Predictive Patient Care: Survival Model to Prevent Medication Non-adherence

T. Janssoone[1(✉)], P. Rinder[1], P. Hornus[1], and D. Kanoun[2]

[1] Semeia, Paris, France
tjanssoone@semeia.io
[2] Clinique Pasteur, Toulouse, France

Abstract. Adherence in medicine is a measure of how well a patient follows their treatment. Not following the medication plan is actually a major issue as it was underlined in the World Health Organization's reports (http://www.who.int/chp/knowledge/publications/adherence_full_report.pdf). They indicated that, in developed countries, only about 50% of patients with chronic diseases correctly follow their treatments. This severely compromises the efficiency of long-term therapy and increases the cost of health services.

In this paper, we report our work on modeling patient drug consumption in breast cancer treatments. We test a statistic approach to predict medication non-adherence with a special focus on the features relevant for each approach. These characteristics are discussed in view of previous results issued from the literature as well as the hypothesis made to use this model.

Keywords: Adherence · Survival risk analysis · SNIIRAM

1 Introduction

During the last decades, patient-administered oral medications have become more and more prevalent [5,13]. This shift in anticancer treatments has increased the focus on adherence [14] defined as *"the extent to which patients take their medications as prescribed by their healthcare providers"*.

A common solution is to set patient support programs that include, for example, (1) providing information, patient counseling, (2) support and coaching sessions delivered by nurses (by phone or face-to-face), and (3) sending information to health professionals treating the patient. These programs have been shown to be effective, for example, pharmacist coaching has improved adherence by 12% [8] and an SMS based recall system showed a 10% improvement in adherence [15]. Yet, there are two main limitations to these interventions: (1) The use of human intervention is effective but very expensive limiting their reach, (2) The use of digital technologies (notifications and explanations) is too generic and sometimes too intrusive (daily reminders) which leads to patients losing interest. To optimize the relevance of these interventions, we propose using machine

I. Rekik et al. (Eds.): PRIME 2018, LNCS 11121, pp. 42–50, 2018.
https://doi.org/10.1007/978-3-030-00320-3_6

learning techniques on the consumption data of patients with breast cancer. We aim to give to each of them a risk index reflecting their adherence to the treatment. This allows us to predict the most appropriate moments to notify the ones really needing help. Thus, people will benefit from support adapted to their profiles and needs, and human interventions will be reserved for situations that are really critical.

To determine the categories of patients at risk and the appropriate moments to contact them, we develop predictive models built on anonymized data. These predictive models are trained on the reimbursement data of the French Health Insurance (SNIIRAM - the French National Health System).

In the rest of this paper, we first review previous approaches, then we introduce our models and discuss our results.

2 Related Work

Given the importance of the phenomenon of non-observance, many surveys have tried to identify their determining factors. This allows us to improve interventions and, therefore, compliance with the treatment. The review of many observation based scientific publications provides an interesting quantitative assessment of the research conducted on the subject [3,10]. This meta-analysis indicates that the increasing age of patients, and the treatment complexity level (multiple drugs, injections, ...) are essential factors of non-adherence. Similarly, low education and, more importantly, low income are correlated with lower adherence. Another study highlights the impact of patients' mental health shows that depressive episodes have a very negative impact on the patient's compliance to the prescriptions of health professionals [4].

Yet, other DiMatteo studies show that other factors also influence adherence. For example, in distinguishing between the objective severity of the patient's illness and their awareness of the severity of their pathology, they point out that the patient's beliefs influences the level of compliance, and not the actual severity of the condition. This enforces the importance of the role of patient education in strengthening their adherence to their treatment. Similarly, other analyses highlight the effects of modifiable factors in non-compliance. A meta-analysis thus shows the influence of the patient's entourage (support of their spouse, family, relatives and the wider social environment) in the proper monitoring of his treatments [2].

These studies provide a priori indications for detecting risk profiles of non-adherence. At the same time, they highlight the interest of identifying and accompanying these patients in taking their medication.

However, Franklin et al. underline the difficulty to use this information to predict adherence [6]. They evaluate different approaches, using logistic regression and boosted logistic regression, to define three categories of adherence predictors. Hence, they show that using census information or transaction data leads to poor prediction. However, they point out that using adherence observations during the first month significantly increases the accuracy of the results.

