Predicting Low Information Laboratory Diagnostic Tests

Shivaal K Roy,[1] Jason Hom,[2] Lester Mackey,[3] Neil Shah,[4] Jonathan H Chen[2]

1 Department of Computer Science, Stanford University, Stanford, CA, USA;
2 Department of Medicine, Stanford University, Stanford, CA, USA;
3 Microsoft Research New England, Cambridge, MA US;
4 Department of Pathology, Stanford University, Stanford, CA, USA

Abstract

Escalating healthcare costs and inconsistent quality is exacerbated by clinical practice variability. Diagnostic testing is the highest volume medical activity, but human intuition is typically unreliable for *quantitative* inferences on diagnostic performance characteristics. Electronic medical records from a tertiary academic hospital (2008-2014) allow us to systematically predict laboratory pre-test probabilities of being normal under different conditions. We find that low yield laboratory tests are common (e.g., ~90% of blood cultures are normal). Clinical decision support could triage cases based on available data, such as consecutive use (e.g., lactate, potassium, and troponin are >90% normal given two previously normal results) or more complex patterns assimilated through common machine learning methods (nearly 100% precision for the top 1% of several example labs).

Introduction

Unsustainable growth in health care costs is in part due to waste that does not actually improve health.[1] The Institute of Medicine estimates that over two hundred billion dollars a year are spent on tests and procedures that are unnecessary.[2] Given this sobering amount of misallocated resources, there has been an increasing emphasis on high value care, notably with the American Board of Internal Medicine's (ABIM) Choosing Wisely guidelines.[3] Lab testing is an important target for high value care efforts as it constitutes the highest volume medical activity.[4] Prior research suggests that in the inpatient setting, unnecessary labs constitute a quarter to half of all testing.[5,6] Consequently, guidelines regarding high value care and lab testing have been emphasized by numerous professional medical societies in both general inpatient and critical care settings (http://www.choosingwisely.org).

A major barrier to high value care is variability in clinical practice among providers. A meta-analysis of lab testing reports both over- and under-utilization, depending on the clinical setting.[4] When used properly, diagnostic testing yields information that positively impacts decision-making. The consequences of inappropriate ordering are not merely financial. Poorly chosen diagnostics may lead to misinterpreted results and treatment. For example, cardiac stress testing often does not reclassify patients beyond risk assessments achievable from a basic history, physical exam, and laboratory tests. More so, for a patient complaining of chest pain with a high pre-test probability of cardiac disease (e.g., elderly male with a history of diabetes and cigarette smoking), accepting a "normal" cardiac stress test as reassurance of health reflects inappropriate medical care due to the high risk of a false negative test.[7] Complementary false positives include incidental findings that subject patients to the risk of additional follow-up testing, procedures, and complications.[8] Serially repeated lab draws, for example, contribute to adverse outcomes including iatrogenic anemia,[9] impaired sleep, risk of delirium and overall decreased patient comfort and satisfaction.[5]

Given the aforementioned issues, there is increasing emphasis on high value care education[10–13] to avoid diagnostics without reasonable expectation to change management decisions. Human doctors typically have heuristic impressions for diagnostic utilities, but an over-reliance on anecdotal data subject to availability, framing, and anchoring bias as well as fatigue and excess responsibilities can confound decision making. The result is that most human (physicians) have a poor intuition for quantitatively applying diagnostic test performance with gross mis-estimates of pre- and post-test probabilities.[14]

With the rising prevalence of electronic medical record data sources, the potential now exists to inform medical decisions with concrete data and high-throughput predictions models.[15,16] For example. prior studies have illustrated high accuracy in imputing the value of ferritin laboratory results given all other labs co-ordered, indicating redundant information.[17] Others have reviewed historical data to develop risk scores to estimate the yield of blood cultures.[18] Finding extremes of high (or low) pre-test probabilities of normal laboratory test results reflect low "entropy" (and thus low "information" content) in a diagnostic test.[19] These

point to opportunities in clinical decision support to systematically identify low yield diagnostics and interventions.[20,21]

Objective

Here we describe the prevalence of common laboratory tests in a hospital environment and the rate of "normal" results to quantify pre-test probabilities under different conditions. Pre-test probabilities are estimated as a function of repeated testing as well as through a battery of machine learning methods applied to existing electronic medical record data to illustrate the potential for predictable lab results based only on existing data available *before* the laboratory tests were ever ordered.

Methods

We extracted deidentified patient data from the (Epic) electronic medical record for all inpatient hospitalizations at the Stanford tertiary academic hospital via a clinical data warehouse.[22] The structured data covers patient encounters from their initial (emergency room) presentation until hospital discharge. With five years of data spanning 2008-2014, the dataset includes structured data elements (primarily laboratory orders and results in this study) for >71K patients. Numerical lab results were binned into categories based on "abnormal" flags established by the clinical laboratory. We aggregated patient problem list and admission ICD9 diagnosis codes based on the Charlson comorbidity categorizations.[23] The study was deemed non-human subjects research by the Stanford IRB.

