# Risk Prediction Methods Based on Electronic Medical Records and Social Surveys for Improving Patient Outcomes and Enabling Targeted Care Services[1]

*Charis Kaskiris* [a], *Jakka Sairamesh* [a], *Ram Rajagopal* [b],
*Ravi Nemana* [a], *Keith Argenbright* [c] *and Paula Anderson* [c]

[a] *Advisory Board Company*
*San Francisco, CA 94108, USA*

[b] *Department of Civil and Environmental Engineering*
*Stanford University*
*Stanford, CA 94305, USA*

[c] *Moncrief Cancer Institute*
*University of Texas Southwestern*
*Fort Worth, TX 76104, USA*

## ABSTRACT

This paper focuses on effectiveness of methods for improving patient quality (e.g. improving treatment adherence, reducing adverse events) outcomes and targeted interventions based on psychosocial and clinical risk factors embedded structured and unstructured elements in medical records. Current methods on outcomes analysis such as adherence to treatment regimen largely rely on survey instruments, and provide lagging indicators that inhibits timely intervention and care services. In this paper we present a novel early-warning method that can predict patients at risk of non-adherence based on clinical rules, natural language processing techniques and predictive algorithms applied to risk factor information embedded in electronic medical records. We conducted studies on the effectiveness of our risk estimation methods across 2.5 million patient-visit records from a community cancer clinic that spans a 14 year time-horizon. We identified 2 distinct patient groups, between 26 and 38 (mean risk score, r=0.77, s=0.22), and 75 and 90 (r=0.81, s=0.19) years of age respectively, who exhibited a strong likelihood of non-adherence to treatment regimen. We obtained a reasonably high C-statistic (> 0.77) on predicting outcomes based on the risk factors. The dominant risk-factors, not surprisingly, included psychosocial (e.g. depression and lack of support), medical (e.g. side-effects) and financial (e.g. co-pay). We finally discuss the effectiveness of the methods for targeted and improved health care services.

**Keywords**: Electronic Medical Records, Medical Adherence, Cancer, Predictive Analytics, Text Mining

[1] Funding provided by the National Cancer Institute under National Institute of Health SBIR Grant number: 1R43CA141899-01.

https://openaccess.cms-conferences.org/#/publications/book/978-1-4951-2091-6

# INTRODUCTION

At least one in five cancer patients fails to adhere to a recommended treatment plan of care or medication regimen. For specific cancers this proportion can be substantially higher. These non-adherence leads to higher costs through unplanned hospitalizations, adverse outcomes, and increased risk of preventable death. Early warning of risk of patient non-adherence can provide clinicians with useful information for attempting suitable and targeted interventions. Understanding the factors that drive medical non-adherence is a long standing research effort (DiMatteo, 2004).

We utilize a retroactive dataset of cancer patient visitation information at an outpatient clinic, the Moncrief Institute at University of Texas Southwestern, to create a rich information model based on attributes cleaned from the unstructured fields in electronic medical records. This was achieved through the use of data dictionaries and text mining techniques. These factors are combined with structured information from medical systems are used in calculating the risk of non-adherence using predictive modeling. The models are evaluated and validated through a combination of blinded validation datasets and ultimately by clinical evaluation and validation conducted by clinical staff.

# STUDY DESIGN

Predicting patients at risk of non-adherence to their hospital administered treatment plan through canceling and no-shows of their scheduled visits to the hospital. Specifically our models will attempt to score whether the patient will be a no-show or cancel her next treatment visit. These risk scores will help guide the provision of additional resources towards patients with high-risk profiles of non-adherence in order to maintain their treatment plans and avoid potentially life-threatening situations.

A retrospective study is conducted on past cancer patients who were participants with the UTSW Cancer Institute with patient cases and treatment visits ranging from the end of 1978 through 2012. The focus of the study is on three main categories of cancer, namely breast, colon, and lung cancers as well as an overall one. We utilize two types of model features; first a set of features based on structured information which include frequency and recency of events within different time horizons, medical condition of the patient, prior treatment history, presence of comorbidities, adverse effects; secondly a set of features extracted from medical clinical records utilizing text-mining, clinical rules, and natural language processing which include social factors, financial factors, domestic issues, emotional issues and related concepts (Christiansen and Ehlers, 2002). The design of these text-mining algorithms utilizing dictionaries is described in (Sairamesh 2009a, 2009b).

A critical step in our approach is to validate the terms and variations of terms and phrases in the risk factor dictionary. The validation is done using the frequency and distance vector methods of terms and phrases used by clinicians in various records. We measure error rates between the text in the new records and risk factor terms in the dictionary. A crucial next step is to validate the risk algorithms by estimating the risk based on the risk factor dictionary and verifying with the help of a clinician if the patients projected to be at risk are indeed the patients who are at risk. This process requires comparing the actual versus the computed value of risk.