This nuance on the weight of each adherence prediction variable is confirmed in [9]. They use random survival forests highlights to find patient specific adherence thresholds to discriminate between hospitalization risks. Here again, the major variables are linked to patient history and previous transactions.

We propose in this paper to explore these solutions to predict the risk of a illegitimate stop during a treatment.

3 The SNIIRAM Database

3.1 Introduction

In order to optimize the use of human intervention and improve the use of digital technologies, we propose the use of machine-learning techniques on breast cancer patients' consumption data. The goal is to categorize patients into risk classes according to their characteristics. The long-term goal is to know the most appropriate moments to contact them for support. Thus, people will benefit from support adapted to their profile and their needs, and human interventions will be reserved for the situations for which they are really necessary. To determine these categories of patients at risk and these appropriate moments, we develop predictive models built on anonymized data.

These predictive models are trained and tested on the reimbursement data of the French Health System (SNIIRAM). SNIIRAM is one of the largest structured databases of health data in the world. The use of this massive data allows the application of complex models and the detection of weak signals. Useful data are, for example, hospitalizations, drug purchases or contextual patient information (age, government services, geographic information, ...). More details can be found in [16]. Previous work has already shown the value of massive data mining to aid diagnosis, either by taking all the information for a "static" approach [12], or, more recently, by also incorporating dynamic information [11]. Other studies have been conducted on the determinants of compliance, particularly for breast cancer.

Our study focuses on women's breast cancer on part of the SNIIRAM data. The cohort of the study consists of 50% of women (drawn randomly) who meet the following criteria:

- diagnosed with breast cancer
- having purchased at least one of the following molecules for the studied period: *Anastrozole*, *Capecitabine*, *Cyclophosphamide*, *Etoposide*, *Everolimus*, *Exemestane*, *Lapatinib*, *Letrozole*, *Megestrol*, *Melphalan*, *Tamoxifen*, *Toremifene* and *Vinorelbine*

Extraction concerns consumptions between 2013 and 2015 and is made up of three main categories:

- Pharmacy transactions (molecule, number of doses, date, ...)
- Hospitalizations (diagnosis, start date, end date, ...)

- Patient information (age, department, date of the diagnostic of eventual long-term illness (referred as *ALD*), pathologies, ...)

The aim of the study is to follow the entire care course, so the studied population must be representative. A discussion with the experts of the CNAMTS (French National Fund for the Health Insurance of Employees) allowed to fix a threshold: a period of 6 months without consumption of at least one of the target molecules is sufficient to consider that the person was not receiving a treatment.

A preprocessing has been done on the 'seniority of ALD' variable which represents the number of days since the diagnosis of the disease (stated in ALD 30). This variable has the characteristic of containing some extreme values, which bias the estimation for models assuming a linear effect. Thus, a common logarithm is used for the study to eliminate this bias. This still allows to keep the order of magnitude of the duration (in days, weeks, months or years).

3.2 Phases of Treatment

The raw data has been reworked to show the different phases of the treatment. A phase is a period of continuous intake of a molecule or hospitalizations for chemotherapy or radiotherapy. This allows the reconstruction of the patient's care path.

The criterion for identifying an end of a phase is the existence of a period of two months after the median time covered by the last purchase (or hospitalization) without a new purchase of the molecule (or hospitalization of the same type). For medication, days of hospitalization are excluded from this period. For example: the median time between two chemotherapies is three weeks. If there is a period of 2 months and 3 weeks without chemotherapy, this is considered a break in the phase.

The phase of treatment is regarded as censored by one of the legitimate stops (death, switch of treatment, some kind of serious cardiac issue or beginning of palliative care) if this event occurs less than two months after the date of the last theoretical dose. For example, if a patient bought a box of 30 pills on January 1^{st}, the event has to occur before March 31 ($30 + 2 \times 30$). The date of the last theoretical dose is obtained by calculating the median interval between two purchases of the molecule or two hospitalizations of the same type: this median behavior is considered to be in conformity with the posology. Thus, the median time is 30 days between two box purchases of 30 doses of tamoxifen. The end of this period after the last box purchased corresponds to the date of the last theoretical take. The date of death is present in the initial data, the switches are identified by the beginning of a new phase of treatment and palliative care as a main diagnosis (which is spotted with a "Z515" tag in the database). Censorship of data caused by the end of the extraction period (end of December 2015) is also considered a legitimate stop. If the data extraction end date is less than two months from the last theoretical consumption, then, in the same way as for legitimate stops, the processing phase is considered censored.