Inpatient Population Information

Total Number of Patients	71051
Total Number of Hospitalizations	114296
Age – Mean (Years)	60.12652
Age – Standard Deviation (Years)	18.68509
Percentage Female / Male	51.1 / 48.9

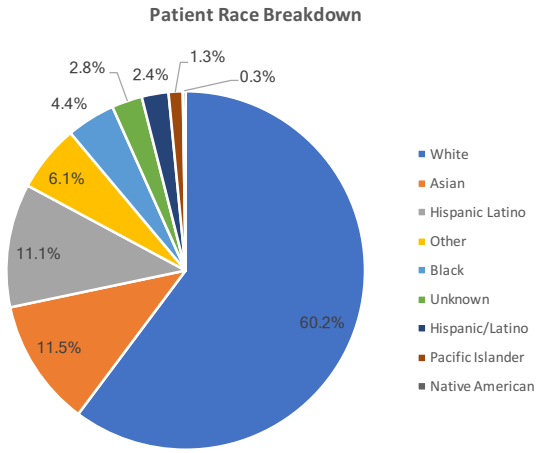


Table 1a (left) – Population information for the 71K sampled patients. Table 1b (right) – Breakdown by race of the sampled patients.

We restricted ourselves to laboratory diagnostics that the clinical lab pre-specified as "Normal" or "In Range" vs. not (e.g., "Abnormal," "High," or "Low") to use an established reference for binary classification of result values. We reviewed serial normal results within several prior window periods (1, 2, 4, 7, 30, and 90 days) to assess how the normal rate shifts after *k* consecutive normal results are observed (e.g., daily lab tests). When *k* is 0, this reflects no previous results (normal or abnormal) exist within the window period. When *k* is 1, this reflects there is exactly one consecutive prior normal result.

Pseudocode for Counting Consecutive Normal Lab Results:

```
Initialize counts dictionary to zeros
Initialize queue to empty
For every lab result (chronological order):
       Remove all results in the queue outside the current window
       If the previous lab result is outside the current window:
              k = length of queue
       Else:
```

```
        k = NULL
Increment total count for k
If result is normal:
        Add the result to the queue
        Increment normal count for k
Else:
        Empty the queue
```

The length of the queue represents how many consecutive normal results fall within the window. Thus, when we see an abnormal result, the queue is cleared. The queue can be empty for two reasons: a) no prior result falls within the window, or b) the prior result is abnormal (but falls within the window). In order to distinguish between the two cases, we check whether the previous lab result falls within the window. In order to align with the definition of $k = 0$ given earlier, we disregard counts where k equals 0 as given by the pseudocode and instead use the values where k equals NULL.

Note that it is common for clinicians to order "panel" tests with multiple component results. For example, a common order is for a basic metabolic panel (BMP), which yields results for Sodium (Na), Potassium (K), Chloride (Cl), HCO3 (Bicarbonate), Blood Urea Nitrogen (BUN), Creatinine (Cr), Calcium (Ca) and Glucose. Individual results register as normal or abnormal, while the panel test is labeled normal if and only if all component results are normal.

For an example set of common laboratory tests (e.g., TSH=Thyroid Stimulating Hormone, SPLAC=Sepsis Protocol Lactate, FER=Ferritin, NTBNP=N-Terminal pro-Brain Natriuretic Peptide) with single result values (unlike panel tests that rarely have *all* normal results) we trained machine learning models to predict the binary result of whether a random sampling of orders in 2013 would yield a "normal" result. Data features extracted included:

- Patient Age
- Gender
- Race (White Non-Hispanic-Latino, Asian, Black, Pacific Islander, etc)
- Time of Day lab ordered (reflecting daily patterns)
- Time of Year lab ordered (reflecting seasonal patterns)
- Presence and time since occurrence of a diagnosis in one of the Charlson comorbidity groups
    - (e.g., Moderate-Severe Liver Disease, Rheumatic Disease, Congestive Heart Failure, etc.)
- Presence and time since primary or consulting treatment by hospital specialty teams
    - (e.g., Cardiology, Hematology, Oncology, Medical Intensive Care Unit, etc.)
- Presence and time since occurrence of the same lab test being predicted, with different window sizes (1 day, 7 days, 30 days, etc.)
- Summary statistics as below during the window period for results of the laboratory test being predicted, vital signs (blood pressure, pulse, respiration rate, temperature, urine output, Glasgow Coma Scale, FiO2) and laboratory result values (white blood cells, hemoglobin, sodium, potassium, bicarbonate, blood urea nitrogen, creatinine, lactate, ESR, CRP, Troponin, blood gases) commonly used in severity of illness prognosis systems.[24,25] Missing values were simply imputed with the overall median observed values for each measure, while the "count" feature effectively acts as an indicator variable to reflect the informativeness of certain lab results *not* being available.
    - Count
    - Count In Range / Normal
    - Min
    - Max
    - Median
    - Mean
    - Standard deviation
    - First
    - Last
    - Difference = Last – First
    - Slope = Difference / (Last Time - First Time)

