Clinical-based Prediction of Side Effects in Colon Cancer Chemotherapy

Mouna Kacimi
Faculty of Computer Science
Free University of Bozen-Bolzano
I-39100 Bozen-Bolzano, Italy
Mouna.Kacimi@unibz.it

Ognjen Savković
Faculty of Computer Science
Free University of Bozen-Bolzano
I-39100 Bozen-Bolzano, Italy
Ognjen.Savkovic@unibz.it

Manfred Mitterer
Hospital Franz Tappeiner
I-39012, Meran-Merano, Italy
Manfred.Mitterer@asbmeran-o.it

***Abstract*—Chemotherapy is used to treat cancer by killing malignant cells or stopping them from multiplying. However, it can also harm healthy cells, which causes side effects. Some of the side effects of chemotherapy do not pose a serious threat to patients' health. But, some others can be very serious such as the rapid fall in white blood cells making the patient vulnerable to serious infections. Different approaches have been proposed in the literature to predict the probability of experiencing a certain side effect on a specified day of each cycle of the chemotherapy. In our work, we are interested in predicting the side effects a patient is more likely to experience in each cycle. To this end, we take a different approach where we propose a predictive model based on patient clinical information, such as concomitant diseases and medications. We have analyzed data from from 75 patients under FOLFOX chemotherapy for colon cancer. The results show that our model improves the prediction accuracy compared to previously proposed time-based approaches. The goal of this work is to help healthcare professionals in identifying possible side effects before starting a chemotherapy and taking the necessary actions to improve the quality of the treatment.**

I. Introduction

A. Motivation

Cancer is a class of diseases characterized by a growth of abnormal cells that divide uncontrollably and manage to move throughout the body destroying healthy tissue. According to the World Health Organization [20], cancer is becoming a leading cause of death worldwide accounting for 7.6 million deaths (around 13% of all deaths) in 2008. These numbers are projected to continue rising, with an estimated 13.1 million deaths worldwide in 2030. Hence, research targeting cancer prevention and treatment is of a vital importance. For this reason, many efforts were invested over the last decades to improve the quality of medical treatments leading to an increase of the event free and overall survival rate. More specifically, adjuvant chemotherapy, employed after tumor removal surgery, was proven to help reducing the recurrence risk of some types of cancer [18], [19]. However, chemotherapy can damage healthy cells along with cancer cells. Patients receiving chemotherapy may experience significant side effects, especially if polychemotherapy is used, that could have a negative impact on their quality of life and daily living [17], [1]. Moreover, side effects can influence the overall survival [5] leading in some cases to mortality [10]. To address this problem, a comprehensive study of the causes of side effects is necessary to provide an effective prediction of their occurrence. In this way, health care professionals can take the necessary actions to alleviate the discomfort and anxiety of patients. Moreover, they can tailor the treatment depending on the patient's vulnerability to side effects for having a better control on his reactions to the chemotherapy.

B. Related Work

The problem of chemotherapy side effects have called for the use of technology as a mean of communication between professionals and patients. The goal is to improve the symptom control and achieve a better quality of life by reducing the hospitalizations rate and cost [12], [11], [7]. A practical example of the use of technology in cancer care is the Advanced Symptom Management System [8], [14], [6], [15]. It consists of a mobile telephone-based remote symptom monitoring system which can be used to register, monitor and predict the side-effects of chemotherapy while the patient is not with a healthcare professional [4]. In this kind of systems, patients are asked to complete a symptom questionnaire on a mobile phone on daily basis and sent this information directly to their hospital-based healthcare professional. Self-care advice is then given on the basis of the reported symptoms. All information collected from patients during their chemotherapy is then exploited to develop predictive models for side effects. Examples of this models include the work by Maguire et. al., [13] that aims at estimating the probability of experiencing each symptom on a specified day of treatment, for patients with breast cancer. This study shows that the probability of experiencing a specific symptom is time dependent. Its tendency is to have a peak effect, around the day in which the treatment is received by patients, and an inverted U-shape effect rising from a low on the day after treatment to a peak around mid-cycle before falling again. This work has been improved and generalized to other types of cancer by Mazzocco et. al., [16].

