An Early Warning System that combines Machine Learning and a Rule-Based Approach for the Prediction of Cancer Patients' Unplanned Visits

H. F. Witschel[1], E. Laurenzi[1], S. Jüngling[1], Y. Kadvany[2] and A. Trojan[2]

[1]*FHNW University of Applied Sciences and Arts Northwestern Switzerland, Riggenbachstrasse 16, CH-4600 Olten*
[2]*mobile Health AG, Falkenstrasse 21, CH-8008 Zürich*

In this paper, we present the early results of a hybrid intelligent approach that consists of an interpretable rule-based machine learning model for the prediction of unplanned visits of cancer patients. The approach is contextualized within the area of personalized medicine and will contribute to the development of an early-warning system (EWS) whose goal is to support cancer patients to cope with their daily symptoms remotely, by avoiding as much as possible physician visits. The interpretability of rules makes it possible to involve medical experts in the learning process who can accept, reject or modify rules, e.g. by adding conditions to increase their precision. The results appear to be promising as the discovered rules provide value in the identification of critical situations of patients. Experts also suggested the modification of rules for recommending not only visits to a physician, but also other (less costly) actions, such as increasing the dosage of pain killers – an extension to the EWS that would not have been possible without our hybrid approach. Overall, our first experiments showed how a new form of "dialogue" between the experts and the machine learning algorithm started to emerge.

1. Introduction

In this position paper, we describe ongoing work in the area of personalized medicine. Our goal is to develop an early-warning system that can help cancer patients get additional medical insights – above all judge the criticality of their current health status – making them aware of when exactly contacting or visiting a physician is required.

We intend to reach this goal by implementing a rule-based warning system, where rules are constructed via a combination of knowledge engineering, based on the expertise of physicians, and machine learning applied to patient diary data.

2. Related Work

There is ample work around the automated prediction of medical events and conditions, e.g. septic shock [1], cardiac arrest [2], suicide attempts [3] or hospital re-admission [4]. A survey

of clinical risk prediction approaches in general [5], including a wide variety of approaches based on data analyses, concludes that machine learning (ML) is generally the most successful of these approaches.

While ML seems to be successful in improving predictions, early-warning systems (EWS) based on such predictions are not necessarily considered helpful by clinicians [1], e.g. because the reasons for raised alerts are not understood. The study revealed a lack of trust, caused primarily by poor transparency of ML models. As argued by Rudin [6], one should not rely on explanation models applied to black-box ML models, but rather on genuinely human-interpretable models to create such transparency. Rudin also points out that such interpretable models can be competitive with e.g. deep learning when meaningful structured features exist and that their transparency can result in better models through improved insights from the testing phase. In fact, other researchers point out that the value of transparent or directly interpretable ML models might be limited or even harmful when human experts are trying to challenge an ML model based on a (sometimes overwhelming) explanation. [7, 8]. Instead, as Ghassemi et al. [7] point out, "these methods are incredibly useful for model troubleshooting and systems audit, both of which can be used to improve model performance." That is, there is a strong belief that studying an ML model can be an inspiration for experts and expert feedback can lead to improved models.

Expert knowledge is frequently used to improve ML models, e.g. by weighting features [9] or augmenting training data [10]. Gennatas et al. [11] have shown how to improve a rule-based ML model by integrating human expertise, by studying discrepancy in human and ML assessment of clinical risk.

Our study is based on a similar idea. However, contrary to [11] and following the arguments of Rudin [6], we choose to work with a learning algorithm that produces directly interpretable rules. With this, a rich set of possibilities for interaction between human and ML emerges. Of course, other directly human-interpretable models exist that would lend themselves to such interactions, above all Bayesian Networks, where experts may intervene by estimating prior probabilities, and which have been successfully used for such purpose [12, 13]. However, as pointed out e.g. by Botsas et al. [14], Bayesian approaches address and require knowledge about (internal) parameters of the ML model whereas rules can be interpreted, formulated and enhanced by experts based solely on an understanding of input-output associations. This is why we focus on rule-based models.