For each example lab test evaluated, we randomly split the data into training and test sets, with 80% of the data

going to the training set and the remainder being held out for testing. We used several popular machine learning models implemented in the scikit-learn Python library (http://scikit-learn.org). We performed a grid search on various hyperparameters to identify favorable configurations, using area under the receiver operating characteristic curve (ROC AUC) accuracy as an objective function to guide the hyperparameter search. Methods and hyperparameters used include:

- Decision Tree
  - Max_depth: Maximum allowed depth of the decision tree
- Random Forest - Ensemble classifier that fits multiple decision trees and averages their results
  - N_estimators: Number of decision trees to use in the ensemble
  - Max_depth: Maximum allowed depth of any decision tree
- Ada Boost - Ensemble classifier that starts with a base model (Decision Tree) and then fits additional copies of the model to the data, but uses an adaptive learning rate to put emphasis on misclassified examples
  - N_estimators: Number of decision trees to use in the ensemble
  - Learning_rate: Multiplier for the contribution of each successive classifier in the ensemble
- Logistic Regression
  - C: Inverse of regularization strength (smaller means stronger regularization)

Results

Figure 1a and 1b report the most prevalent lab tests ordered and component results in the hospital along with the rate of "normal" = "In Range" values for each, indicating a general pre-test probability. Figure 2 illustrates how this pre-test probability "normal rate" in most cases rapidly converges given a previous consecutive series of normal results for the same test (indicating reducing information value in repeated testing). Table 2 reports the prediction accuracy in terms of ROC AUC (c-statistic) of several machine learning methods for binary classification of "normal" results for example laboratory tests based on patient demographic and derived time series summary features to illustrate the overall discriminatory power and prediction accuracy. Figure 3 "zooms in" on the tail of example cases predicted most likely to yield normal results and the actual normal rate ("precision at K").

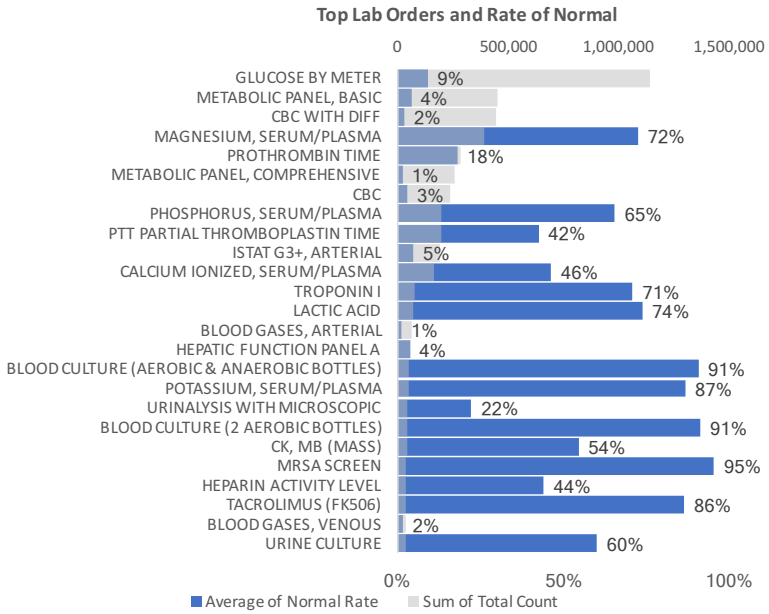


Figure 1a – Top 25 most prevalent lab orders 2009-2014. Quantity noted by gray bars. Blue bars and data labels reflect percent of lab orders yielding all "Normal" = "In Range" results. For example, over there were over 1,000,000 orders to check a patient's Glucose by Meter, with 9% of such test yielding a value in the normal range.