For each phase, the following data is calculated:

- Start and end dates, number of intakes or hospitalizations, molecule or type of hospitalization
- End of treatment type (switch, death, stop, right censorship)
- Patient information (comorbidities, number of consultations in the first year of treatment, age, ...)
- Interventions on the breast (mastectomy) during the three months before the studied phase

In this paper, we propose evaluating different ways to model whether the end of a phase is legitimate or not. We focus on the consumption of *Tamoxifen* as it is the most used molecule. In addition, this molecule is prescribed for up to 10 years, so no patient is supposed to have stopped their treatment during the observation period (3 years) because due to the end of their prescription.

4 Our Model: Survival Analysis

To measure the rate of non-persistence over time, we use the Kaplan-Meier estimator [7]. This estimator uses non-parametric statistic to evaluate the survival function on a state takes. For example, it is used to estimate the amount of patients living for an amount of time after a treatment, the time-to-failure of machine parts,... Its force is to take into account the censored data, in particular by right censorship, each observation being weighted according to the number of observations censored previously. The four factors of censorship are: switch, death, palliative care and end of the extraction. The duration of 'survival' in the treatment phase is thus estimated by taking into account censorship factors such as legitimate end of treatment as well as censorship linked to the end of extraction (end of 2015).

The Fig. 1 shows the variations of the hazard function representing the treatment dropout rates as a function of time. There is a high drop-out rate at the beginning of the phase, during the first 150 days. During the first 5 months, the curve is significantly higher than the rest of the values. This period of high risk will therefore be the most beneficial period to help patients. Kaplan-Meier estimator allows us to analyze survival but we need to use a regression model to examine the factor influence of the different variables.

We use a Cox model [1] to identify the characteristics related to poor adherence. The Cox regression estimates a fixed effect of each variable in relation to the patients' average behavior. It is based on two strong assumptions:

(1) the expected effect of each variable is linear
(2) the effect of each variable does not vary over time. An example in our case is that, if the weight found for the *CMU-C* variable is 1.40, we assume that a person who benefits from *CMU-C* has 40% more risk to discontinue their treatment than someone who doesn't benefit from *CMU-C*

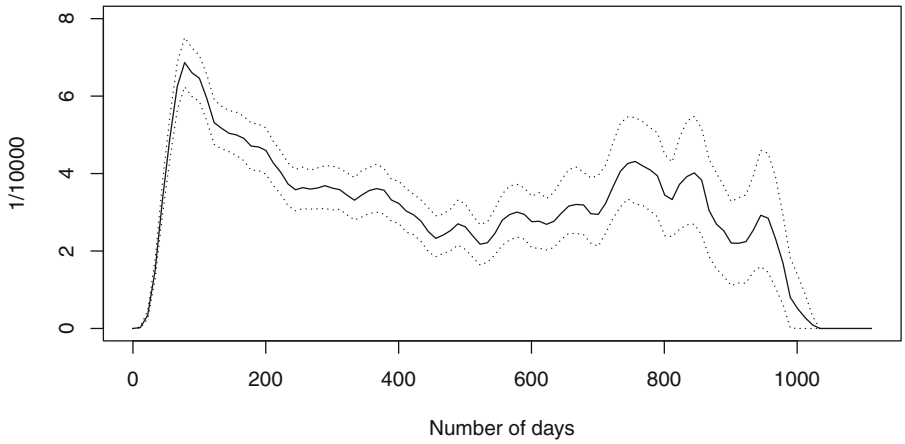

Fig. 1. Kaplan-Meier hazard function for Tamoxifen which represents the rate of failure (here drug drop-out) at time t. The plain curve represents the estimated value and the dotted ones represent the 95% confidence intervals (computed with bootstrapping).

The explanatory variables in the Cox model are characteristics of the phase, the patient pathway and patient profile. In order to extract the most significant variables and to robustly estimate the associated coefficient, the model was estimated with the following iterative process:

- estimation of the coefficients for the set of variables,
- selection of the variables having a p-value lower than threshold of 0.05
- new estimation of the coefficients for the model limited to the selected variables

The most influential coefficients are presented in Table 1. The characteristics highlighted by the literature are found to be influential to the patient's adherence to their treatment. We can then evaluate the impact of the age, social support, or previous illness (psychiatric, mastectomy, ...). We also underline that the treatment preceding the current phase has a major influence.