Very few approaches have looked at patient characteristics and clinical information to predict side effects. The work by Dranitsaris et. al., [3] proposes a regression model based on patient information to estimate cardiotoxic risk for patients with metastatic breast cancer. Their study has shown that characteristics such as patient age and weight and the number of cumulative cycles are good predictors for having cardiac toxicity. Similarly, Kuderer et. al., [10] show that patient characteristics, type of malignancy, comorbidities, and infectious complications are useful in identifying patients at increased risk of serious medical complications and mortality associated

978-1-4673-5801-9/13/$26.00 ©2013 IEEE

with febrile neutropenia. Authors in [9] investigate different approaches to retrieve top k patients with a chemotherapy history that is most similar to a given patient, which is a first step towards automatizing side effects prediction.

C. Goals

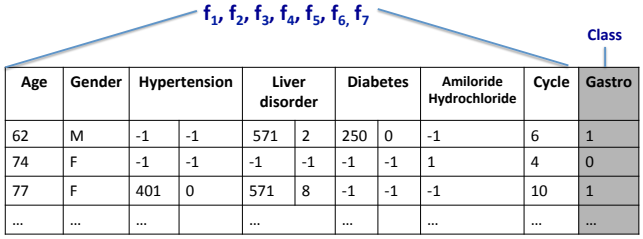
The goal of this study is to propose a comprehensive model for side effect prediction based on patient information. This includes personal data, such as age and gender, and clinical data such as concomitant diseases and prescribed medications. The model can be used to predict the side effects a patient is most likely to experience in each cycle of the chemotherapy. While time-based approaches presented previously achieve a good prediction accuracy, they require a permanent monitoring of patients and the availability of data on a day-by-day basis which is not always possible. In the worst case, it can lead to non-objective conclusions if the input provided by patients is not accurate. Moreover, they are conceived to predict the probability of a side effect on a given day based on the probability of having it on the day before. These restrictions make the adaptation of these approaches not suitable for our purpose. The new model is independent from monitoring systems and allow the prediction of side effects on cycle-basis. Additionally, it automatically identifies for each side effect the most important risk factors that could trigger its occurrence. Thus, the model is flexible and can be adapted to any type of chemotherapy. The ultimate goal is to use it as a tool for providing before-hand information to patients and health care professionals what can possibly happen during the chemotherapy. This prior knowledge would (1) help patients to be psychologically prepared for possible discomfort which reduces their anxiety and (2) help health care professionals to take the necessary actions and possibly provide tailored medications based on patient individual needs.

II. Side Effects Prediction

In our work, we are interested in predicting the side effects a patient would possibly have in each cycle of the chemotherapy. To this end, we introduce a personalized approach, by taking into account patient personal and clinical information, in contrast to time-based approaches proposed in the literature. We use two different predictive models based on binary classification, where class 0 means the absence of a given side affect and class 1 means its presence. The first model performs prediction by exploiting the similarity between patients while the second model maps observations about patient features to conclusions about related side effects. In the following, we describe the training data used to build out the two models and how the prediction is performed in each of them.

A. Training Data

Each patient is described by a set of features including personal and clinical information. In our datasets, personal information consists of the *age* of the patient, *gender*, and *age* when the cancer was diagnosed, while clinical information consists of the list of concomitant deceases of the patient, such as *diabetes*, *bacterial pneumonia*, and *thyrotoxicosis*. In addition to concomitant disease, we also have the list of medications the patient is taking. Formally, each patient P is described by a set of features $\{f_1, f_2, \ldots, f_n\}$. Based on these features our model predicts possible side effects in a personalized way.



Age	Gender	Hypertension		Liver disorder		Diabetes		Amiloride Hydrochloride	Cycle	Gastro
62	M	-1	-1	571	2	250	0	-1	6	1
74	F	-1	-1	-1	-1	-1	-1	1	4	0
77	F	401	0	571	8	-1	-1	-1	10	1
...	...	...		...		...		...	...	...