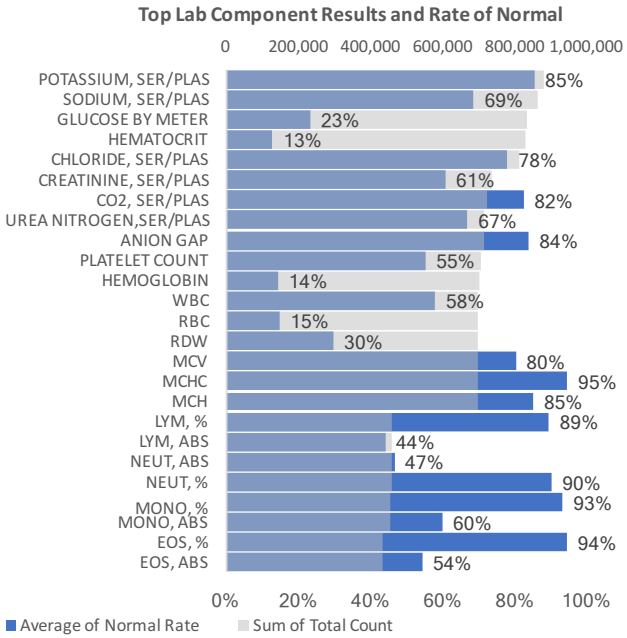


Figure 1b – Top 25 most prevalent lab result components 2009-2014. Quantity noted by gray bars. Blue bars and outlying data labels reflect percent of lab results denoted as "Normal" = "In Range" results.

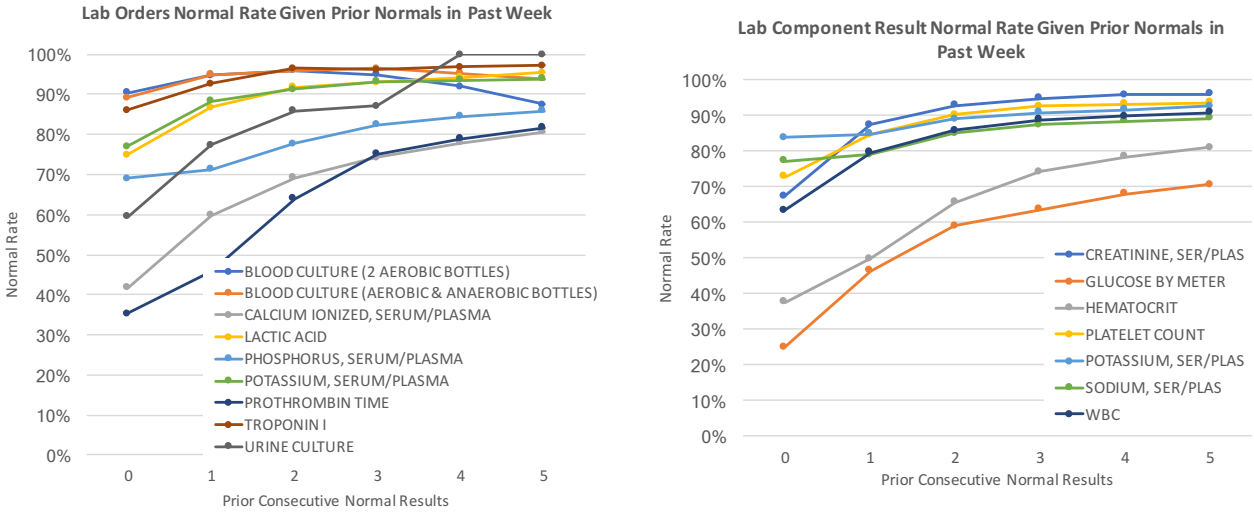
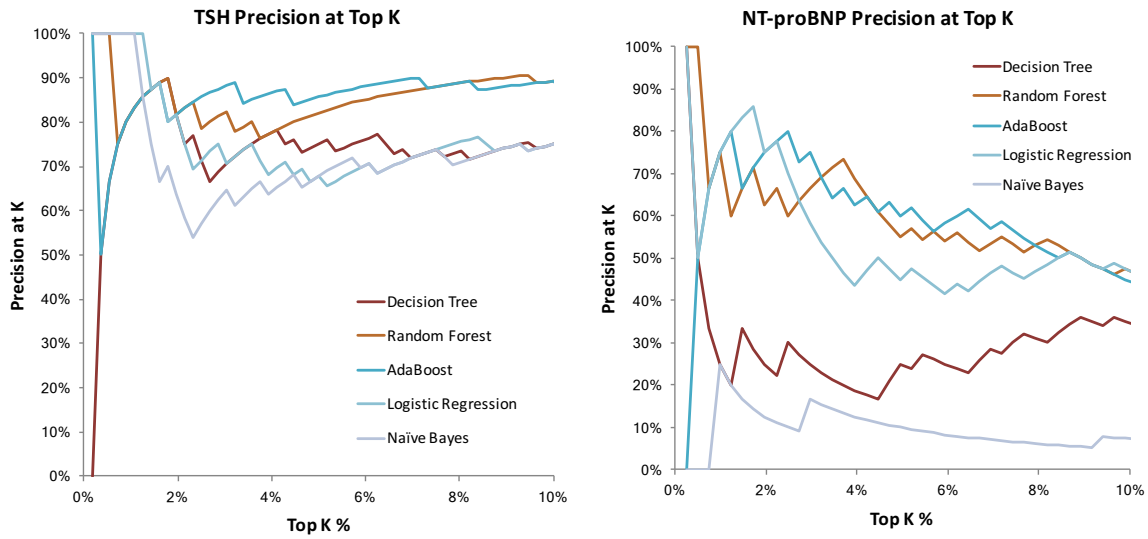


