[41]) examine factors related to decision support compliance. Moreover, numerous prior works have incorporated machine learning models trained based on a clinical outcome (eg, will the patient develop sepsis) into alerting logic.[42] However, few studies (Chen et al noted above being a rare exception) have explored the development of a predictive model to determine whether to fire a traditional rule-based alert as we do here. To our knowledge, this is the first study to examine this approach in the context of an alert related to laboratory or diagnostic test ordering.

Our study has several limitations. While we considered data across 8 hospitals, they were all within the same health system with each hospital using the same set of alerts and duplicate rules. Future work at other health systems may further support the generalizability of the approach. In addition, while we anticipate that similar approaches may be useful for a variety of CDS alerts related not only to laboratory testing but also many other areas of care, additional work will be needed to adapt and generalize to these other applications. Another consideration is that even our best-performing predictive model was only able to achieve moderate performance (AUroC 0.82); nonetheless, as we show, even this moderate performance could be sufficient to substantially reduce provider alert burden. Our analytic dataset reflected the clinician's response to each alert and not the ultimate set of test orders that were in place at the end of the encounter. Thus, in cases where clinicians accepted the alert and then subsequently reordered the test or conversely rejected the alert and then subsequently canceled the test order, our analysis may not reflect the final test orders. However, we suspect that these cases comprise a small proportion of the total and would not be particularly impactful with regard to the proposed application of our models. Nonetheless, we may as future work quantifies the impact of orders changes made subsequent to alert acceptance or rejection. Finally, while we were able to model the impact of the predictive algorithm-based alerting strategy we developed, future work applying this approach into practice will be needed to fully validate it.

CONCLUSIONS

Taken together, we present this work both to illustrate the factors that impact compliance with CDS for duplicate testing and to demonstrate that a predictive model may enable smart alerting that reduces overall alert burden. We hope and anticipate that future work will confirm and generalize our findings to a wide range of CDS applications. With clinicians facing severe alert fatigue, substantial EHR-related time demands, a high level of overall administration burden and in some cases, burnout, such strategy may be critically important in optimizing patient care, improving clinician efficiency, and enabling better patient and clinician experiences.

FUNDING

AUTHOR CONTRIBUTIONS

Data extraction and/or assembly: JMB, ASD, and DM. Statistical analysis: JMB. Project conception and analytic approach: JMB and ASD. Data interpretation: JMB, RH, and ASD. Manuscript drafting: JMB and ASD. Manuscript revision: all authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

Dr. JMB is a computational pathology consultant for Roche Diagnostics; however, this manuscript does not directly relate to his work with Roche.

DATA AVAILABILITY

The raw patient data underlying this article cannot be shared as doing so would compromise patient privacy and would violate institutional policies and legal requirements.

REFERENCES

1. Bennett P, Hardiker NR. The use of computerized clinical decision support systems in emergency care: a substantive review of the literature. *J Am Med Inform Assoc* 2017; 24 (3): 655–68.
2. Baron JM, Dighe AS. Computerized provider order entry in the clinical laboratory. *J Pathol Inform* 2011; 2 (1): 35.
3. Baron JM, Dighe AS. The role of informatics and decision support in utilization management. *Clin Chim Acta* 2014; 427: 196–201.
4. Weilburg JB, Sistrom CL, Rosenthal DI, *et al*. Utilization management of high-cost imaging in an outpatient setting in a large stable patient and provider cohort over 7 years. *Radiology* 2017; 284 (3): 766–76.
5. Payne TH, Hines LE, Chan RC, *et al*. Recommendations to improve the usability of drug-drug interaction clinical decision support alerts. *J Am Med Inform Assoc* 2015; 22 (6): 1243–50.
6. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. *Yearb Med Inform* 2016; 25 (Suppl 1): S103–16.
7. Monteiro L, Maricoto T, Solha I, Ribeiro-Vaz I, Martins C, Monteiro-Soares M. Reducing potentially inappropriate prescriptions for older

patients using computerized decision support tools: systematic review. *J Med Internet Res* 2019; 21 (11): e15385.

8. Baron JM, Lewandrowski KB, Kamis IK, Singh B, Belkziz SM, Dighe AS. A novel strategy for evaluating the effects of an electronic test ordering alert message: optimizing cardiac marker use. *J Pathol Inform* 2012; 3 (1): 3.

9. Bates DW, Kuperman GJ, Wang S, *et al*. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc* 2003; 10 (6): 523–30.

10. Krasowski MD, Chudzik D, Dolezal A, *et al*. Promoting improved utilization of laboratory testing through changes in an electronic medical record: experience at an academic medical center. *BMC Med Inform Decis Mak* 2015; 15 (1): 11.

11. Procop GW, Keating C, Stagno P, *et al*. Reducing duplicate testing: a comparison of two clinical decision support tools. *Am J Clin Pathol* 2015; 143 (5): 623–6.

12. Procop GW, Yerian LM, Wyllie R, Harrison AM, Kottke-Marchant K. Duplicate laboratory test reduction using a clinical decision support tool. *Am J Clin Pathol* 2014; 141 (5): 718–23.

13. Jackups R Jr. The promise-and pitfalls-of computerized provider alerts for laboratory test ordering. *Clin Chem* 2016; 62 (6): 791–2.

