Iran J Public Health, Vol. 50, No.3, Mar 2021, pp.598-605

Original Article

Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation

Hadi LOTFNEZHAD AFSHAR [1], Nasrollah JABBARI [2], *Hamid Reza KHALKHALI [3], Omid ESNAASHARI [4]

1. *Department of Health Information Technology, School of Paramedical, Urmia University of Medical Sciences, Urmia, Iran*
2. *Department of Medical Physics, Solid Tumor Research Center, School of Paramedical, Urmia University of Medical Sciences, Urmia, Iran*
3. *Department of Biostatistics and Epidemiology, Patient Safety Research Center, School of Medicine, Urmia University of Medical Sciences, Urmia, Iran*
4. *Omid Treatment and Research Center, Urmia, Iran*

***Corresponding Author:** Email: hamid.reza.kh2420@gmail.com

(Received 07 May 2020; accepted 27 Jul 2020)

Abstract

Background: The low breast cancer survival rates in less developed countries are critical. The machine learning techniques predict cancers survival with high accuracy. Missing data are the most important limitation for using the highest potential of these techniques to predict cancers survival. Multiple imputation (MI) was implemented and analyzed in detail to impute the missing data of a breast cancer dataset.

Methods: The dataset was from The Omid Treatment and Research Center Urmia, Iran between Jan 2006 and Dec 2012 and had information from 856 women. The algorithms such as C5 and repeated incremental pruning to produce error reduction were applied on the imputed versions of the original dataset and the non-imputed dataset to predict and extract clinical rules, respectively.

Results: The findings showed the performance of C5 in all the evaluation criteria including accuracy (84.42%), sensitivity (92.21%), specificity (64%), Kappa statistic (59.06%), and the area under the receiver operator characteristic (ROC) curve (0.84), was improved after imputation.

Conclusion: The dataset of the present study met the requirements for using the multiple imputation method. The extracted rules after the application of MI were more comprehensive and contained knowledge that is more clinical. However, the clinical value of the extracted rules after filling in the missing data did not noticeably increase.

Keywords: Breast neoplasms; Survival; Observer variation; Imputation; Machine learning

Introduction

Breast cancer is one of the most common types of cancer in women worldwide. It is the leading type of cancer in Iranian females, accounting for 24.6% of all cancers (1). Breast cancer is also a typical cause of death in the world (2). The mortality rate of this disease was 4.33 per 100,000 in 2010 that had been dramatically increased from 0.96 per 100,000 in 1995 (3). Mean of 5-yr sur-

vival rate of breast cancer in Iran was 69.5% between 2004 and 2014 (4). However, this rate in United States was 90.6% in 2013 (5).

The prediction of a disease outcome (death or survival) is discussed in the medical prognosis field (6). The establishment of treatment plan, intensity and type of drugs are dependent to prognosis (7, 8). Survival analysis as a subset of medical prognosis applies methods on patients' previous medical data to predict the survival of them (7). In the recent years, frequency of machine learning methods has been increased in the cancer survival analysis. The high accuracy of these methods over traditional statistical methods in predicting cancer survival and discovering hidden patterns among cancer datasets seem to be main reasons (6, 9, 10).

Machine learning methods generally have been designed to mine large datasets (11). However, large datasets in health domain were not always accessible. For example, a cancer related dataset needs at least a five-year period that be in a desired condition to apply machine learning algorithms on it for predicting breast cancer survival (6, 10). SEER (Surveillance, Epidemiology, and End Results Program) cancer dataset containing most information about different types of cancers is the largest dataset used by machine learning researchers. Availability of SEER dataset freely from the internet and good volume of its data were main reasons to use this dataset. However, SEER dataset reflects the properties related to cancer patients of USA (United States of America).

The lack of a national cancer registry has caused that conducting the countrywide machine learning researches are impossible for predicting cancers survival in Iran. Another option in these situations was using datasets containing province-wide information. The Omid Treatment and Research Center (OTRC) has been located in Urmia (the largest city in West Azerbaijan province of Iran). The patients of different cancers were referred to this center. Cancerous patients' information was documented on paper records in OTRC.

