Adaptive Sequential Laboratory Diagnostic Tests: Joint Bayesian Analysis for Optimality

Zeyneb Guenfoud, Péter Antal
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary
Email:{guenfoud, antal}@mit.bme.hu

***Abstract*—Laboratory diagnostic tests provide a fundamental, traditional level in clinical practice and biomedical research. Despite the detailed diagnostic characterization of individual laboratory tests, their overall interdependences has not been investigated. We summarize a probabilistic framework to define optimal measurements, which relies on comprehensive, multivariate probabilistic models of laboratory diagnostic tests formalized as probabilistic graphical models. Within the probabilistic framework we propose sequential inference schemes to improve requested tests. We discuss the challenges and propose scenarios for the integrated application of a decision support system optimizing the selection of laboratory tests with an interplay of clinicians and the laboratory. We also present results of preprocessing a dataset from a central laboratory of a large medical center, which will be the basis of later real-world evaluations.**

***Index Terms*—artificial intelligence, probabilistic graphical models, Bayesian learning, optimal decision, laboratory diagnostic test, cost efficiency**

I. Introduction

The availability of massive, electronic health datasets provides an unprecedented opportunity in early diagnosis and personalized medicine. Modern cornerstones of personal health information are the various molecular biological datasets corresponding to different biological levels and corresponding measurement technologies, such as genetics, transcriptomics or proteomics. However, clinical laboratories are still have a central role in clinical practice: for myriads of clinical requests about diagnostic tests from a wide range, they provide standardized, high-quality information with strict time-constraints. On the one hand, this separated service model focusing on the measurement of the requested tests may increase efficiency at the population level, e.g. optimizing the measurement process of multiple requests according to the laboratory infrastructure and workload. But on the other hand, it can decrease efficiency at the level of patient, e.g. the measurement process in an actual case could be guided by the measured information, optionally involving the clinical expert as well. Specifically, a sequential, adaptive measurement process of laboratory tests, potentially with an interaction between the laboratory and clinical diagnostician could result in the following:

1) *Canceled measurements* The laboratory could inform the clinician that certain requested tests are confidently predictable based on earlier measurements from the patient's history and from current measurements. Depending on the suspected disease and corresponding diagnostic protocol, the clinician could decide that the in silico predictions are sufficient and could cancel the pending requests.
2) *Extended measurements* The laboratory could inform the clinician that the value of certain not requested tests are abnormal with high confidence, predicted based on earlier measurements from the patient's history and from current measurements. Depending on the suspected disease and corresponding diagnostic protocol, the clinician could decide that the measurements could be indeed valuable for those variables and expand the request.

We investigate the following scenario of adaptive, sequential laboratory diagnostic tests, for which the assumptions are motivated by our real-world cooperation with the Central Laboratory of the Semmelweis University:

1) *Separated laboratory information* The laboratory has no access to the patient's medical history and current suspected diseases, but basic demographic information, such as gender and age and earlier laboratory tests for the patients may be available.
2) *Requested tests with urgent/compulsory subsets and suggested ordering* The clinician could ask the measurement of test(s) indicating also that certain measurements are urgent and/or compulsory. Suggested ordering for their sequential measurement can be also indicated.
3) *Predictable tests* The laboratory could inform the clinician that the value of certain tests are confidently predictable based on earlier measurements from the patient's history and from current measurements.

The central assumption of our approach is that laboratory diagnostic tests have a robust probabilistic dependency structure, in fact, the redundancy of the current set of tests is one of the main challenge in laboratory medicine [15]. Because in our scenario information about indications and diseases are not available, we focus on the separated, standalone dependency structure of the tests. Note that this property excludes the usage of information about well-known disease-specific multivariate tests. Furthermore, to simplify our task, as a first approximation, we ignore the temporal aspect of laboratory tests, e.g. we do not perform a time-series analysis and do not model that certain tests are used to monitor the result of a surgery.

Based on these assumptions, the main questions of our work is twofold:

1) *Predictable measurements* We try to estimate the distribution, variance and expected value of the number of correctly predictable measurements.
2) *Unmeasured abnormalities* We try to estimate the distribution, variance and expected value of the number of unmeasured tests with abnormal values.

Two notes are in order. Note that certain measurements are prescribed by medical protocols, e.g. to exclude vital conditions. Thus, the expected value of predictable measurements provides only an upper bound for the potentially avoidable laboratory tests. Analogously, certain measurements are trivially abnormal in certain medical conditions, so they are not requested to measure, consequently the expected value of unmeasured abnormalities is only an upper bound for missed, medically relevant measurements.