Fig. 1. Example of Training Data for *Gastro Side Effect*

For each side effect S_i, we prepare a training data consisting of a set of patients $D_{train} = \{P_1, P_2, \ldots, P_n\}$ where each patient P_j is described by the set of features presented above and is assigned a class label. The class label is equal to 1, if the patient P_j had the side effect S_i and 0 otherwise. The training data is then used to build a classifier that, for each new patient, predicts whether the patient is going to have the side effect S_i or not. In Figure 1, we give a simplified example of the training data related to the side effect of the gastrointestinal system (*Gastro*) in the FOLFOX chemotherapy of colon cancer consisting of 12 cycles. The complete training data contains more information about other concomitant diseases and prescribed medications. Each row of the training data corresponds to one patient during one cycle. The class field gives information whether the patient had gastrointestinal side effects (*Gastro*) in that cycle or not. The diseases are described by the International Classification of Diseases (ICD-9-CM) coding system [2]. Each disease code has a prefix and suffix. As shown in Figure 1, non alcoholic liver disorder has the code 571.8 and the alcoholic liver disorder has the code 571.2. We note that the two diseases fall into the same family of liver disorder, thus they share the same prefix. This coding system helps finding similarities between diseases which is very helpful for the prediction task. The medications features are binary features where the value 1 means that the patient is taking a given medication, such as Amiloride Hydrochloride in Figure 1. If the patient does not have a given disease or is not taking a given medications, their corresponding values are set to -1 as shown in Figure 1.

B. Similarity-based Prediction

Our first model falls into the category of lazy classifiers where the effort of prediction is done at the moment of the arrival of a new patient. It is known in the literature as the *K-Nearest Neighbor Classifier (KNN)*. The training phase in this case consists only of storing the feature vectors of patients and the class labels of the training data. When a new patient arrives, we extract his personal and clinical information as described above. Once we have the feature vector, we can perform the prediction of a given side effect S_i in any cycle C_l of the chemotherapy. The query would consist of two main components: (1) a query feature vector formed using the patient information and the cycle number C_l and (2) the name of side effect S_i. The similarity-based model finds the k most similar patients to the new patient from the training

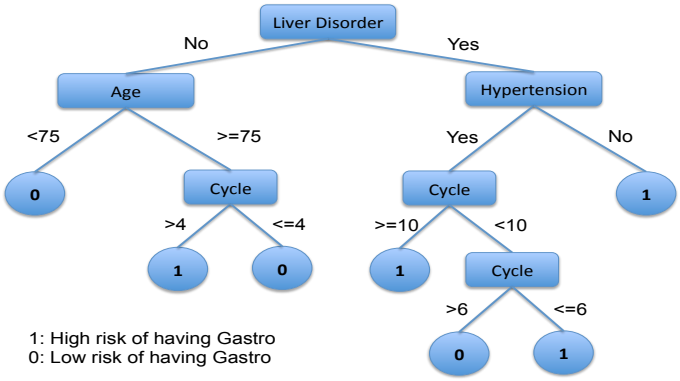


Fig. 2. Example of a Decision Tree *Gastro Side Effect*

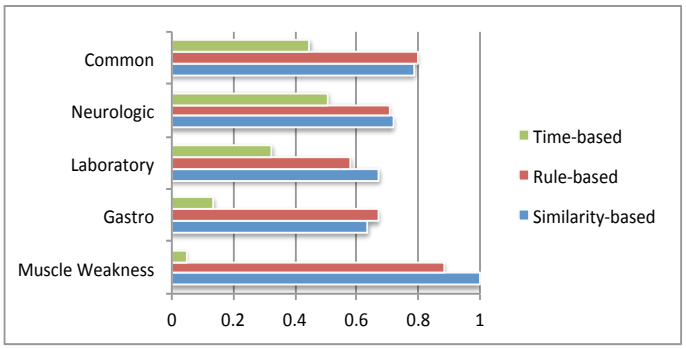


Fig. 3. Comparison of Performances for Each Side Effect (ROC Area)

data, where k is a user specific parameter. Each of the similar patients has a binary class label that indicates if the patient had the side effect S_i or not. The most frequent label among the k patients is predicted for the new patient.

