Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Data-driven approach for assessing utility of medical tests using electronic medical records

Stein Olav Skrøvseth [a,b,*,1], Knut Magne Augestad [a,c,d], Shahram Ebadollahi [b]

[a] Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Tromsø, Norway
[b] IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA
[c] Department of Gastrointestinal Surgery, University Hospital of North Norway, Tromsø, Norway
[d] Department of Surgery, University Hospitals Case Medical Center, Case Western Reserve University, Cleveland, OH, USA

ARTICLE INFO

Article history:
Received 29 June 2014
Accepted 23 November 2014
Available online xxxx

Keywords:
Electronic medical records
Medical informatics
Colorectal surgery
Information theory
Pattern recognition

ABSTRACT

Objective: To precisely define the utility of tests in a clinical pathway through data-driven analysis of the electronic medical record (EMR).

Materials and methods: The information content was defined in terms of the entropy of the expected value of the test related to a given outcome. A kernel density classifier was used to estimate the necessary distributions. To validate the method, we used data from the EMR of the gastrointestinal department at a university hospital. Blood tests from patients undergoing surgery for gastrointestinal surgery were analyzed with respect to second surgery within 30 days of the index surgery.

Results: The information content is clearly reflected in the patient pathway for certain combinations of tests and outcomes. C-reactive protein tests coupled to anastomosis leakage, a severe complication show a clear pattern of information gain through the patient trajectory, where the greatest gain from the test is 3–4 days post index surgery.

Discussion: We have defined the information content in a data-driven and information theoretic way such that the utility of a test can be precisely defined. The results reflect clinical knowledge. In the case we used the tests carry little negative impact. The general approach can be expanded to cases that carry a substantial negative impact, such as in certain radiological techniques.

1. Introduction

At any point in the patient pathway, the health care provider can perform a test to increase their knowledge about the state of the patient. The nature of a test can vary widely, from simple examinations to expensive or burdensome patient scans. Examples of tests include blood samples, radiology, invasive procedures, echo-cardiography, endoscopies, and many more. The result of a test is expected to help the physician in the decision-making process on whether to provide a therapy, or perform a procedure. The decision-making process is inherently associated with uncertainties and risks. It is expected that the test results assist the physician in reducing such uncertainties. Physical examination of the patient reveals complimentary information about the patient, such as blood pressure, pulse, temperature, pain, color of skin. Together with prior medical history these present the essential information for making clinical decisions [1]. Often the decision to test is driven by routines, clinical knowledge and suspected issues with the patient. A quantitative way to decide whether or not to test is to employ a Bayesian approach and estimate possible posterior probabilities in the range of possible test results. If there is no posterior that would make the practitioner change their course of action, the test should not be done [2]. Both establishing a prior and compute posteriors accordingly are difficult tasks and render such rules difficult to employ in clinical practice.

Due to the increased use of electronic medical records (EMRs) [3], and the resulting availability of *observational data*, we believe one could find *data-driven* methodologies to define the information content and utility of medical tests. We have devised an approach and conducted a set of experiments to validate this hypothesis. Our experiments conducted in the context of tests and outcomes related to surgery show that one can effectively measure the information content of various medical tests by employing data-driven methods on longitudinal EMR data of patients. By utilizing the large

* Corresponding author at: University Hospital of North Norway, Norwegian Centre for Integrated Care and Telemedicine, P.O. Box 35, N-9038 Tromsø, Norway.
E-mail address: stein.olav.skrovseth@telemed.no (S.O. Skrøvseth).

[1] This research was conducted while the first author was a Visiting Scientist in the Health Informatics department in IBM T.J. Watson Research Center.

http://dx.doi.org/10.1016/j.jbi.2014.11.011
1532-0464/

volumes of data in the EMR, providers can immediately get support that improves the quality of care based on their own data [4,5].

New information about a patient is gathered at a cost of conducting tests. The benefit of a test is in terms of the added information it provides about the patient. This is the *information content* of the test. Here information is used in the sense of the physician gathering more data about the patient, thereby learning more about the patient. There is also a distinct but related mathematical concept of information, phrased in terms of entropic measures that quantifies the surprise of learning an outcome. For an outcome with probability p, the surprise is defined as $-\log p$. Thus, for a distribution $p(x)$ for discrete outcomes x, the expected surprise, also known as Shannon entropy or simply entropy, is $S(X) = -\sum_x p(x) \log p(x)$ with the convention that $0 \log 0 = 0$ [6]. The base of the logarithm is arbitrary, and we use base 2 throughout such that the entropy is measured in bits. This precisely quantifies the expected surprise of learning a piece of information. If the information is known prior to testing ($p(x) = 1$ for one value of x, zero else), entropy is zero. The entropy is maximized when each outcome is equally likely.