Figure 2 – Common laboratory tests and component results with their rate of all "normal" results as a function of consecutive normal results previously observed in the prior week. For example, of all (Ionized) Calcium tests ordered where zero prior results existed in the week prior, just over 40% yielded a result in the normal range. For repeated testing of Ionized Calcium, the prevalence of normal results progressively increases with respect to the number of prior normal results. In this case, Ionized Calcium tests yield ~80% results in the normal range when preceded by 5 consecutively normal results in the past week. WBC:White Blood Cells

Learning Method + Hyperparameters	Ferritin (FER)	Lactate, Sepsis Protocol (SPLAC)	NT-proBNP (NTBNP)	Thyroid Stim. Hormone (TSH)
Ada Boost n_estimators=10 learning_rate=0.01	0.62	0.60	0.68	0.58
Ada Boost n_estimators=10 learning_rate=0.1	0.67	0.82	0.79	0.60

Ada Boost n_estimators=10 learning_rate=1	0.71	0.87	0.80	0.58
Ada Boost n_estimators=100 learning_rate=0.01	0.68	0.82	0.79	0.61
Ada Boost n_estimators=100 learning_rate=0.1	0.73	0.88	0.86	0.62
Ada Boost n_estimators=100 learning_rate=1	0.75	0.86	0.85	0.59
Ada Boost n_estimators=50 learning_rate=0.01	0.65	0.81	0.79	0.60
Ada Boost n_estimators=50 learning_rate=0.1	0.72	0.87	0.83	0.61
Ada Boost n_estimators=50 learning_rate=1	0.74	0.85	0.87	0.60
Decision Tree max_depth=20	0.59	0.69	0.70	0.51
Decision Tree max_depth=5	0.66	0.76	0.78	0.56
Decision Tree max_depth=50	0.61	0.63	0.63	0.53
Gaussian Naive Bayes	0.57	0.79	0.60	0.60
Logistic Regression C=0.1	0.74	0.86	0.83	0.59
Logistic Regression C=1.0	0.73	0.84	0.82	0.59
Logistic Regression C=10.0	0.73	0.82	0.80	0.58
Random Forest n_estimators=10 max_depth=10	0.72	0.88	0.79	0.59
Random Forest n_estimators=10 max_depth=15	0.69	0.85	0.79	0.58
Random Forest n_estimators=10 max_depth=5	0.72	0.85	0.79	0.62
Random Forest n_estimators=30 max_depth=10	0.73	0.87	0.80	0.61
Random Forest n_estimators=30 max_depth=15	0.73	0.87	0.83	0.64
Random Forest n_estimators=30 max_depth=5	0.72	0.86	0.79	0.62
Random Forest n_estimators=5 max_depth=10	0.70	0.88	0.70	0.56
Random Forest n_estimators=5 max_depth=15	0.68	0.84	0.71	0.59
Random Forest n_estimators=5 max_depth=5	0.70	0.83	0.76	0.59
Max Value	**0.75**	**0.88**	**0.87**	**0.64**

Table 2 – Prediction accuracy (ROC AUC) for example lab tests using electronic medical record data to train models by several machine learning methods and range of hyperparameters. The high discrimination power overall illustrates the significant predictive power of existing structured clinical data towards preemptively predicting these example laboratory results.