3. Background

The focus of our early-warning system (EWS) is on cancer patients, undergoing diverse treatments with sometimes considerable side effects. While some of these side effects are "normal", we want to predict actually critical situations, which may be due to side effects or other conditions in which the patients should quickly consult a physician.

3.1. Data set

Our starting point is a diary kept by cancer patients in the form of a mobile app where they enter, on a daily basis, their general well-being, current symptoms, treatments applied and

Attribute(s)	number	description	type, value range
Birth year	1		numeric
Sex	1		{male,female}
Primary tumor	1		{breast, gut, blood/lymph, lung, prostate}
Wellbeing	1	subjective wellbeing of patient	numeric [0...100]
Therapy form	1	frequency of treatment	{daily, weekly, bi-weekly, 3-weekly, 4-weekly}
Drugs	88	cancer and other drugs	numeric [1,nan]
Symptom strengths	52	strength of relevant symptoms	numeric [0...1,nan]
Diagnosis terms	246	Terms occurring in patient's diagnosis	numeric (TF/IDF)
Note terms	311	Terms from patient notes	numeric (TF/IDF)
Unplanned visit	1	Class attribute	{yes,no}

Table 1
Attributes used to train our EWS

free-text notes. When patients entered the study, their initial diagnosis is captured as free text, together with their date of birth, sex, primary tumor, diagnosis and therapy start date, as well as frequency of therapy (from daily to 4-weekly, see also Table 1). For more details regarding the origin of the data, see [15, 16]. On the whole, our dataset comprises 16,670 diary entries of 266 patients, most of them suffering from breast cancer.

3.2. Task definition

The data also contains information regarding unplanned visits of these patients to their physician or to the hospital, which we equate with situations where a warning should have been raised and which we thus want to learn to predict. The ground truth for training our early-warning system (EWS) comes from these unplanned visits of patients.

In our data, each combination of patient and day represents an instance, i.e. a training example. Based on our knowledge of unplanned visits, we associate a class attribute "unplanned visit" with each such instance. We set the value of this attribute to "yes" not only on the day the unplanned visit occurred but also on the three days before – based on the goal of constructing an *early* warning system that should foresee problems at least some days ahead. The medical expert involved in our study estimated that a time horizon of three days should be realistic.

3.3. Attributes

All attributes are summarised in Table 1.

Note that drugs and symptom strengths are encoded via a series of attributes, each representing the presence or strength on the given day. For encoding symptom strengths on a scale from 0 to 100, patients receive a guideline with definitions and descriptions of values that have been carefully developed by oncologists and based on the Common Terminology Criteria for Adverse Events (CTCAE[1]).

---

[1] https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm

Since patients will only actively enter symptoms and medication that have actually occurred on a given day, symptom strengths and values for drugs are mostly unavailable. We have chosen to represent these as missing values instead of assigning a value of 0 because discovered rules will otherwise contain conditions such as "stomach ache = 0". While one can envision situations in which such conditions might be useful, representing the non-presence of symptoms or drugs by 0s resulted in too many useless rule conditions in our first experiments.

We have used the free-text attributes "diagnosis" and "note" which represent a patient's diagnosis details written by a physician upon entering the study (has the same values for all days where the given patient was part of the study) and notes (optionally) captured by the patients on a daily basis. We have vectorised these string attributes using TD/IDF weights, resulting attributes were prefixed with "$diag\_$" and "$note\_$", respectively.

Although enhancing semantics through e.g. word embeddings might have improved the results, such approaches were deliberately not used to ensure a maximum of readability of discovered rules.

Obviously, the set of attributes used here represents a first "starting point" for the analysis. Advanced feature engineering may e.g. also incorporate a certain history of wellbeing or symptom development over time or look at missing patient entries of previous days etc. However, such feature engineering is not in the center of the current study and will be done at a later stage.