In short, our assumptions are twofolds. Firstly, it relies on a non-temporal learning phase, in which laboratory test measurements in a given, recent period are merged into a vectorial description as current state, and earlier measurements are neglected. This first phase results in a *a posteriori* distribution over models. Secondly, we will approximate inference for a given patient using this *a posteriori* distribution over models and perform a sequential inference in a given model using laboratory tests as evidences of the current state of a patient.

II. Earlier works

The broader context of our formalization and approximation for an adaptive, sequential use of laboratory diagnostic tests encompasses the full-fledged Bayesian decision theoretic framework, including actions for the selection of tests for measurements, for the rejection of a measured value and an action for stopping with the measurements and possibly suggesting interventions based on the diagnosis. The optimal selection of actions resulting in interventions and observations, leads to the concept of expected value of an experiment (EVE) [18].

Ignoring the potential interventional consequences, the utilization of the temporal sequence of measurements for a given patient can be seen as a time-series analysis, especially in the prequential and online learning approaches for non-stationer processes. This scenario corresponds to the monitoring of the result of a surgery using a panel of biomarkers for example in oncology.

The real-world constraints on measuring laboratory tests, especially the financial and temporal constraints, can be also investigated in the frameworks of active learning and budgeted learning.

The measurement itself of a laboratory test usually indicates an increased belief for a potential abnormal value, i.e. the informativeness of the mere availability of a measured test. Thus, the usual laboratory test datasets violate the missing-at-random (MAR) assumption and require special approaches [5].

Assuming that the inductive part results in a limited number of models, the adaptive, sequential use of laboratory tests in this phase corresponds various types of inferences in a complex, fixed model to explore the current state of a given patient. If utility functions are available, then value of information calculations could be used to support information gathering [4], [6], [7], [9], [11]. Lacking informative utility functions, general domain specific functions can be constructed and/or sensitivity of inference could be used to support information gathering. Another approach is to use explanation generation methods [3], [10].

The joint analysis of all laboratory tests is a natural extension of the network paradigm from diseases, genes, drugs, phenotypes and symptoms [20]. Indeed, both data mining and deep analysis of electronic health records are among the top priorities for improving health care [2], [8], with special emphasis on using laboratory diagnostic tests [12]–[17], [19].

Unfortunately, these earlier frameworks and methods are not directly applicable for this problem, so there are no available benchmark results.

III. Domain and datasets

The raw dataset contains all results for the measurements of 225 most relevant laboratory blood tests between 2011 and 2015 October at the Central Laboratory of the Semmelweis University. From this 4-year period, the original dataset contains 13,754,888 measurements from 1,376,759 orders for 1,392 laboratory tests and 202,976 persons.

The measurements are grouped by their orders, which usually correspond to a visit to a medical professional and a respective blood sampling. For each order, the urgency of the measurement and the institute of the doctor ordering the tests are indicated, but not used in the current analysis. For each patient, gender and age will be available, but could not yet accessed. The identifiers and the abbreviated names of the laboratory diagnostic tests are the World Health Organization (WHO) codes and the Logical Observation Identifier Names and Codes (LOINC) codes. The reference interval and the measurement unit for each measurement is separately indicated in the database, which allows the following semi-quantitative coding and interpretation:

1) **Non-measured (0)**: not suspicious or relevant, default assumption for the unmeasured value is normal.
2) **Measured-normal (1)**: suspicious and relevant, but measured test value is in the reference range.
3) **Abnormally-low (2)**: the measured test value is below the lower bound of the reference range.
4) **Abnormally-high (3)**: the measured test value is above the upper bound of the reference range.

In the current analysis, for each patient the measurements in a given, most recent period are merged and earlier measurements are neglected. We treat these merged tests vectorial representation as the *current state* of the patient.

The applied combinations for the window size and merge function are as follows. We split the data for 5 sub-data as last month (1m), last 3 months (3m), last half year (6m), last year (1y) and last 2 years (2y). In the continuous version of the merge, we aggregate all the sub-data calculating the

maximum (max), minimum (min), average (avg) and median (med) values or keeping only the last measurements (last). In the discretized version of the merge, we first convert each value into a binary normal(0)/abnormal(1) value, then we aggregate all the sub-data calculating their AND (and) and OR (or) combinations or keeping only the last measurements (last, there is no difference in this case between the continuous and discretized version).

For each aggregation, we generated the following three types of matrices with semi-quantitative values, when the rows are the patients and the columns are the laboratory tests.

1) *Measurement indicator* matrix (I_M) has binary values, where 1 means that the laboratory test is performed for the patient in the given period, and 0 if not performed.
2) *Abnormality data* matrix (D) The value *2* means that lab-test is performed for patient and its merged value is *outside* the reference/normal range in the continuous case and 1 in the discretized case. The value *1* means the analogous case and empty means that the lab-test is *not performed* for the patient in the given period.
3) *Ternary data* matrix (T). Technically, it is a completed version of the *Abnormality data* matrix by setting its all empty items to 0. An intuitive interpretation is that a non-measured test indicate an *a priori* normal value, which is in certain cases suggest a smaller risk then a measured, i.e. suspicious, but normal value.

