Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)
© 2015 European Federation for Medical Informatics (EFMI).
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-512-8-190

Outcome-driven Evaluation Metrics for Treatment Recommendation Systems

Jing MEI[a,1], Haifeng LIU[a], Xiang LI[a], Yiqin YU[a] and Guotong XIE[a]
[a] *IBM Research – China*

Abstract. Treatment recommendation systems aim to providing clinical decision supports, e.g. with integration of Computerized Physician Order Entry (CPOE). One of the most significant issue is the quality of recommendations which needs to be quantified, before getting the acceptance from physicians. In computer science, such evaluations are typically performed by applying appropriate metrics that provides a comparison of different systems. However, a big challenge for evaluating treatment recommendation systems is that ground truth is only partially observed. In this paper, we propose an outcome-driven evaluation methodology, and present five metrics (i.e. precision, recall, accuracy, relative risk and odds ratio) with highlight of their statistic meanings in clinical context. The experimental results are based on the comparison of two well-developed treatment recommendation systems (one is knowledge-driven and based on clinical practice guidelines, while the other is data-driven and based on patient similarity analysis), using our proposed evaluation metrics. As a conclusion, physicians are less prone to comply with clinical guidelines, but once following guideline recommendations, it is much more likely to get good outcomes than not following.

Keywords. Clinical Decision Support Systems, Health Care Evaluation Mechanisms, Treatment Outcome, Intervention Studies

Introduction

In literature, a variety of treatment recommendation systems have been proposed and implemented to provide clinical decision support. For instance, the computerization of clinical practice guidelines (CPG) paves the way for knowledge-driven treatment recommendation systems [1]. Clinical statements including recommendations could be computerized as decision rules by rule engines, or more advanced, by process engines. Meanwhile, data mining is the process of finding previously unknown patterns and trends in a large volume of data. [2] is a survey on data mining approaches for healthcare, in which various algorithms, such as Bayesian Network and K-Nearest Neighbour (KNN), are introduced to give personalized intervention options for patients.

To widely populate the usage of treatment recommendation systems, a key is convincing domain experts of their benefits. The most straightforward way is to ask domain experts to answer questionnaires for expressing their agreement or disagreement with the recommendations [4]. However, it requires a lot manual work, and such a labour consuming job unfortunately degrades the acceptance of treatment

1 meijing@cn.ibm.com.

recommendation systems. A work-around solution is to compare recommendations with prescriptions in real data. However, is the prescription really the ground truth (a.k.a. gold standard, or best practice)? The answer should be NO. Suppose that drug A was recommended by a decision engine, and the physician did choose drug A as the prescription, but unfortunately, the patient outcome of using drug A was bad. In this respect, could we mark the recommendation of drug A as correct? Another story is that drug B was recommended by an analysis module, but the physician chose drug C as the prescription, and the patient outcome of using drug C appeared good. Thus, could we mark the recommendation of drug B as incorrect? What if the patient outcome of using drug B becomes better than using drug C?

To address the evaluation problem of treatment recommendation systems in absence of ground truth (or saying, only with partially observed ground truth), we propose an outcome-driven evaluation methodology. In the rest of this paper, we first introduce two recommendation systems. One is knowledge-driven and based on clinical practice guidelines (abbr. CPG). The other is data-driven and based on patient similarity analysis (abbr. PSA). Then, we present five evaluation metrics (i.e. precision, recall, accuracy, relative risk and odds ratio) and highlight their statistic meanings in clinical context from an outcome-driven perspective. For experimental results, we compare the two recommenders using these five metrics, and CPG outperforms PSA as expected. In particular, the compliance rate of CPG is lower than PSA, but the adoption of CPG recommendations have a much higher positive relevance with good outcomes.

1. Methods

1.1. Treatment recommendation systems

As mentioned above, there are two ways for treatment recommendation systems. One is knowledge-driven, and clinical practice guidelines provide the comprehensive domain knowledge. Taking the NICE clinical guideline for Type 2 diabetes as an example, it recommends to start metformin treatment in a person who is overweight or obsess and whose blood glucose is inadequately controlled by lifestyle interventions alone. Via the computerization of clinical practice guidelines, rule engines or process engine could be utilized to provide decision supports, especially for treatment recommendation. Our previous work [5] (abbr. CPG) computerized guidelines into standard (XPDL)-based business processes, where clinical conditions were represented using GELLO expressions. At run-time, a process engine would invoke a query adaptor to retrieve clinical data and a GELLO engine to evaluate clinical conditions whenever a decision-making was needed during the care process. As a result, clinical recommendations were generated for lifestyle intervention, and drug therapy, etc.