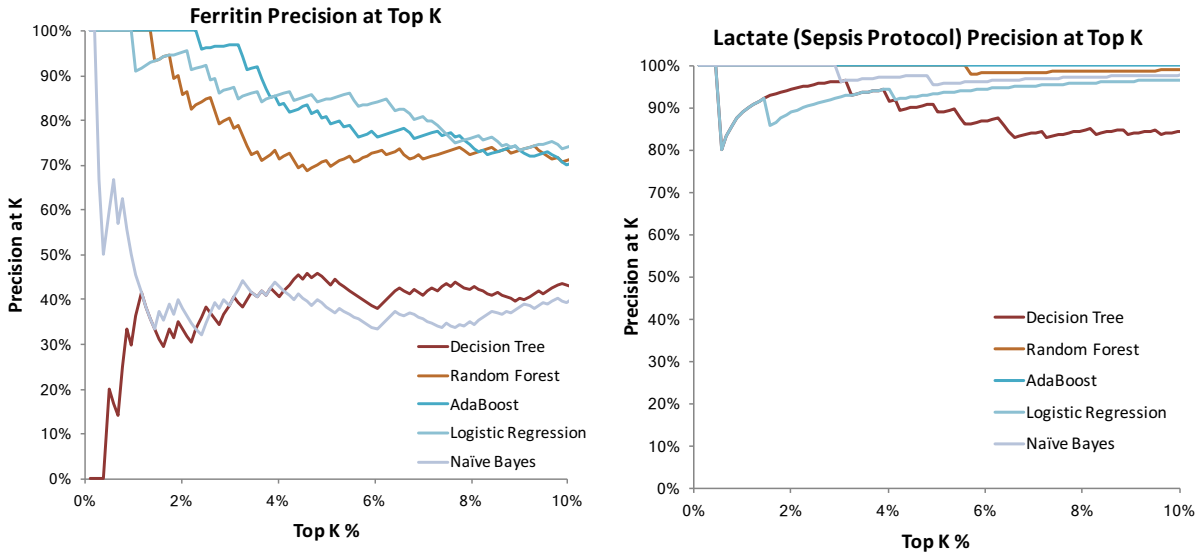


Figure 3 – Top 10% tail distribution of several example lab results based on how likely they are to yield "normal" results as predicted by machine learning methods. Curves plot the precision (positive predictive value) = rate of normal lab results when assessing only the top K% of results). For many methods and many labs, the top few percent of lab results are essentially 100% predictable to be normal given existing data. Example hyperparameters illustrated: Decision Tree Max Depth = 20, Random Forest 30 Estimators with Max Depth =5, AdaBoost with 50 Estimators and Learning Rate = 0.1, Logistic Regression with Regularization C = 1.0, and Gaussian Naïve Bayes).

Discussion

Low (information) yield laboratory tests are commonly ordered in the hospital, with Figure 1a highlighting blood cultures with very high (~90%) pre-test probabilities of being normal. Figure 2 and Figure 3 illustrate how pre-test probabilities can be further refined by integrating additional information on prior consecutive results or more complex patterns of clinical data assimilated through machine learning methods. For cases where the pre-test probability of "normal" approaches 50%, this reflects high information entropy where the tests are most valuable to confirm information that is difficult to predict. We focus our attention on the tails of the distribution where test instances have an extremely high (or inversely an extremely low) pre-test probability. Figure 3 in particular illustrates the distribution of predictability for individual laboratory test instances at the extreme ends, where a percentage of cases is essentially 100% predictable to be normal.

Our approach provides a data-driven, systematic method to identify cases where the incremental value of testing is worth reconsidering. The low information entropy in these cases reflects low expected diagnostic value for medical decision making. When the pre-test probability of a diagnostic result is extremely high/low, finding the opposite test result may even reflect incorrect findings (false positive or false negative) rather than a surprising clinical result.[14] For example, the ~90% of normal blood cultures does not even include false positive blood cultures growing (skin) contaminant organisms that do not reflect a clinically relevant infection. While this approach systematically identifies possible low yield diagnostic cases for review, it is important to acknowledge that a high pre-test probability of a normal result does *not* obviate the value of the test. Low *information* does not entail low *importance* or low *value*. A test with 99% pre-test probability of being normal may still be worthwhile if the consequence of missing the 1% event (e.g., a preventable cancer or bacteremia) is itself extreme. In such cases, a more nuanced cost-effectiveness or decision analysis is more appropriate than relying on prediction accuracy alone.

We focus on diagnostic laboratory testing in this work as a well-structured target, though the approach can be expanded to many other application (e.g., predicting low yield imaging tests). This

approach can also easily predict which laboratory tests have a high pre-test probability of being *ab*normal. We do not target these cases however, as there are many contexts where it is clinically important to monitor abnormal lab results (e.g., tracking hemoglobin values in a bleeding patient or creatinine values in a patient recovering from kidney injury). Notably, normal (negative) results are often quite valuable and necessary to trigger subsequent care decisions. For example, confirming negative blood cultures is often a key decision point to guide antibiotic treatment, and our results should not be interpreted as a blanket argument that blood cultures are wasteful tests.

While normal results are often valuable, the more nuanced argument our approach presumes is that confirming a *highly predictable* "normal" lab result is often a waste. Clinicians, patients, and family members commonly argue that it is worth it "just to document" the normal results to "rule out all possibilities." While a common rationale is that it "never hurts to have more information" that you might need "just in case," information theory would explain that there is essentially no information gained from confirming results that are reliably predictable from available data. As a result, common practice may well reflect both an over-estimation of the benefits of lab testing and an under-estimation of the risks (iatrogenic anemia, erroneous or incidental findings leading to further unnecessary work-up, risks of delirium, etc.). Returning to our blood culture example, really we do not care about whether the culture is positive or negative. What we care about is whether the patient has a bloodstream infection. The diagnostic test is imperfectly correlated with the clinical status we actually care about, but provides enough predictive power to satisfy a decision making threshold. If diagnostic tests have imperfect accuracy and "it never hurts to have more information," arguably we should always order more tests. Indeed we do routinely order blood cultures in sets of 4 bottles at a time (or even 12 bottles for suspected heart infections). We would not continue this indefinitely however, as once we have achieved enough confidence based on available information, further testing is not expected to change our assessment or decision making.