- First, the algorithms are trained on a large enough training set from the retrospective data. We considered a variety of risk factors (e.g. treatment cycle, side effects, costs, social status and others) and patient attributes (e.g. age, genetics code, vital signs and others) for validation.
- Once trained, the algorithms are then tested on a “testing set” consisting of patient records different from the training set to verify that the risk factors extracted are indeed the dominant factors for patient non-compliance. 10-fold cross-validation is also used to better assess the prediction error.
- A third set of patient records, called the validation set, is then chosen from the newly created patient records for further validation and correction.
- Finally a clinical validation study is contacted to compare the results of the algorithmic models with the results skilled clinicians would identify and compare the two. Models that predict close to the results of the clinical staff are in general better.

# RISK MODELING

The risk modeling component of this study involved the use of different predictive methodologies to predict the risk of non-compliance. The study was done in two parts. One part involved utilizing text-mining features alone and the second part involved the generation of a more extensive set of features utilizing structured information and coupling it with the text-based risk factors. Prior predictive model efforts include adjuvant treatments studies for women with primary breast cancer (Partridge et al, 2003)(McGowan, 2008), and Anastrazol (Partridge et al, 2010) but without utilizing text-mined risk factors.

## Risk Models using Text-Based Risk Factors

We conducted studies on the effectiveness of our risk estimation methods across 2.5 million patient-visit records from a community cancer clinic that spans a 14 year time-horizon. Multiple modeling techniques were utilized and evaluated. Bayesian, Neutral Networks, and nearest neighbor techniques were utilized against the text-based risk factors as well as some basic demographics to predict no-shows and cancellations.

These models allowed us to identify 2 distinct patient groups, between 26 and 38 (mean risk score, r=0.77, s=0.22), and 75 and 90 (r=0.81, s=0.19) years of age respectively, who exhibited a strong likelihood of non-adherence to treatment regimen. We obtained a reasonably high C-statistic (> 0.77) on predicting outcomes based on the risk factors. The dominant risk-factors, not surprisingly, included psychosocial (e.g. depression and lack of support), medical (e.g. side-effects) and financial (e.g. co-pay).

## Risk Models using Combined Features

A subset of the visit data based on type of cancer was used in this modeling effort. We utilized two predictive modeling methodologies namely regularized logistic regression and boosted decision trees to building prediction models. We evaluate both methodologies utilizing performance metrics. We utilized 10-fold cross validation as a way of calculating performance metrics. Models are built for colon, breast, lung and combined (for all three).

Table 1: Population Characteristics of Prediction Models

| | | Breast | Colon | Lung | Combined |
|---|---|---|---|---|---|
| Patients | Patients since onset of cancer | 3,735 | 945 | 1,904 | 6,584 |
| | Total patient visits since onset | 262,681 | 64,505 | 123,206 | 450,392 |
| Adherence | % of visits non-adherent | 8.3% | 7.5% | 5.9% | |
| | % of patients non-adherent | 81.8% | 79.1% | 70.7% | |
| Gender | Female | 3,702 | 452 | 864 | 5,018 |
| | Men | 23 | 492 | 1,036 | 1,551 |
| | Unknown | 10 | 1 | 4 | 15 |
| Age at onset | 65- | 2,543 | 492 | 939 | 3,974 |
| | 65+ | 1,192 | 453 | 965 | 2,610 |
| Survival Rate | | 90% | 80% | 62% | 81% |

The C-Statistic for each of the boosted tree versions of the models are over 0.77 with the most prominent features being the length of time between the last visit and current visit, days since the treatment plan began, stage of cancer, and prior incompletions of treatment visits. The most prominent text-based features were fatigue, nutrition, emotional, and domestic issues.

# CLINICAL VALIDATION STUDY

We aimed to study the risk factors computationally gleaned from clinical notes that predict for administrative noncompliance and overall risk of deviation from a treatment plan as measured by appointment keeping. All datasets and study protocols were submitted to the UT Southwestern Institutional Review Board (IRB) for approval and certification of the study protocol, its human subjects protections and provisions, and aims. Such approval was obtained prior to the project commencement.

Patients who had a confirmed diagnosis of Breast Cancer and Colorectal Cancer as indicated by ICD-9 code were included. We excluded male breast cancer patients because of the different course and indications for therapy. Further, charts of patient who were not living, not receiving therapy (e.g. in survivorship), or participating in a clinical trial during the course of the study period, 2008-2009, were excluded from the sample. This enrollment yielded 3,400 breast cancer patients and 1,879 colorectal cancer patients. This subset was then randomly sampled for three groups of breast (n=125), lung colorectal (n=125) cancer patients that underwent both computational risk scoring and manual, human risk scoring. Specifically 125 patients from each of the three cancer areas were randomly selected from the population of patients within each category provided that the following criteria were in effect: The patient age at point of cancer diagnosis was 65 and above; the patient visit information included a treatment plan, diagnosis, visits, and visit notes; and the patient was diagnosed with either a breast, colon, or lung cancer.

These patients' records and visit notes were made available to clinical staff for evaluation solely through the use of clinical notes. These are retroactive evaluations for patient visit activity as the clinicians were provided the visit note history for each patient and asked to evaluate whether a particular set of risk factors was present.