C. Rule-based Prediction

Our second model falls into the category of eager classifiers, more specifically it consists of a *Decision Tree Classifier* which is commonly used in data mining. The goal of this model is to make the prediction of side effects based on several patient features. A simplified example of a decision tree related to gastrointestinal side effect (*Gastro*) is shown in Figure 2. Each interior node corresponds to one of the patient features; there are edges to children for each of the possible values of that feature. Each leaf represents a value of the class given the values of the patient features represented by the path from the root to the leaf. As shown in Figure 2, if a patient does not has a liver disorder and his age is less than 75 years old then he has a low risk to have gastrointestinal symptoms as a side effect. By contrast, if the patient has liver disorder and hypertension then he has a high risk of having gastrointestinal side effects in the last three cycles of the chemeotherapy (i.e., 10, 11, and 12). In this model, the training phase consists of splitting the training data into subsets based on the feature values. This process is repeated recursively until each subset has the same class label or when splitting no longer adds value to the predictions.

III. Experiments

A. Setup

In our experiments, we have used data from FOLFOX chemotherapy for colon cancer with 75 patients. As test cases, we have chosen the 3 most common types of side effects, namely: (1) *Common Effects* that include *nausea*, *vomiting*, *tiredness*, and *anemia*, (2) *Neurologic Effects* that include *paresthesia*, *dysgeusia*, and the *sensation of limbs disturbance*, and (3) *Laboratory Effects* related to lab results such as *hypercalcemia*, *hyperpotassemia*, and *neutropenia*. Additionally, we have used 2 other side effects that are more specific including (4) *Gastro* and (5) *Muscles Weakness*. To assess the accuracy of our predictive models, we have performed a *10-fold stratified cross-validation* where the original sample of patients is partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is used for testing and the remaining 9 subsamples are used for training. The cross-validation process is then repeated 10 times. The results from the different folds are finally averaged to produce a single estimation.

B. Approaches Under Comparison

We have tested both the similarity-based and the rule-based predictive models. As a baseline, we have implemented a time-based model following the same principle proposed in the pilot model [13]. However, due to the lack of daily information in our data, we have investigated the applicability of the time-based model on cycles information. In other words, we investigated whether the occurrence of a side effect in a given cycle n depends on its occurrence in previous cycles. However, we could not find any meaningful tendencies over time to learn a probability function as proposed in [13], [16]. Thus, we have used a discrete prediction model for each side effect S_i on cycle C_l. To build the model, we used a set of binary features $g_{C_1}, g_{C_2}, \ldots, g_{C_{l-1}}$ where g_{C_i} equals 1 if a patient has the side effect S_i in cycle C_l and 0 otherwise. We have used both decisions trees and KNN classifiers and the two models have given very similar results, so we have chosen decision trees to perform the time-based prediction.

C. Performance Metrics

To assess the performance of the different approaches, we have chosen the three precision measures commonly used in binary classification: *Sensitivity*, *Specificity*, and overall *Precision*. Sensitivity, also called the *true positive rate*, measures the proportion of actual positives which are correctly identified as such. In our study, it represents the percentage of patients who have a given side effect and are correctly identified as having it. By contrast, specificity, also called the *true negative rate*, measures the proportion of negatives which are correctly identified. In our study, it represents the percentage of patients who do not have a given side effect and are correctly identified as not having it. The overall precision is a weighted average of the sensitivity and the specificity values. The weights are computed based on the proportion of patients in each class.