We define the *utility* of a test as the increase in information content over the cost of the test. Cost can be in terms of economy, time, risk or discomfort to the patient, with a corresponding range of quantifications. Also, there may be secondary costs, e.g., in the case of false positives leading to unnecessary follow-ups or patient distress. Our goal is to precisely define the information gain due to a test and measure this with the information available in the EMR. Having a precise quantification of the expected information content of a test at a given time in a patient's pathway can be useful in designing data-driven decision support systems for use in clinical practice. Thereby one accomplishes using real-world data to continuously improve clinical practice [7].

1.1. Prior work

The use of information theory in medical testing dates back to 1973 [8]. In 1981 Diamond et al. used information theoretic ideas to quantify the value of ECG tests [9]. Rifkin used an approach that utilized the clinical value of multivalued test results rather than just binary ones [10]. Both approaches relied on theoretical justification or known clinical properties of tests such as their specificity and sensitivity. The information theoretic approaches were compared to the newly introduced concept of *expected value of perfect information* [11], for which an efficient computational algorithm was recently developed by Strong and Oakley [12]. A simple analysis of relative entropy as a measure of the information in diagnostic tests was done by Benish [13], and more recently McCabe et al. used information gain to develop risk scores [14]. Other approaches to using information theory to quantify the information content of tests has been suggested several times [15–19]. Our contribution is providing a data-driven methodology for quantifying the information gain and the utility of medical tests at different stages in a patient's clinical pathway.

There is a general similarity between this problem and experimental design. In experimental design, usually there is an underlying model, and the task is to choose the experiment that provides the most information at optimal cost. The model has a certain parameter set, θ and a choice has to be made as to what experiment $\xi \in \Xi$ needs to be conducted to gain the most amount of knowledge about the parameters, where Ξ denotes a set of available experiments. Mutual information, defined as

$$I(X, \Theta) = S(\Theta) - S_X(\Theta|X),$$

has been suggested as a useful measure for correlating test results to outcomes [20,21]. Liepe et al. argue that mutual information is the best measure for determining the optimal experiment [22]. Here, Θ is the random variable associated with the model parameter θ, X is the output data from a particular experiment and $S(\Theta)$ is the entropy of the distribution $\pi(\theta)$. A definition of what it means for an experiment to be preferable to another experiment was made clear in statistical terms by Goel and Ginebra [23].

Notably, by using the information theoretic approach, the test result X can be of any type, and the information content of various tests would be comparable in terms of the entropy irrespective of the original parameters. It is acknowledged that improving parameter estimates do not necessarily improve model predictions, and an alternative is to maximize the mutual information between predictions and data rather than parameters and data.

2. Materials and methods

2.1. Information gain

In this chapter we define the methods we use to quantify the information gain in a test at a given point in a patient's pathway. The methods are influenced by other techniques [19,22–24], but have been adapted to the context of medical tests. To our knowledge, the framework presented here has not been used in this context before. In order to quantify the information gain for a given medical test, a patient model is needed. The patient model is described by a set of parameters θ. Prior information about these parameters, are encoded in the prior distributions $\pi(\theta)$. The purpose or complexity of the patient model is highly problem specific. In the simplest example it might be a single, binary parameter where $\theta \in \{0, 1\}$ depends on a binary outcome, or the patient model may be arbitrarily complex with a large number of heterogeneous parameters. Obviously, for any chosen patient model, there must be sufficient data available to support the modeling framework.

At any point in time for a given patient model, and based on the prior distribution $\pi(\theta)$ the prior entropy is $S(\Theta)$. When a certain test ξ is chosen, the result of the test y has a distribution $\pi(Y|\Theta, \xi)$. Using Bayes' rule the posterior is

$$\pi(\Theta|Y, \xi) = \frac{\pi(Y|\Theta, \xi)\pi(\Theta)}{\pi(Y|\xi)}.$$

Thus we are able to compute the expected information content. When little is known about the particular patient, the prior can be computed using population averages, or using a flat or uninformative prior. As we perform tests and gain more knowledge about the patient, the prior is updated accordingly. To learn how a test result couples to an outcome, we need to employ a decision rule. Using retrospective data, one can compute the decision rule by observing (θ, y) for every patient, and use a classifier (in case θ takes discrete values) or regression model (when θ takes continuous values), such that the classifier or regression model yields $\pi(\Theta|Y, \xi)$. Another option is to inform the decision rule by clinical knowledge, where the domain expert gives the distributions for each test, which corresponds to a knowledge-driven approach. Notably, the posterior can serve as a prior for subsequent tests, though care should be taken if the tests are not independent. E.g., if the tests are performed using the same or related method (e.g., two computed tomography tests), reasons for false positives might be recurrent in the second test.

In the following we will use data-driven decision rules, where we have access to some retrospective test results from the EMR and the outcome associated with those results. We will denote the decision rule as a classifier, but it is understood that if θ takes continuous values, it will be a regression model [25].

We define the information gain in a similar fashion to how information gain was defined in terms of choosing the best experiment

Please cite this article in press as: Skrøvseth SO et al. Data-driven approach for assessing utility of medical tests using electronic medical records. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.11.011