4. Approach

While our work is ongoing, we have already gathered insights during a workshop with a medical expert and while analysing the diary data with a rule-based classifier [17]: In our initial experiments and discussions, we found that

- expert-defined rules tend to be more generic than ML-provided ones, i.e. ML may be able to contribute specific patterns that experts do not readily think of
- it is possible to derive a manageable set of interpretable rules with the above-mentioned ML algorithm including also textual features from e.g. notes in patients' diaries. Some of these rules are shown and discussed in Section 5 below.
- building an early-warning system usually implies working with a heavily imbalanced data set. This is also true in our case: critical situations are rare, i.e. out of the available 16,670 patient-date combinations, only 166 (1%) represent cases where the patient had an unplanned visit within the next 3 days, i.e. a situation where a warning should be raised.

To account for the last of these points, we use cost-sensitive classification [18] to make the rule learner more sensitive, assigning a far higher cost to false negatives (i.e. unplanned visit not recommended when it is necessary) than false positives (i.e. unplanned visit recommended when it is not necessary). However, we do configure the rule learner to generate rules with a minimum support of 5, such that rules originating from just one unplanned visit of one patient are ruled out.

However, this will still result in a possibly high number of rules that have rather low support and may not generalize well, requiring human inspection. For rules with medium or low support,

one of the major contributions of the medical expert is thus to judge whether a rule describes a critical situation that is rare, but valid (i.e. may occur also in other patients and should lead to raising a warning) or whether the rule is based on peculiarities of the training data that will not generalize to other patients.

Based on these findings, we propose the following division of labor between ML and expert:

1. The human expert states a set of rules A.
2. We apply the rule learner [17] to generate a set of rules B.
3. Rules from both A and B are evaluated based on a cost matrix where false negatives (FNs) have higher cost than false positives (FPs, false alarms). Rules are ranked by cost.
4. ML rules are inspected by the human expert in the ranked order. The expert can suggest to drop a rule, but also to modify it, e.g. dropping or adding a condition. Modified rules will be evaluated and accepted if their cost on the test set is acceptable.
5. A detailed error analysis is performed on each resulting rule, eliciting causes of both FNs and FPs. We expect that this can result in e.g. creation of new features or re-sampling of training data.
6. Modifications are made based on insights from step 5 and the entire process is repeated from step 2 until no further improvement results in step 3. Given that the human expert might receive new insights in the process, iteration may even start from step 1.

While several steps in this process are common in interactive ML, the novelty of our approach lies in the explicitness and degree of interpretability of the ML model, enabling fine-grained interventions in steps 4 and 5, where ML and human can exchange knowledge using the same language (of rule conditions). These interventions will be illustrated in the next sections.

5. Preliminary Findings

As a first step towards validating our suggested approach, we performed its first 4 steps. Step 1 of the process was done by means of a workshop where the medical expert was invited to first look at a number of example cases where unplanned visits had happened. In a second step, the expert formulated some rules, which are shown in Figure 5. The expert developed these rules by thinking of the most frequent categories of problems that his patients tend to have, e.g. lung, cardiac, respiratory, kidney problems, infections or side effects of cancer drugs – see last column of the table. Interestingly, although the rules were at least in some part inspired by the examples that we previously discussed, the rules themselves – when applied to the data – do not match any of the situations where an unplanned visit occurred within 3 days. This may indicate several issues, e.g. that there might be critical situations in the data that did not lead to an unplanned visit – but where raising a warning might have nevertheless been beneficial. On the other hand, it may also indicate how hard it can be to formulate rules that are capable of predicting unplanned visits.

In the second step of our process, we applied the rule learner to our data set. We applied cost-sensitive classification with a range of different cost matrices. The results we report here were obtained by assigning a cost of 1 to false positive predictions ("false alarms") and a cost of 10 to false negatives, i.e. critical situations that are not recognised. This approach was confirmed