We have also used the ROC curve to compare the performances of the different models. The idea behind the ROC curve is that classifiers use threshold values to make a decision about the class label. So, for each possible value of the decision threshold, a pair of true-positive and false-positive performance

TABLE I. MODELS PERFORMANCES USING FOLFOX CHEMOTHERAPY

	Common		Neurologic		Laboratory		Gastro		Muscle Weakness	
	Precision	ROC Area	Precision	ROC Area	Precision	ROC Area	Precision	ROC Area	Precision	ROC Area
Time-based	0.688	0.447	0.761	0.505	0.877	0.325	0.865	0.136	0.995	0.047
Similarity-based	0.857	0.787	0.76	0.718	0.848	0.673	0.930	0.636	0.995	0.997
Rule-based	0.865	0.796	0.760	0.705	0.839	0.580	0.928	0.670	0.997	0.883

rates are represented on the ROC curve. Even though the classifiers we have used are discrete (i.e., they are designed to output only a class label from each test instance), we can generate the curve and not just a single point by generating scores. For example, a decision tree determines a class label of a leaf node from the proportion of instances at the node; the class decision is simply the most prevalent class. These class proportions serve as a score. Once the curve is presented, we use the area under the curve as a performance metric which indicates the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

D. Results

Table I shows precision and *ROC* values of the three models described previously. We can see clearly that our proposed models outperform the time-based model. In terms of overall precision, we are doing particularly well for the *Common* and *Gastro* side effects where we achieve a precision of 86% with the rule-based model and a precision of 93% with the similarity-based model, respectively. Even if there is no big improvement in the precision for the other side effects, the ROC values show that we are outperforming the time-based model by reducing the number of false negatives and false positives. Particularly, the ROC value is high for the *Common*, *Neurologic*, and *Muscles Weakness* side effects with respective values of 79%, 71%, and 99%, as shown in Figure 3.

From the results shown in Table I, we can see that the similarity-based model and the rule-based model have very similar performances. However, if we take a closer look at the specificity and sensitivity values for *Common* and *Muscles Weakness* side effects, we can see that rule-based model performs better than the similarity-based model. The improvement significantly varies depending on the type of the side effect. For example, we can see in Figure 4 that the rule-based model improves the prediction of the presence of Muscles Weakness side effect by more than 30% compared to the similarity-based model, while for the common side effects there is an improvement of 1%.

IV. DISCUSSION

The main purpose of this study was to propose a side effect prediction model that takes into account patient information including the patient personal and clinical data. The proposed model can be applied to any type of cancer chemotherapy. Our experimental results show that both similarity-based and rule-based models outperform the time-based model by a significant margin. Sometimes, however, the gains are small and generally depend on the type of the side effect. In the following, we discuss some of the specific strengths of our approach and we also point out the limitations.

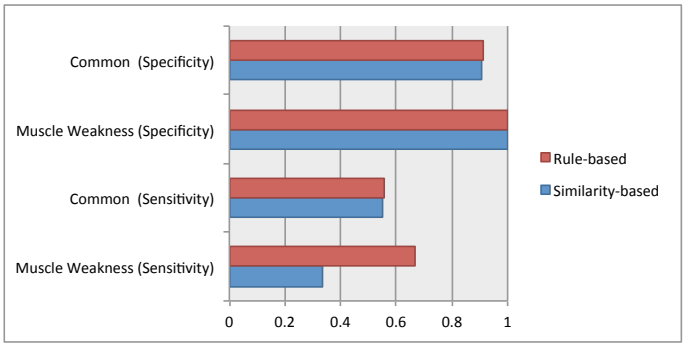


Fig. 4. Comparison of Rule-based and Similarity-based Models (Sensitivity and Specificity