## Clinical Validation Results

We have contacted a quick analysis to determine the risk factors that are possibly driving the decision of a clinician to assess the level of risk that a patient would have given their medical record information. In our particular study we utilized two expert clinicians for validations to expedite the processing of records. This approach necessitated checking the inter-rater reliability between the two clinicians.

## Inter-rater Reliability

Items such as clinical record evaluations and conclusions with regards to the overall diagnosis often rely on some degree of subjective interpretation by clinicians. Studies that measure the agreement between two or more clinicians should include a statistic that takes into account the fact that observers will sometimes agree or disagree simply by chance. Since the data we have are ordinal in nature (there is an ordering in the categories) we are making use of Cohen's Kappa [2]and Krippendorff's alpha[3]. In both case a value of 1 denotes complete agreement. No agreement between the raters unless by chance when the value is close to 0. Negative values denote disagreement that is worse than random.

What we observe is that over certain areas the two raters had higher agreement namely with factors that pertain to their expertise, namely behavioral health, pain, fatigue, mobility, non compliance and missed appointments, nutrition, depression and anxiety as well as employment. There was disagreement on issues of self image. Overall however the two raters are fairly consistent.

## Clinician Model

We have contacted a CART based model to evaluate the different levels of risk of non-adherence assessments against evaluated risk factors and get some insight into the factors that seem to be affecting the clinician's evaluation on whether the patient is Low (1), Medium (2), or High(3) risk. We achieved this by aggregating the evaluations with the highest agreement between the clinicians.

We run a CART model on the assessments generated by the clinicians using their identification of risk factors and

[2] http://en.wikipedia.org/wiki/Cohen%27s_kappa

[3] http://en.wikipedia.org/wiki/Krippendorff's_alpha

their overall assessment at the end on a per patient basis. The model demonstrates that once the clinician has identified patient information regarding mobility, transportation, nutrition, companionship, pain, and prior missed appointments, ability to pay, anxiety and depression then a determination is assessed on the risk of non-adherence. The most important features as extracted by model are mobility, transportation, nutrition, companionship and pain management.

### Comparisons with Prediction Model

The clinical study was conducted using all visit information while the prediction models are scored on a per visit basis. As such a particular logic needed to be implemented to translate the clinical per-patient assessments to the per-patient-visit prediction scores.

For each patient identified to have one or more of the risk factors by the clinicians should be compared with all the patient visits (encounters) you have our prediction models gleaned from the .

- If the risk factors identified by the model and the risk score for a single patient across all encounters is over 60% matched with risk factors identified by the clinicians for a patient then we call this a match.
- If the risk factors are very different and match less than 60% then we don't have a match on the risk factors. However is the risk prediction score shows that a patient is likely to be non-adherent, and the clinicians say the same then it is a match.

Comparing the prediction results of the overall boosted tree model against the evaluation provided by the clinicians we observe that applying the logic above the text extraction and models have at least 67% overlap between what the clinicians identified and what the text tool and models predicted.

## CONCLUSIONS

We proposed a novel, real-time risk estimation method that uses data-mining and text processing of unstructured text embedded in electronic medical records to detect whether patients are at risk of non-adherence to treatment regimens**.** Our results indicated that over 30% of the patients were likely to drop off treatment based on several risk factors. We validated the predictions of these models against blinded datasets as well as clinical validation with clinicians with a fairly high overlap in their assessments. We also showed that two distinct patient groups (ages less than 40 and over 75) were at a high risk of non-adherence compared to other age groups. Further work needs to be done in leveraging the gleaned risk-factors to target interventions over multiple patient groups for improving adherence, care and quality of life for the patients. These risk scores can now be integrated into early-warning systems for clinicians to use. They allow them to identify the high risk of non-adherence patients and implement proactive protocols to reduce non-adherence through better understanding of the risks posed by the patient and through the improvement of overall care.

## REFERENCES

Christensen AJ and Ehlres SL. (2002), “Psychological Factors in End-Stage Renal Disease: An Emerging Context for Behavioral Medicine Research” J Consult Clinical Psychology 70:712-24

DiMatteo, MR. (2004), “Variations in Patient’s Adherence to Medical Recommendations: A Quantitative Review of 50 Years of Research” Med Care, 42:200-209

McGowan et al, (2008), “Cohort study examining Tamoxifen adherence and its relationship to mortality in women with breast cancer” British Journal of Cancer 99: 1763-1768

Partridge A. et al. (2004), “Non-adherence to adjuvant Tamoxifen therapy in women with primary breast cancer” Journal of Clinical Oncology, 21.4:602-606

Partridge A. et al (2010), “Adherence to initial adjuvant Anastrazole therapy among women with early-stage breast cancer” Journal of Clinical Oncology, 26.4:556-562

Sairamesh J et al. (2009a), “Early  Warning Methods and Risk Assessments for Improving Cancer Patient Care” Proceedings of the AMIA Conference, San Francisco

Sairamesh J et al. (2009b), “Risk Estimation Methods for Improving Patient Care” Proceedings of the INFORMS DM-SI Workshop, San Diego.