This prompted us to analyze this value in particular. The Fig. 2 shows the different lifelines depending on the treatment preceding the current Tamoxifen phase. We see three types of influences. First, the classic pathway: a hospitalization (here chemotherapy or radiotherapy) preceding the current phase shows the lowest risk of abandonment. Second, a hormonotherapy other than Tamoxifen was used, corresponds to a *switch* of treatment and shows a higher risk. Third a Tamoxifen phase was used before the current phase. This case suggests that an illegitimate stop has happened before the current phase, and could explain the highest risk this case has. Nevertheless, this underlines the interest of our model to find additional information to predict the evolution of Tamoxifen intake phases. As illustrated in the Fig. 3. We can use the background information of the patient to compute a score at the beginning of a *Tamoxifen* phase. Then, we

Table 1. Weights computed with the Cox-regression showing their different influence. The odd ratio indicates the impact of the variable on the average risk (1.5 means 50% more risk of non-adherence). The first ones are indicators of income (CMU-C (Supplementary universal health cover) and ACS (Assistance with the acquisition of supplementary health insurance) and their p-value show their influence.

	Cox computed coefficient	Odd ratio (exp(coefficient))	p-value
CMU-C	3.84e−01	1.47e+00	6.62e−04
ACS	4.16e−01	1.52e+00	1.87e−03
Time since ALD status (log)	7.88e−02	1.08e+00	3.91e−02
Number of medical consultation	−5.70e−03	9.94e−01	1.19e−02
Psychiatric illness	1.78e−01	1.19e+00	9.31e−03
Recent hospitalization with diagnostic C50: Malignant neoplasms of breast	−3.83e−01	6.82e−01	3.59e−07
Last treatment - tamoxifen	9.12e−01	2.49e+00	<1e−10
Last treatment - radiotherapy	−7.09e−01	4.92e−01	<1e−10
Last treatment - chemotherapy	−8.62e−01	4.22e−01	<1e−10
Menopause	1.17e−01	1.12e+00	2.76e−02

select the most accurate survival function which gives the probability of abandonment during the specific number of days of treatment, which allows us to predict the abandon risk over time.

Yet, we based our model on the strong assumption that the effect of each variable does not vary over time as explained in (2). This proportional hazard assumption can be checked with a Schoenfeld individual test and visualized with the log-log plot of survival displayed in the Fig. 4. Using Schenfeld residual, we obtain p-values that verify this hypothesis except for the assumption concerning the previous treatment. This indicates that the previous treatment has different

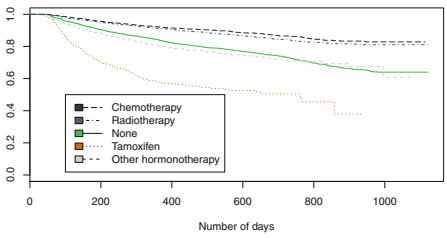

Fig. 2. Kaplan-Meier plot of Tamoxifen survival function for each previous treatment phase (hospitalization are chemotherapy or radiotherapy)

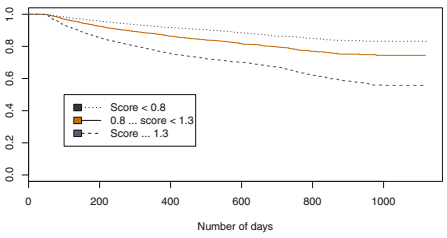

Fig. 3. Evolution of the different survival curves given a specific score computed at the beginning of the phase

influences over the duration of the current phase and that our assumption was too strong for this variable. However, the Fig. 4 illustrates whether the hazards are approximatively proportional throughout. We can see that this assumption remains valid after the 20 first days. This analysis gives us insights to improve our model with a special focus to put on the beginning of the phase.

The next step of this study is to challenge our model with machine-learning approaches and other statistical models to improve our predictions.