14. 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 (1): 17.

15. Page N, Baysari MT, Westbrook JI. A systematic review of the effectiveness of interruptive medication prescribing alerts in hospital CPOE systems to change prescriber behavior and improve patient safety. *Int J Med Inform* 2017; 105: 22–30.

16. Vogel EA, Billups SJ, Herner SJ, Delate T. Renal drug dosing. Effectiveness of outpatient pharmacist-based vs. Prescriber-based clinical decision support systems. *Appl Clin Inform* 2016; 7 (3): 731–44.

17. Jha AK, Chan DC, Ridgway AB, Franz C, Bates DW. Improving safety and eliminating redundant tests: cutting costs in U.S. hospitals. *Health Aff (Millwood)* 2009; 28 (5): 1475–84.

18. Magid S, Forrer C, Shaha S. Duplicate orders: an unintended consequence of computerized provider/physician order entry (CPOE) implementation: analysis and mitigation strategies. *Appl Clin Inform* 2012; 3 (4): 377–91.

19. Rubinstein M, Hirsch R, Bandyopadhyay K, *et al*. Effectiveness of practices to support appropriate laboratory test utilization: a laboratory medicine best practices systematic review and meta-analysis. *Am J Clin Pathol* 2018; 149 (3): 197–221.

20. Rudolf JW, Dighe AS. Decision support tools within the electronic health record. *Clin Lab Med* 2019; 39 (2): 197–213.

21. Kwon JH, Reske KA, Hink T, Jackups R, Burnham CD, Dubberke ER. Impact of an electronic hard-stop clinical decision support tool to limit repeat Clostridioides difficile toxin enzyme immunoassay testing on test utilization. *Infect Control Hosp Epidemiol* 2019; 40 (12): 1423–6.

22. 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 (1): 42–9.

23. Peiffer-Smadja N, Rawson TM, Ahmad R, *et al.*. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clin Microbiol Infect* 2020; 26 (5): 584–95.

24. Chaparro JD, Hussain C, Lee JA, Hehmeyer J, Nguyen M, Hoffman J. Reducing interruptive alert burden using quality improvement methodology. *Appl Clin Inform* 2020; 11 (1): 46–58.

25. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. *AMIA Annu Symp Proc* 2007; 2007: 26–30.

26. Carroll AE. Averting alert fatigue to prevent adverse drug reactions. *JAMA* 2019; 322 (7): 601.

27. Bazari H, Palmer WE, Baron JM, Armstrong K. Case records of the Massachusetts General Hospital. Case 24-2016. A 66-year-old man with malaise, weakness, and hypercalcemia. *N Engl J Med* 2016; 375 (6): 567–74.

28. Huck A, Lewandrowski K. Utilization management in the clinical laboratory: an introduction and overview of the literature. *Clin Chim Acta* 2014; 427: 111–7.

29. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Med Inform Decis Mak* 2017; 17 (1): 36.

30. Lam JH, Ng O. Monitoring clinical decision support in the electronic health record. *Am J Health Syst Pharm* 2017; 74 (15): 1130–3.

31. Kane-Gill SL, O'Connor MF, Rothschild JM, *et al*. Technologic distractions (part 1): summary of approaches to manage alert quantity with intent to reduce alert fatigue and suggestions for alert fatigue metrics. *Crit Care Med* 2017; 45 (9): 1481–8.

32. Yoshida E, Fei S, Bavuso K, Lagor C, Maviglia S. The value of monitoring clinical decision support interventions. *Appl Clin Inform* 2018; 9 (1): 163–73.

33. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; 2013.

34. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag; 2016.

35. Lipitz-Snyderman A, Sima CS, Atoria CL, *et al*. Physician-driven variation in nonrecommended services among older adults diagnosed with cancer. *JAMA Intern Med* 2016; 176 (10): 1541–8.

36. Shahbazi S, Woods SJ. Influence of physician, patient, and health care system characteristics on the use of outpatient mastectomy. *Am J Surg* 2016; 211 (4): 802–9.

37. Doctor K, Breslin K, Chamberlain JM, Berkowitz D. Practice pattern variation in test ordering for low-acuity pediatric emergency department patients. *Pediatr Emerg Care* 2018; doi: 10.1097/PEC.0000000000001637.

38. Nguyen LT, Guo M, Hemmelgarn B, *et al*. Evaluating practice variance among family physicians to identify targets for laboratory utilization management. *Clin Chim Acta* 2019; 497: 1–5.

39. Szymanski JJ, Qavi AJ, Laux K, Jackups R Jr. Once-per-visit alerts: a means to study alert compliance and reduce repeat laboratory testing. *Clin Chem* 2019; 65 (9): 1125–31.

40. Chen J, Chokshi S, Hegde R, *et al*. Development, implementation, and evaluation of a personalized machine learning algorithm for clinical decision support: case study with shingles vaccination. *J Med Internet Res* 2020; 22 (4): e16848.

41. Wong A, Amato MG, Seger DL, *et al*. Prospective evaluation of medication-related clinical decision support over-rides in the intensive care unit. *BMJ Qual Saf* 2018; 27 (9): 718–24.

42. Baron JM, Kurant DE, Dighe AS. Machine learning and other emerging decision support tools. *Clin Lab Med* 2019; 39 (2): 319–31.