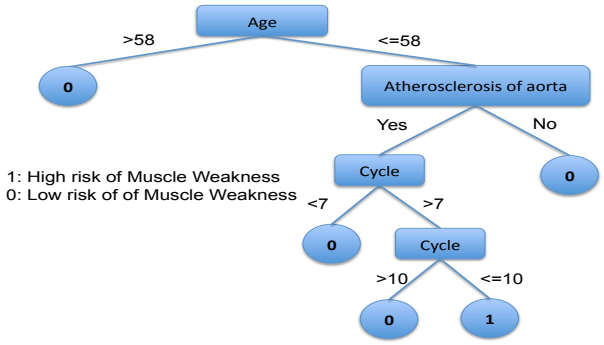


Fig. 5. Decision Tree for Muscles Weakness Side Effect

A. Specific Strengths

The main advantage of our approach is threefold. The first aspect is that we use patient history for side effect prediction so we allow our models together with health care professionals to learn from past experiences. This learning process helps to discover correlations between patient characteristics and side effects. Hence, the chemotherapy can be tailored for specific needs to improve the patient experience. The second aspect is that we do not require additional infrastructures to collect data such as the mobile remote monitoring systems which are not always available for or used by patients. This makes our model feasible to any hospital that does not provide the remote monitoring service. Moreover, it can easily be implemented over any standard database system. The third aspect regards the simplicity of the approach and its ease to understand by health care professionals. In particular, the rule-based model allows a very concise description of the important risk factors for each side effect. The visual representation of the model is very easy to understand and the interpretation of the results is straightforward. We take as an example the decision tree of the *Muscles Weakness* side effect shown in Figure 5. Even though patients are described by tens of features, it is straightforward to see that, according to our data, what matters for *Muscles*

Weakness are the *age* and *Atherosclerosis of aorta*. Basically, if the age of the patient is higher than 58 then the patient has little chance to get *Muscles Weakness*, otherwise if the patient has *Atherosclerosis of aorta* the patient has high chances of getting *Muscles Weakness* between cycles 7 and 10. The decision tree can thus be used for predicting side effects as well as a concise summary for important risk factors of the side effect.

B. Limitations

The main limitation of our model is the type of data we are dealing with. Medical data has a very similar structure to transactional databases where each patient is described by a variable number of features. Some patients have many diseases, and some others have few. Some patients have many medicines and some others do not take any. Given that predictive models require that all patients are described by a fixed number of features, we have build our training data based on all the exiting features in the dataset. This resulted in a sparse data. For example, if there are 100 diseases, and each patient has a maximum of three diseases, then each instance would have at least 97 features with the value -1 which represents the null value. The sparsity of the data has an impact on the effectiveness of the classifiers which explains why, in some cases, our models have a slight decease in terms of performance. Moreover, the high number of features opens up the problem of the curse of dimensionality. While decision trees deal well with high dimensions, similarity-based approaches are less tolerant. Two main directions should be explored to solve this problem. The first one is by reducing the dimensionality of the space using features selection techniques, while the second one is to combine frequent pattern techniques with classification to get better results.

V. Conclusions

We have presented a clinical-based approach for side effects prediction in cancer chemotherapies. Our model automatically selects the most relevant risk factors for each side effect which makes our approach to be applicable to any type of chemotherapy. We have proposed a rule-based model and a similarity-based model. The rule base model has shown to be more effective in predicting specific side effects. Moreover, it can also be used to provide a concise representation about the important factors of a given side effect. We have run experiments using data from the FOLFOX chemotherapy of colon cancer which have shown that our proposed models improve significantly the prediction effectiveness compared to time-based approaches. This work can be used as a tool to help healthcare professionals to understand better the patient experience and risks, so they can provide more tailored treatments to the patient needs and improve the patient quality of life.

Acknowledgment

This work was supported by the "RARE" project, funded by the province of Bolzano.

References

[1] H. Anderson and M. K. Palmer. Measuring quality of life: impact of chemotherapy for advanced colorectal cancer. experience from two recent large phase iii trials. *British Joumal of Cancer*, 77:9–14, 1998.

[2] CD-9-CM. http://icd9cm.chrisendres.com/.

[3] G. Dranitsaris, D. Rayson, M. Vincent, J. Chang, K. Gelmon, D. Sandor, and G. Reardon. The development of a predictive model to estimate cardiotoxic risk for patients with metastatic breast cancer receiving anthracyclines,. *Breast Cancer Research and Treatment*, 107:443450, 2008.

[4] L. Forbat, R. Maguire, L. McCann, N. Illingworth, and N. Kearney. The use of technology in cancer care: applying foucaults ideas to explore the changing dynamics of power in health care. *Journal of Advanced Nursing*, 65:306315, 2009.