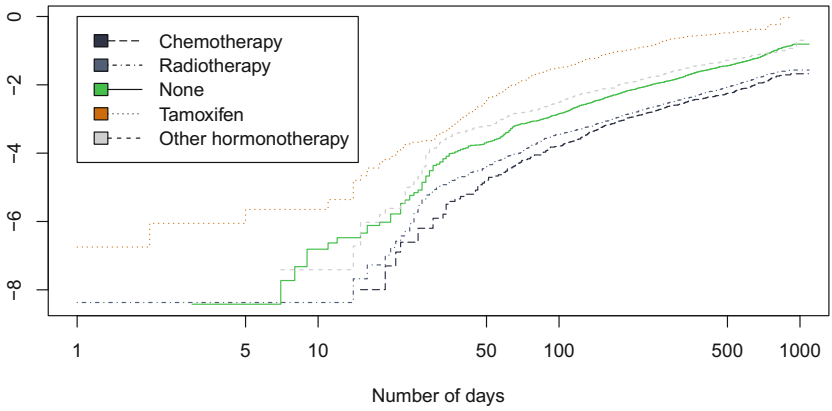

Fig. 4. Log-log plot of survival for each previous treatment. We observe a validation of our model after the 20 first days

5 Conclusion and Future Work

In this paper, we show the importance of the survival function analysis to predict the adherence of a patient to their treatment during of a phase of medication. The study of explanatory variables gives more insights about the different cases of patient courses. We applied this method to the study of *Tamoxifen*: we validates our approach by retrieving information in agreement with the literature. We also found other explanatory variables that we can use to compute more accurate risk estimations for patients. These risks could trigger alerts that indicate the patient's need of support. With an appropriate response, this could lead to improve the patient's adherence to their treatment.

This predictive model allows us to validate the possibility of evaluating the risk of abandonment during a phase of the treatment. We plan to challenge this approach with other algorithms, especially sequence-mining and deep-learning which are very suitable to the amount of data provided by SNIIRAM. One major advantage to our statistic-based approach is that they are less parameters to fine tune compare to machine-learning approaches. We are planning to survey these different kind of analysis to provide a meta-comparison of this tools to the community.

References

1. Cox, D.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**(2) (1972)
2. DiMatteo, M.R.: Social support and patient adherence to medical treatment: a meta-analysis. Health Psychol. **23**(2), 207 (2004)
3. DiMatteo, M.R.: Variations in patients adherence to medical recommendations: a quantitative review of 50 years of research. Med. Care **42**(3), 200–209 (2004)
4. DiMatteo, M.R., Lepper, H.S., Croghan, T.W.: Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch. Intern. Med. **160**(14), 2101–2107 (2000)
5. Fallowfield, L., et al.: Patients' preference for administration of endocrine treatments by injection or tablets: results from a study of women with breast cancer. Ann. Oncol. **17**(2), 205–210 (2005)
6. Franklin, J.M., et al.: Observing versus predicting: initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. Health Serv. Res. **51**(1), 220–239 (2016)
7. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**(282) (1958)
8. Krolop, L., Ko, Y.D., Schwindt, P.F., Schumacher, C., Fimmers, R., Jaehde, U.: Adherence management for patients with cancer taking capecitabine: a prospective two-arm cohort study. BMJ Open **3**(7), e003139 (2013)
9. Lo-Ciganic, W.H., et al.: Using machine learning to examine medication adherence thresholds and risk of hospitalization. Med. Care **53**(8), 720 (2015)
10. Mann, D.M., Woodward, M., Muntner, P., Falzon, L., Kronish, I.: Predictors of nonadherence to statins: a systematic review and meta-analysis. Ann. Pharmacother. **44**(9), 1410–1421 (2010)
11. Morel, M., Bacry, E., Gaïffas, S., Guilloux, A., Leroy, F.: ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. arXiv preprint arXiv:1712.08243 (2017)
12. Neumann, A., Weill, A., Ricordeau, P., Fagot, J., Alla, F., Allemand, H.: Pioglitazone and risk of bladder cancer among diabetic patients in france: a population-based cohort study. Diabetologia **55**(7), 1953–1962 (2012)
13. O'neill, V., Twelves, C.: Oral cancer treatment: developments in chemotherapy and beyond. Br. J. Cancer **87**(9), 933 (2002)
14. Osterberg, L., Blaschke, T.: Adherence to medication. N. Engl. J. Med. **353**(5), 487–497 (2005)
15. Spoelstra, S.L., et al.: An intervention to improve adherence and management of symptoms for patients prescribed oral chemotherapy agents: an exploratory study. Cancer Nurs. **36**(1), 18–28 (2013)
16. Tuppin, P., De Roquefeuil, L., Weill, A., Ricordeau, P., Merlière, Y.: French national health insurance information system and the permanent beneficiaries sample. Revue d'epidemiologie et de sante publique **58**(4), 286–290 (2010)