Rule	matches	incorrect (false positives)	correct	Precision	Recall	F measure	Cost	Problem Category
diag_nikotine_abuse >= 0) and (Chest_Pain >= 0) => UnplannedVisit=yes	26	26	0	0.0%	0.0%	0	1686	Lung
(Weight_Gain >= 0) and (Fatigue >= 0) and (Shortness_of_Breath >= 0) => UnplannedVisit = yes	14	14	0	0.0%	0.0%	0	1674	Cardiac / respiratory
(Weight_Gain >= 0) and (Fatigue >= 0) and (Chest_Pain >= 0) => UnplannedVisit=yes	55	55	0	0.0%	0.0%	0	1715	Cardiac / respiratory
(Weight_Gain >= 0) and ((Shortness_of_Breath >= 0) or (Cough >= 0)) => UnplannedVisit = 0	31	31	0	0.0%	0.0%	0	1691	Kidney
(Fever >= 0) and (Fatigue >= 0) and (Joint_Pain >= 0) and ((Cough >= 0) or (Diarrhea >= 0) or (Burning_during_Urination >= 0) => UnplannedVisit=yes	7	7	0	0.0%	0.0%	0	1667	Infection
(Loss_of_Taste/Appetite >= 0) and (Fatigue >= 0) and (Weight_Loss >= 0) => UnplannedVisit=yes	27	27	0	0.0%	0.0%	0	1687	Side effects of cancer drugs

Figure 1: The rules provided by the medical expert

by our medical experts who stated that having several more false alarms is acceptable when one can discover additional critical situations and thus alleviate patients' problems more effectively.

While using the same cost for both types of errors (i.e. a ratio of 1:1) did not produce *any* rules – i.e. the rule learner would always predict "no" – the number and properties or rules were rather similar when using e.g. a ratio of 1:20 instead of 1:10.

We then computed the cost of each single rule, as well as its precision, recall (which is expectly small for each individual rule) and F-measure. Following step 3 of our process, we then sorted the rules by cost. Figure 5 shows the 6 rules with lowest cost.

Rules	matches	incorrect (false positives)	correct	Precision	Recall	F measure	Cost
(diag_clip_marker >= 4.661854) and (Wellbeing <= 74) => UnplannedVisit=yes	53	41	12	22.6%	7.2%	0.110	1581
(Wellbeing <= 46) and (Endoxan >= 1) and (diag_lymphangiosis >= 5.296978) => UnplannedVisi	15	8	7	46.7%	4.2%	0.077	1598
(Pain >= 0.38) and (note_today >= 2.30896) and (Wellbeing <= 60) => UnplannedVisit=yes	17	10	7	41.2%	4.2%	0.077	1600
(Pain >= 0.37) and (Loss_of_Appetite <= 0.43) and (Fatigue >= 0.44) => UnplannedVisit=yes.	14	8	6	42.9%	3.6%	0.067	1608
(Wellbeing <= 41) and (Sleep_Disorder <= 0.61) and (Shortness_of_Breath <= 0.58) and (diag_	8	3	5	62.5%	3.0%	0.057	1613
(BirthYear <= 1954) and (Wellbeing <= 65) and (Fever <= 0.18) => UnplannedVisit=yes	24	18	6	25.0%	3.6%	0.063	1618

Figure 2: The 6 best rules, according to their cost, discovered by the rule learner

5.1. Performance of machine-learned rule set

The results in Figure 5 were obtained by learning rules from and applying them to the entire training set. To get an impression of the performance of the corresponding model, we additionally performed a 10-fold cross-validation. The confusion matrix obtained in this evaluation is shown in the lower left corner of Figure 5.1.

Overall, our machine-learned rule set discovered 47 out of the 166 critical situations (28.3% recall), while generating 263 false alarms (15.2% precision). By applying the 1:10 cost matrix to the confusion matrix, we see that a cost of 1453 results for our rule set, compared to a cost of 1660 that the baseline achieves. When we use a ratio of 1:20 for the cost-sensitive rule learner, the model will recognise 54 critical situations (i.e. 7 more than with the 1:10 model), but at a cost of an extra 104 false alarms, i.e. overall 367 false positives.