Limitations of this study include its basis at a single medical center. The methods are generalizable, but would be best reproduced with local data if attempting to predict local practice patterns that will constantly evolve.[26] To help cope with model overfitting, we evaluated on randomly separated train-test splits of the data and focused on varying select model hyperparameters. With large values of max_depth, the tree models may strongly overfit the training data. With large values of ensemble n_estimators, we can reduce overfitting at a cost to training time. Based on the results analyzed here, the ensemble based AdaBoost and Random Forest approaches tend to yield better laboratory prediction accuracies, but many other learning methods (e.g., LASSO L1 regularization,[27] support vector machines, and neural networks) and hyperparameter variants can be surveyed for their efficacy. The ensemble based approaches likely perform better in this problem construction when they can better cope with the many highly correlated or irrelevant feature variables. For example, if only a single prior vital sign reading was available for a prediction, the min, max, mean, and median values will all be identical (let alone correlated). Prior studies on predictions using semi-structured healthcare data have also integrated more sophisticated missing data imputation methods for laboratory values,[17] though we focused on summary statistics of the laboratory ordering behavior (e.g., quantity and timing) that are often more predictive of clinical outcomes than the value of the results themselves.[28] Future work can explore not just the overall predictive accuracy of these different methods but identify which data features are most informative, with prior work suggesting the "first derivative" of time series data may be more valuable than the primary data.[29]

We envision implementations where computerized provider order entry (CPOE) systems automatically provide patient-specific pre-test probability estimates whenever a clinician is about to order a diagnostic test. This would be similar to existing frameworks where systems can report the relative cost, turnaround time, and prior result values for diagnostic tests before attempted order entry.[30, 31, 32] Similar to these existing constructs, the clinician remains empowered to make the final decision as to whether the diagnostic test order is appropriate given all contextual information (including patient factors that may not be obvious in the computer record). Explanatory modules for why the system predicts extreme pre-test

probabilities can further increase confidence and adoption. Short of such direct implementation, our approach can simply identify the outlier population of low yield diagnostic testing for expert clinician review to confirm or refute appropriateness. If inappropriate use is confirmed, this can lead to different interventions such as targeted education, standardized practice guidelines, and customized best practice alert rules[33] (perhaps inspired by the relevant feature set learned from the machine learning algorithms).

While identifying low yield medical care via automated methods is necessary, it is insufficient, to deter ineffective use. Factors that contribute toward unnecessary testing ranges from ease of ordering with panels and electronic order entry, inexperience, fear of litigation and perceived pressure from colleagues and patients.[34] Affecting change in clinical practice will require a range of processes. Patient education is necessary to illuminate when "less is more" in healthcare. The respective risks of pressuring providers to administer more diagnostics and treatments is important to clarify, particularly in the context of increasing emphasis on patient satisfaction (scores). Refinement of payment models are necessary to drive incentives for high value care, at both the institutional and provider-specific level. Clinician education remains crucial in the quantitative interpretation of diagnostic performance characteristics, backed by effective decision support. These challenges are significant, but effectively balancing the benefits and risks of diagnostic testing is critical towards simultaneously optimizing the cost, quality, and access in our healthcare system.

Conclusions

Low yield laboratory tests are commonly ordered in the hospital. These can be quantitatively targeted for clinical decision support based on machine learned patterns from readily available structured electronic medical record data.