Fig. 1 shows the data preprocessing pipeline.

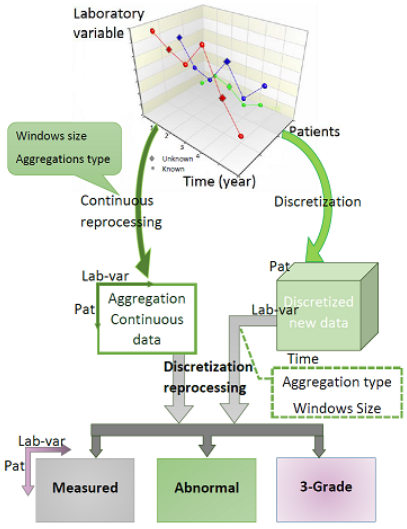

Fig. 1. The preprocessing pipeline resulting discrete data matrices.

We have developed Python scripts for these preprocessing steps and exploratory statistical data analysis.

IV. Methods

The central tasks according to our assumptions are the indication of confidently predictable tests among the requested ones and tests with confidently abnormal values among the unrequested ones. In our non-temporal approach using the described merging and discretization, we conceive these tasks as (1) the indication of confidently predictable tests among the known tests after merging and (2) as the indication of tests with confidently abnormal values among the unknown tests after merging. We introduce concepts for the probabilistic formalization of these questions in case of a new, actual patient with index $N+1$, assuming that the same preprocessing is applied in this case as for the data matrix D_N with N samples. Let K_{N+1} denote the set of the indices of known tests for patient $N+1$: $i \in K$ iff $I_{N+1,i} = 1$. The known set is divided to an evidence set $E \subset K$ and query set $Q = K \setminus E$. Using these index sets, $D_{N+1,K}$ denotes the subvector of known tests. We call the tests in Q as $l - \tau$ predictable iff $|E| = l$ and tests $D_{N+1,Q}$ are predictable with at most τ probability:

$$\forall i \in Q : \tau < \max_{j=1,2} p(D_{N+1,i} = j | D_{N+1,E}, D_N). \quad (1)$$

For a given τ, the minimal value with this property is denoted with $l_{N+1}(\tau)$, i.e. the size of the minimum number of tests from the known set sufficient the predict the rest of the known ones. Note that the set of potentially redundant tests could be defined more precisely using a multivariate approach. Using this univariate formalization, the number of τ-probably redundant tests is defined as

$$r_{N+1}(\tau) = |K_{N+1}| - l_{N+1}(\tau). \quad (2)$$

Analogously, the set $C_{N+1}(\tau)$ of probably abnormal variables with threshold τ is defined as follows

$$i \in C_{N+1} : \tau < p(D_{N+1,i} = \text{"abnormal"} | D_{N+1,K}, D_N).$$

We apply Bayesian network models M_i in the Bayesian model averaging framework to approximate the predictive distributions, i.e. for target Y

$$p(Y | D_{N+1,K}, D_N) \approx \sum_{M_i} p(Y | M_i, D_{N+1,K}) p(M_i | D_N).$$

For this purpose, currently we are extending and evaluating Markov Chain Monte Carlo (MCMC) methods over Bayesian network structures developed earlier in our group [1]. In the general batch case for M patients $D_{N+1:N+M}$, we treat the cases separately, because of computational limitations.

V. Results

The laboratory test measurements are highly incomplete, as they are specific to diseases and clinical conditions. Using only the last laboratory visit for each patient, Fig. 2 shows the proportion of cases with measurement, valid value and normal value for the laboratory tests (proportion of "Valid" is not shown as nearly all measurements have proper syntax, thus valid). The existence of a measurement means its presence in the dataset, its validity means that it has proper syntactic format and the reference interval (a.k.a. normal region) is available, finally, normality means that the measurement of a test has a valid decoding in the reference region.

In the current approach we merge the laboratory visits using varying window size. Table I represents proportions of

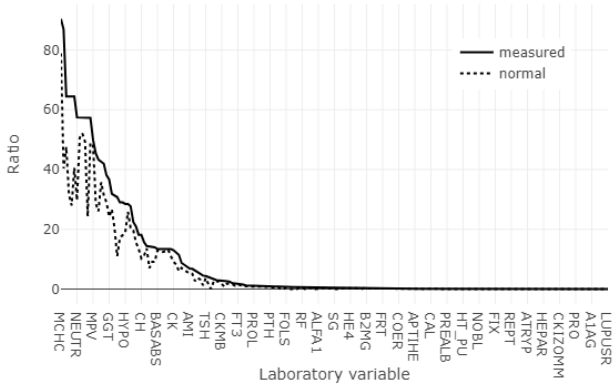

Fig. 2. Proportion of cases that were requested ("Measured") and their value is in the reference range ("Normal").

measured tests, valid values and normal values in the dataset using 1 month (1M), 3 month (3M) and 6 month (6M) window sizes for merging the results per patients.