	Predicted visit	
	Yes	No
Unplanned visit — Yes	0	10
No	1	0

	Predicted visit	
	Yes	No
Unplanned visit — Yes	47	119
No	263	16241

	Predicted visit	
	Yes	No
Unplanned visit — Yes	0	166
No	0	16504

Figure 3: The cost matrix used for evaluation (top), as well as the confusion matrix of our rule learner (bottom left) and a baseline that always predicts "no" (bottom right)

5.2. Interpretation and modification by experts

Finally, we performed step 4 of our process by discussing the 12 best rules (according to their cost) with two medical experts. The findings of our discussion can be summarised as follows:

- Out of the 12 rules, 3 were accepted in their original version
- Another 3 rules were rejected. This was because symptoms were not deemed critical and in some cases, with unclear interpretation.
- The remaining 6 rules were accepted with modifications. In 2 cases, these modifications were **additional conditions** (e.g. additional symptoms that would make a situation truly critical). The additional conditions additionally involved new attributes in both cases and these were both related to the trend, i.e. will require to construct new features that take the historical development of symptom strengths into account. In the remaining 4 cases, we discovered an interesting new insight: initially, we built rules that predict unplanned patient visits to their physician or a hospital. However, the medical experts suggested that sometimes rules do predict a situation that requires action, but not necessarily a visit. Thus, a **different kind of alert** could be raised advising a patient or the home care institution to e.g. increase the dose of pain killers – instead of seeing a doctor. It could also advise patients to further monitor certain symptoms and only contact the doctor when they get worse. This is an adjustment to the system that would not be possible when working with black-box machine learning models.

To illustrate these two types of modifications (see bold-printed terms above), let us look more closely at the first two rules in Figure 5:

- The first rule suggests that an unplanned visit will occur if the term "clip marker" appears in the diagnosis details of a patient and if her wellbeing drops below a level of 75 (which is still relatively high). This was explained by our medical experts by saying that breast cancer patients receive a neoadjavant chemotherapy before a surgery, during which the tumor shrinks (which is why its position is marked with a "clip marker" – i.e. that term correlates with a specific treatment). The chemotherapy impacts the wellbeing negatively. Since this is to be expected, the rule was judged as only partially useful. However, the experts remarked that a warning should be raised if two additional conditions apply,

namely when a) the trend of wellbeing is negative over several days and when b) nausea or fatigue appear as accompanying symptoms. This serves as an example of human-recommended **additional conditions** in ML-discovered rules.

- The second rule recommends to raise a warning when the drug "Endoxan" is taken by a patient, the term "lymphangiosis" appears in the diagnosis and the patient's wellbeing drops below 47. The experts identified this as a situation of palliative care. Again, a visit to the physician did not seem necessary. However, raising an alert may make sense indicating – e.g. to the home care service – to intensify measures to alleviate the suffering and ensure a higher wellbeing. This is an example of a newly discovered **type of alert**.

In summary, we can see that a) there is hope that a machine learning approach to the discovery of rules may provide value and that a corresponding model will be able to discover several critical situations, that b) the utility of a machine-learned rule set will be limited because increasing its coverage (recall) is possible, but comes at the cost of lower precision, i.e. more false alarms and that c) the analysis of rules by medical experts not only results in rule modifications and suggestions for feature engineering, but also – in our case – in specific types of actions entailed by predictions that allow for more fine-grained recommendations to be made by the Early-Warning System.

6. Conclusions

In this paper, we presented our work-in-progress toward the development of an Early-Warning System (EWS) that aims to detect critical situations for cancer patients, based on a diary of symptoms. For this, we developed a hybrid intelligent approach that takes the form of an interpretable rule-based ML model. On the one hand, rules are learned based on the correlation between symptoms and unplanned visits. On the other hand, the rules are checked and improved or discarded by medical experts. Results are promising: the discussions with medical experts have shown that ML-based rule discovery makes experts think of more specific contexts than when prompted for rules in a general way. Several of the discovered rules are not only capable of predicting unplanned visits in our data, but were also accepted by experts as being generally valid. Other rules had to be modified by adding further conditions. The most interesting discovery was that some rules were considered not to predict truly critical situations, but nevertheless useful to trigger the recommendation of certain actions other than a visit to the physician or hospital. This shows how a "knowledge exchange" between humans and machine can lead to a better overall understanding of how an EWS can be optimised.