Bibliography

1. Hackbarth, A. D. Eliminating Waste in US Health Care. *JAMA* **307,** 1513 (2012).
2. Smith, M. D. *et al.* Transformation of Health System Needed to Improve Care and Reduce Costs. *Institute of Medicine News* (2012). Available at: http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=13444.
3. Colla, C. H., Morden, N. E., Sequist, T. D., Schpero, W. L. & Rosenthal, M. B. Choosing Wisely: Prevalence and Correlates of Low-Value Health Care Services in the United States. *J. Gen. Intern. Med.* **30,** 221–228 (2014).
4. Zhi, M., Ding, E. L., Theisen-Toupal, J., Whelan, J. & Arnaout, R. The landscape of inappropriate laboratory testing: A 15-year meta-analysis. *PLoS One* **8,** 1–8 (2013).
5. Konger, R. L. *et al.* Reduction in unnecessary clinical laboratory testing through utilization management at a us government Veterans Affairs hospital. *Am. J. Clin. Pathol.* **145,** 355–364 (2016).
6. Yeh, D. D. A clinician's perspective on laboratory utilization management. *Clin. Chim. Acta* **427,** 145–150 (2014).
7. Diercks, D. B. *et al.* Incremental value of objective cardiac testing in addition to physician impression and serial contemporary troponin measurements in women. *Acad. Emerg. Med.* **20,** 265–270 (2013).
8. Hom, J., Smith, C. (daisy), Ahuja, N. & Wintermark, M. R-SCAN: Imaging for Low Back Pain. *J. Am. Coll. Radiol.* **13,** 1385–1386.e1 (2016).
9. Salisbury, A. C. *et al.* Diagnostic blood loss from phlebotomy and hospital-acquired anemia during acute myocardial infarction. *Arch. Intern. Med.* **171,** 1646–1653 (2011).
10. Levy, A. E., Shah, N. T., Moriates, C. & Arora, V. M. Fostering value in clinical practice among future physicians: time to consider COST. *Acad. Med.* **89,** 1440 (2014).
11. Moriates, C., Soni, K., Lai, A. & Ranji, S. The value in the evidence: teaching residents to 'choose wisely'. *JAMA Intern. Med.* **173,** 308–310 (2013).
12. Moriates, C., Mourad, M., Novelero, M. & Wachter, R. M. Development of a hospital-based program focused on improving healthcare value. *J. Hosp. Med.* **9,** 671–677 (2014).
13. Korenstein, D. Charting the Route to High-Value Care: The Role of Medical Education. *JAMA* **314,** 2359–2361 (2015).
14. Manrai, A. K., Bhatia, G., Strymish, J., Kohane, I. S. & Jain, S. H. Medicine's uncomfortable relationship with math: calculating positive predictive value. *JAMA Intern. Med.* **174,** 991–993 (2014).
15. Chen, J. H., Podchiyska, T. & Altman, R. B. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J. Am. Med. Inform. Assoc.* **23,** 339–348 (2016).
16. Chen, J. H., Goldstein, M. K., Asch, S. M., Mackey, L. & Altman, R. B. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. *J. Am. Med. Inform. Assoc.* ocw136 (2016).

17. Luo, Y., Szolovits, P., Dighe, A. S. & Baron, J. M. Using Machine Learning to Predict Laboratory Test Results. *Am. J. Clin. Pathol.* **145,** 778–788 (2016).
18. Shapiro, N. I., Wolfe, R. E., Wright, S. B., Moore, R. & Bates, D. W. Who Needs a Blood Culture? A Prospectively Derived and Validated Prediction Rule. *J. Emerg. Med.* **35,** 255–264 (2008).
19. Mezard, M. & Montanari, A. Introduction to Information Theory. in *Information, Physics, and Computation* (Oxford University Press, 2009).
20. Kawamoto, K., Houlihan, C. a., Balas, E. A. & Lobach, D. F. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* **330,** 765 (2005).
21. Varghese, J., Kleine, M., Gessner, S. I., Sandmann, S. & Dugas, M. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *J. Am. Med. Inform. Assoc.* (2017). doi:10.1093/jamia/ocx100
22. Lowe, H. J., Ferris, T. A., Hernandez, P. M. & Weber, S. C. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu. Symp. Proc.* **2009,** 391–395 (2009).
23. Quan, H. *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **43,** 1130–1139 (2005).
24. Lemeshow, S. & Le Gall, J. R. Modeling the severity of illness of ICU patients. A systems update. *JAMA* **272,** 1049–1055 (1994).
25. Le Gall, J.-R., Lemeshow, S. & Saulnier, F. Simplified Acute Physiology Score (SAPS II) Based on a European / North American multicenter study. *JAMA* **270,** 2957–2963 (1993).
26. Chen, J. H., Alagappan, M., Goldstein, M. K., Asch, S. M. & Altman, R. B. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Inform.* **102,** 71–79 (2017).
27. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. *Elements* **1,** 337–387 (2009).
28. Benik, N. *et al.* Visualizing Healthcare System Dynamics in Biomedical Big Data. *AMIA Joint Summits on Translational Science* (2017).
29. Küffner, R. *et al.* Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat. Biotechnol.* **33,** 51–57 (2015).
30. Fang, D. Z. *et. al.* Cost and turn-around time display decreases inpatient ordering of reference laboratory tests: a time series. *BMJ Qual Saf.* **23,** 994-1000 (2014).
31. Goetz, C. *et al.* The effect of charge display on cost of care and physician practice behaviors: a systematic review. *J. Gen. Intern. Med.* **30,** 835-842 (2015).
32. Silvestri, M. T. *et al.* Impact of price display on provider ordering: A systematic review. *J. Hosp. Med.* **11,** 65-76 (2016).
33. Luo, R. F. *et al.* Alerting physicians during electronic order entry effectively reduces unnecessary repeat PCR testing for Clostridium difficile. *J. Clin. Microbiol.* **51,** 3872-3874 (2013).
34. Hom, J. *et al.* A High Value Care Curriculum for Interns: A Description of Curricular Design, Implementation and Housestaff Feedback. *Postgrad. Med. J.* **(Accepted),** (2017).