TABLE I
PROPORTIONS OF MEASURED TESTS BY ACCUMULATING RESULTS.

	Measured	***Valid***	***Normal***
1M	10.46%	10.44%	8.03%
3M	10.66%	10.64%	8.39%
6M	21.75%	21.71%	17.13%

As results in Table I show the effect of merging laboratory test results in a 3 month period is negligible, which probably related to clinical protocols limiting the repetition of certain tests. These results indicate that the high level of incompleteness of the laboratory test data remains a major challenge, as the ratio of valid data is still around 20% after merging results in a 6 month period (homogeneity assumptions of the clinical state for longer periods usually cannot be expected). However, incompleteness is informative for laboratory tests, as indicated by the proposed semi-quantitative coding and interpretation, which property suggest the use of respective, complete datasets.

Currently, we are investigating the effect of discretization, approaches to cope with incomplete data and computational schemes to perform Bayesian inference using Monte Carlo methods jointly over the missing part of the dataset and predictive models.

VI. Conclusion and Future Work

The prediction of unknown tests could be used both in actual clinical decision support and in evaluation of health policies. From clinical point of view, the investigated methods aim to support the cost-effective use of laboratory capacities, as the set of the requested tests can be adaptively modified. Additionally, these functionalities can also support quality control implementing professional protocols, but it could also help the design and refinement of diagnostic protocols.

References

[1] Péter Antal, András Millinghoffer, Gábor Hullám, Csaba Szalai, and András Falus. A bayesian view of challenges in feature selection: feature aggregation, multiple targets, redundancy and interaction. In *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, pages 74–89, 2008.

[2] David Blumenthal and Marilyn Tavenner. The "meaningful use" regulation for electronic health records. *N Engl J Med*, 2010(363):501–504, 2010.

[3] Urszula Chajewska and Joseph Y Halpern. Defining explanation in probabilistic systems. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 62–71. Morgan Kaufmann Publishers Inc., 1997.

[4] Clifford Champion and Charles Elkan. Visualizing the consequences of evidence in bayesian networks. *arXiv preprint arXiv:1707.00791*, 2017.

[5] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[6] David Heckerman, Eric Horvitz, and Blackford Middleton. An approximate nonmyopic computation for value of information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):292–298, 1993.

[7] Finn V Jensen and Thomas Dyhre Nielsen. Probabilistic decision graphs for optimization under uncertainty. *Annals of Operations Research*, 204(1):223–248, 2013.

[8] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[9] Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

[10] Carmen Lacave and Francisco J Díez. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

[11] Wenhui Liao and Qiang Ji. Efficient non-myopic value-of-information computation for influence diagrams. *International Journal of Approximate Reasoning*, 49(2):436–450, 2008.

[12] Giuseppe Lippi, Antonella Bassi, and Chiara Bovo. The future of laboratory medicine in the era of precision medicine. *Journal of Laboratory and Precision Medicine*, 1(3), 2016.

[13] Giuseppe Lippi, Chiara Bovo, and Marcello Ciaccio. Inappropriateness in laboratory medicine: an elephant in the room? *Annals of translational medicine*, 5(4), 2017.

[14] Giuseppe Lippi and Camilla Mattiuzzi. The biomarker paradigm: between diagnostic efficiency and clinical efficacy. *Pol Arch Med Wewn*, 125(04):282–288, 2015.

[15] Giuseppe Lippi and Mario Plebani. False myths and legends in laboratory diagnostics. *Clinical chemistry and laboratory medicine*, 51(11):2087–2097, 2013.

[16] Giuseppe Lippi and Mario Plebani. Laboratory economics. risk or opportunity? *Clinical Chemistry and Laboratory Medicine (CCLM)*, 54(11):1701–1703, 2016.

[17] Martina Montagnana and Giuseppe Lippi. The risks of defensive (emergency) medicine. the laboratory perspective. *Emergency Care Journal*, 1(1), 2016.

[18] Changwon Yoo and Gregory F Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine*, 31(2):169–182, 2004.

[19] Zhongheng Zhang. The role of big-data in clinical studies in laboratory medicine. *Journal of Laboratory and Precision Medicine*, 2(6), 2017.

[20] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature communications*, 5:4212, 2014.