Next steps comprehend advanced feature engineering in which we will e.g. consider the history of well-being, symptom development over time, and missing patient entries of previous days. It will also be interesting to investigate the effect of further knowledge engineering activities, such as grouping symptoms in a meaningful way and including the resulting symptom categories as new features.

References

[1] J. C. Ginestra, H. M. Giannini, W. D. Schweickert, L. Meadows, M. J. Lynch, K. Pavan, C. J. Chivers, M. Draugelis, P. J. Donnelly, B. D. Fuchs, et al., Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock, Critical care medicine 47 (2019) 1477.

[2] M. Chae, H.-W. Gil, N.-J. Cho, H. Lee, Machine learning-based cardiac arrest prediction for early warning system, Mathematics 10 (2022) 2049.

[3] L. Zheng, O. Wang, S. Hao, C. Ye, M. Liu, M. Xia, A. N. Sabo, L. Markovic, F. Stearns, L. Kanov, et al., Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records, Translational psychiatry 10 (2020) 1–10.

[4] M. K. Lodhi, R. Ansari, Y. Yao, G. M. Keenan, D. Wilkie, A. A. Khokhar, Predicting hospital re-admissions from nursing care data of hospitalized patients, in: Industrial Conference on Data Mining, Springer, 2017, pp. 181–193.

[5] L. M. Bull, M. Lunt, G. P. Martin, K. Hyrich, J. C. Sergeant, Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods, Diagnostic and prognostic research 4 (2020) 1–16.

[6] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[7] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (2021) e745–e750.

[8] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, H. Wallach, Manipulating and measuring model interpretability, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–52.

[9] J. Shi, S.-y. Zhang, L.-m. Qiu, Credit scoring by feature-weighted support vector machines, Journal of Zhejiang University SCIENCE C 14 (2013) 197–204.

[10] A. Mollaysa, A. Kalousis, E. Bruno, M. Diephuis, Learning to augment with feature side-information, in: Asian Conference on Machine Learning, PMLR, 2019, pp. 173–187.

[11] E. D. Gennatas, J. H. Friedman, L. H. Ungar, R. Pirracchio, E. Eaton, L. G. Reichmann, Y. Interian, J. M. Luna, C. B. Simone, A. Auerbach, et al., Expert-augmented machine learning, Proceedings of the National Academy of Sciences 117 (2020) 4571–4577.

[12] M. J. Flores, A. Nicholson, A. Brunskillc, K. Korbb, S. Mascarod, Incorporating expert knowledge when learning Bayesian network structure: Heart failure as a case study, Technical Report, Technical Report 2010/3, Bayesian Intelligence, 2010, http://dx. doi. org/10 . . . , 2010.

[13] C. Deng, X. Ji, C. Rainey, J. Zhang, W. Lu, Integrating machine learning with human knowledge, Iscience 23 (2020) 101656.

[14] T. Botsas, L. R. Mason, O. K. Matar, I. Pan, Rule-based evolutionary bayesian learning, arXiv preprint arXiv:2202.13778 (2022).

[15] A. Trojan, N. Leuthold, C. Thomssen, A. Rody, T. Winder, A. Jakob, C. Egger, U. Held, C. Jackisch, The effect of collaborative reviews of electronic patient-reported outcomes on the congruence of patient-and clinician-reported toxicity in cancer patients receiving

systemic therapy: Prospective, multicenter, observational clinical trial, Journal of Medical Internet Research 23 (2021) e29271.

[16] A. Trojan, B. Bättig, M. Mannhart, B. Seifert, M. N. Brauchbar, M. Egbring, et al., Effect of collaborative review of electronic patient-reported outcomes for shared reporting in breast cancer patients: descriptive comparative study, JMIR cancer 7 (2021) e26950.

[17] W. W. Cohen, Repeated incremental pruning to produce error reduction, in: Machine Learning Proceedings of the Twelfth International Conference ML95, 1995.

[18] P. Domingos, Metacost: A general method for making classifiers cost-sensitive, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 155–164.