[5] M. Gianni Bonadonna, B. Pinuccia Valagussa, M. Angela Moliterni, M. Milvia Zambetti, and M. Cristina Brambilla. Adjuvant cyclophosphamide, methotrexate, and fluorouracil in node-positive breast cancer: The results of 20 years of follow-up. *The New England Journal of Medicine*, 332:901906, 1995.

[6] N. Kearney, L. Kidd, M. Miller, M. Sage, J. Khorrami, and M. McGee. Utilising handheld computers to monitor and support patients receiving chemotherapy: results of a uk- based feasibility study. *Supportive Care in Cancer*, 14:742 752, 2006.

[7] N. Kearney, L. McCann, J. Norrie, L. Taylor, P. Gray, M.-L. M., M. Sage, M. Miller, and R. Maguire. Evaluation of a mobile phone-based, advanced symptom management system (asyms) in the management of chemotherapy-related toxicity. *Supportive Care in Cancer*, 17:437444, 2009.

[8] N. Kearney, . Muir, L, M. Miller, I. Hargan, and P. Gray. Using handheld computers to support patients receiving outpatient chemotherapy. *European Journal of Cancer Supplements*, 1(5):S368, 2003.

[9] M. Khayati, J. M. Anderson, M. H. Böhlen, J. Gamper, and M. Mitterer. Similarity of chemotherapy histories based on imputed values. *IJMEI*, 4(3):282–298, 2012.

[10] N. Kuderer, D. Dale, J. Crawford, L. Cosler, and G. Lyman. Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. *Cancer*, 106:22582266, 2006.

[11] M. E. Larsen, J. Rowntree, A. Young, S. Pearson, J. Smith, O. J. Gibson, A. Weaver, and L. Tarassenko. Chemotherapy side-effect management using mobile phones. *Conf Proc IEEE Eng Med Biol Soc*, 2008:5152–5, 2008.

[12] A. Louis, T. Turner, M. Gretton, A. Baksh, and J. Cleland. A systematic review of telemonitoring for the management of heart failure. *European Journal of Heart Failure*, 5:583590, 2003.

[13] R. Maguire, J. Cowie, C. Leadbetter, K. McCall, K. Swingler, L. McCann, and N. Kearney. The development of a side effect risk assessment tool (asyms-serat) for use in patients with breast cancer undergoing adjuvant chemotherapy. *Journal of Research in Nursing*, 14:2740, 2009.

[14] R. Maguire, L. McCann, M. Miller, and N. Kearney. Nurse's perceptions and experiences of using of a mobile-phone-based advanced symptom management system (asyms) to monitor and manage chemotherapy-related toxicity. *European Journal of Oncology Nursing*, 12:380386, 2008.

[15] R. Maguire, M. Miller, L. McCann, L. Taylor, N. Kearney, M. Sage, and J. Norrie. Application of mobile phone technology for managing chemotherapy-associated side-effects. *Nursing Times*, 104(22):28–29, 2008.

[16] T. Mazzocco and A. Hussain. A side-effects mapping model in patients with lung, colorectal and breast cancer receiving chemotherapy. *13th IEEE International Conference on e-Health Networking Applications and Services (Healthcom)*, pages 34–39, 2011.

[17] A. Molassiotis, C. Stricker, B. Eaby, L. Velders, and P. Coventry. Understanding the concept of chemotherapy-related nausea: the patient experience. *European Journal of Cancer Care*, 17:444453, 2008.

[18] R. Packer, L. Sutton, J. Goldwein, G. Perilongo, G. Bunin, J. Ryan, B. Cohen, G. D'Angio, E. Kramer, and R. Zimmerman. Improved survival with the use of adjuvant chemotherapy in the treatment of medulloblastoma. *J Neurosurg*, 74(3):433–40, 1991.

[19] O. Reporter. Chemotherapy after breast cancer recurrance increases survival rates.http://www.observer-reporter.com/, 2013.

[20] WHO. World health organization: http://www.who.int/.