The other way is data-driven, and a variety of data mining approaches has been explored to uncover new insights in the healthcare domain. Taking our previous work [6] (abbr. PSA) as an example, feature selection algorithms were first employed to identify the factors that affected physicians' prescription decisions. Then, given a patient at an encounter, his/her clinical conditions were represented using a vector of selected features, and we would find out the K most similar prescription instances, where the similarity was measured by the Euclidean distance between the representing feature vectors. Finally, we would choose the most frequently presented medication option (among the K most similar prescription instances) as the recommendation.

1.2. Evaluation metrics

In this paper, we assume there is no manual work to annotate every recommendation as correct or incorrect. Also, we do not regard prescriptions as ground truth, because the prescription does not always bring good outcome. So, the challenge is how to evaluate treatment recommendation systems without ground truth.

Inspired by the cohort study in clinical research, we divide the existing patient encounters into two groups. One group consists of patient encounters where the prescriptions were compliant with the recommendations, namely the exposed group. The other group consists of patient encounters where the prescriptions were not compliant with the recommendations, namely the control group. From a probabilistic view, the exposed group instances with good outcome are true positive (TP), the exposed group instances with bad outcome are false positive (FP), the control group instances with good outcome are false negative (FN), and the control group instances with bad outcome are true negative (TN). As shown in Table 1, m_{11} is the number of instances in the exposed group with good outcome. Likewise, there are explanations for m_{10}, m_{01} and m_{00}. Five evaluation metrics are defined below.

- Precision: $m_{11}/(m_{11}+m_{10})$
- Recall: $m_{11}/(m_{11}+m_{01})$
- Accuracy: $(m_{11}+m_{00})/(m_{11}+m_{10}+m_{01}+m_{00})$
- Relative risk: $(m_{11}*(m_{01}+m_{00}))/(m_{01}*(m_{11}+m_{10}))$
- Odds ratio: $(m_{11}*m_{00})/(m_{10}*m_{01})$

Table 1. An outcome-driven evaluation measure

	Good outcome	**Bad outcome**
Exposed group: prescription compliant with recommendation	(TP) m_{11}	(FP) m_{10}
Control group: prescription not compliant with recommendation	(FN) m_{01}	(TN) m_{00}

The first three metrics are popular used in statistics. In our context, precision is the probability that a (randomly selected) compliant instance has good outcome, and recall is the probability that a (randomly selected) good outcome instance is compliant. Accuracy is the probability that a (randomly selected) instance is expectant, i.e. it is either compliant with good outcome or incompliant with bad outcome.

The last two metrics are widely used in epidemiological research. In our context, relative risk is the ratio of the probability of a good outcome occurring in compliant instances to the probability of a good outcome occurring in incompliant instances. The odds of having good outcome given compliance is $Pr(Good|Compliant)/Pr(Bad|Compliant)=(m_{11}/(m_{11}+m_{10}))/(m_{10}/(m_{11}+m_{10}))$, and the odds of having good outcome given incompliance is $Pr(Good|Incompliant)/Pr(Bad|Incompliant)=(m_{01}/(m_{01}+m_{00}))/(m_{00}/(m_{01}+m_{00}))$. The odds ratio, OR, is the ratio of the two, $OR = (m_{11}*m_{00})/(m_{10}*m_{01})$.

By definition, the odds ratio appears as a better candidate for evaluating treatment recommendation systems. In particular, if the odds ratio is great than 1, then it means following the recommendations is more likely to get good outcomes than not following.

2. Results

For experiment, we compare two treatment recommendation systems, one is knowledge-driven, a.k.a. CPG, and the other is data-driven, a.k.a. PSA, using different

evaluation metrics. The data set is from a customer project in China, after anonymity. It consists of 3150 encounter instances of diabetic patients, and the recommendation options are 7 types of treatments: METFORMIN (metformin alone), ARFA (either insulin secretagogues or α-glucosidase inhibitors), DPP4 (either thiazolidinediones or DPP-IV inhibitors), BI (two oral anti-diabetic drugs), TRI (three oral anti-diabetic drugs), INSULIN (insulin alone), and COMBINED (insulin and oral anti-diabetic drugs). For the outcome measure, we use widely adopted clinical ranges: HbA1c <= 6.4: normal; 6.5 <= HbA1c < 7: well controlled; 7 <= HbA1c<9: moderately controlled and 9 <= HbA1c: poorly controlled. Each patient's treatment outcome is labeled by comparing the next HbA1c test result after treatment with the current one. The outcome is labeled 1 (good) if the HbA1c level moved into a lower range, or remained in the well-controlled range, otherwise it is labeled as 0 (bad). In our data set, there are 2574 instances of 3150 are labeled 1 (good).

Table 2 is the recommendation results. Not surprisingly, CPG has a smaller exposed group (i.e. its compliance rate is 2036/3150=0.646) than PAS (i.e. its compliance rate is 2567/3150=0.815), which means the physicians are less prone to comply with guidelines.

Table 2. Recommendation results

CPG	Good	Bad	Total	PSA	Good	Bad	Total
Exposed	1624	412	2036	Exposed	1832	735	2567
Control	563	551	1114	Control	355	228	583
Total	2187	963	3150	Total	2187	963	3150

Table 3 is the evaluation results. Except the recall, CPG outperforms PSA, and it does make sense since that CPG has a lower rate of compliance. Notably, both CPG and PSA have the odds ratio great than 1, which means the adoption of their recommendations has a positive relevance with good outcomes. The much higher value of CPG's odds ratio means that following CPG recommendations is much more likely to get good outcomes than not following.

Table 3. Evaluation results

	Precision	Recall	Accuracy	Relative risk	Odds ratio
CPG	0.7976	0.7426	0.6905	1.5783	3.8577
PSA	0.7137	0.8377	0.6539	1.1720	1.6008

Finally, we observe that if regarding the prescription as also a kind of recommendation, then all instances are in the exposed group and no one is in the control group. That could be our base line, and its precision is 2187/3150=0.6943, the recall is 3150/3150=1.0, and the accuracy is also 2187/3150=0.6943, while the relative risk and odds ratio are both null.

3. Discussion

As pointed in the survey [3], a proper evaluation metric is crucial for the selection of the recommendation system algorithm that will be employed for deployment. However, previous work does not seriously consider the speciality of evaluating treatment recommendation systems – i.e. we are short of ground truth. In general, the customer actual decision (such as the purchase action for movie/book recommendations) is regarded as the ground truth, but in clinical context, the physician's actual prescription

is NOT the ground truth, because the outcome of following the prescription is not always good. This problem is addressed in this paper, by introducing five evaluation metrics (i.e. precision, recall, accuracy, relative risk and odds ratio), each of which has its clear statistic meanings in clinical context from an outcome-driven perspective. In particular, the odds ratio reveals the relevance of following recommendations to good outcomes.

For future work, we plan to extend our evaluation method for more complex tasks. According to the system output, we present different tasks, as shown in Table 4. The output is formalized, where $C = \{d_1, d_2, \ldots, d_m\}$ is the set of all options, $d \in C$ means one option in C, $D \subseteq C$ means a subset options of C. The ranking means the priority of options, while the scoring means the probability of options, with $p_1 + p_2 + \ldots + p_m = 1$.

Table 4. Tasks of treatment recommendation systems

Task	Output	Sample
Recommend: one option	$d \in C$	Recommend to use metformin
Recommend: multiple options	$D \subseteq C$	Recommend to use metformin or sulfonylurea
Recommend: multiple options with ranking	1^{st}: $d_1 \in C$ …m^{th}: $d_m \in C$	Recommend to use metformin first, and then try sulfonylurea
Recommend: multiple options with scoring	$d_1 \in C : p_1$ …$d_m \in C : p_m$	Recommend to use metformin with probability of 0.6, and sulfonylurea with probability of 0.4

Recommendation of one option is a basic task, and other tasks could be simplified to the basic one. For example, if the system outputs multiple options with ranking, then selects the top one, and if the recommendation system outputs multiple options with scoring, then selects the one with the highest score. For the output of multiple options, we may regard the subset as a whole, and the final recommendation is still D, but the recommendation space becomes the power set of C instead of C, i.e. $D \in U(C) = \{D' | D' \subseteq C\}$. In this paper, we only focused on evaluating the basic task, and future work will address the problem of evaluating complex recommendation tasks.

Besides, another interesting future work is the multi-parametric outcome measures. In literature, multiple-criteria decision analysis (MCDA) is a sub-discipline of operations research, and we may leverage them to medicine domain for clinical decision support.

References

[1] Annette ten Teije, Silvia Miksch, Peter Lucas. *Computer-based Medical Guidelines and Protocols: A Primer and Current Trends*. IOS Press, 2008.

[2] Divya Tomar, Sonali Agarwal. A survey on Data Mining approaches for Healthcare, *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 5 (2013), 241–266.

[3] Asela Gunawardana, Guy Shani. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks, Journal of Machine Learning Research 10 (2009), 2935-2962.

[4] Thean Pheng Lim, Wahidah Hsain, Nasriah Zakaria. Recommender System for Personalised Wellness Therapy, *International Journal of Advanced Computer Science and Applications,* Vol. 4, No. 9 (2013), 54–60.

[5] Haifeng Liu, Jing Mei, Guotong Xie. Towards Collaborative Chronic Care Using a Clinical Guideline-Based Decision Support System, *Proceedings of the 24th European Federation for Medical Informatics,* MIE 2012, 492-496.

[6] Haifeng Liu, Guo Tong Xie, Jing Mei, Weijia Shen, Wen Sun, Xiang Li. An Efficacy Driven Approach for Medication Recommendation in Type 2 Diabetes Treatment Using Data Mining Techniques, *Proceedings of the 14th World Congress on Medical and Health Informatics,* MedInfo 2013, 1071.