The paper health records usually were not documented with enough details by healthcare providers in the Iran's hospitals (12, 13). Poor documentation is one of the most important factors causing missing data in a dataset (11). The datasets created in such conditions do not reflect full potential of saved data for achieving useful knowledge. The imputation of missing data is a proposed solution to overcome the challenges that are created by the poor documentation of the health records (8, 14-17). Methods such as single or multiple imputations of missing data are common ways to filling them (18, 19). Few researchers, to the best of our knowledge, have used methods to handle missing data in the breast cancer prognosis studies done in machine learning domain.

We aimed to develop the models based on C5 algorithm predicting breast cancer survival from a dataset with imputed and removed missing data and to compare performance of developed models. Another goal of this study was to extract rules from the mentioned dataset by Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm and to analyze similarities and differences between them.

Materials and Methods

Data Source and Dataset Characteristics

The data were extracted from paper records of 856 female breast cancer patients (mean age 47.7 yr, standard deviation 9.7) between Jan 2006 and Dec 2012 from OTRC in the city of Urmia, Iran. The Urmia University of Medical Science Research Ethics Board approved the study design (Approval code: IR.UMSU.REC.1392.154).These data included information about the breast cancer patients treated in a cancer charity organization. To prevent the sampling bias, the data that had been matched with mentioned criteria in the Khalkhali's study (17), involved in the dataset. The criteria for collecting breast cancer variables to predict survival were based on consulting with organization oncologists and studying domain literature (6). The survival status of patients

(Alive/Dead) was available in their records and in the case of missing; it was received through phone call to patient's residents. A dataset with 15 variables (four continuous and 11 categorical) and one outcome variable was created by the data of paper medical records. Then dataset was entered into Excel worksheet.

Data Preparation

The dataset was imported from Excel into R software to detect and handling outliers and missing data. The box plots were used to detect outliers and did not show any outlier in dataset. The overall percentage of missingness in the data set was 5.8 that after deleting records that more than 50 percent of their fields had missing data decreased to 3.3 percent. Removing of those records reduced the dataset from 856 to 819 records. Missing data statistics has been showed in the Table 1.

After determining variables with missing data and some basic statistics about them, the pattern of missing data was detected by the mice (20) package.

Table 1: Numbers and percentages of missing data in dataset variables

Variable Name	*Number*	*Percent*
Age	0	0
Primary Site	15	1.8
Histology	9	1.1
Tumor Size	9	1.1
Metastases	12	1.5
Stage	13	1.6
Behavior	12	1.5
Grade	19	2.3
Positive regional node	16	1.9
Removed regional node	22	2.7
Surgery	3	0.4
Radiation	0	0
Her2	173	21.1
ER	66	8.1
PR	66	8.1
Survival	0	0

The pattern showed that "Her2" variable had the highest missing values in the records of the dataset. The results of missing data pattern revealed that 628 records of dataset were without missing data and the total number of missing data on each variable was 435. Next step was determining variables that were missed together and relationships between a variable's "missingness" and the values of the other variables. The variables of ER and PR tended to missing together completely (r=1). Some of variables such as Stage/Behavior (r=0.96) and Removed regional node/Positive regional node (r=0.85) tended to missing together highly. For example, because ER and PR tests always are ordered together by oncologists (21), so their missingness is also together.

Missing data mechanism was identified by the relationships between missing data in a variable and the observed data on other variables. Because the correlations of these variables were not particularly large, mechanism of data deviates minimally from missing completely at random (MCAR) and may be missing at random (MAR). However, the possibility that the data are not missing at random (NMAR) should not be ruled out.

If missing data had been deleted, 23.3 percent of dataset records would have been removed. Deletion of this amount of records could lead to a significant information loss. To keep records with missed fields, missing data were imputed by multiple imputation (MI) method. Approach of this method for filling missing data was based on repeated simulations that are done by Monte Carlo methods. In MI, a set of complete datasets, typically 3 to 10 that in this study was 5, is gener-

ated from the original dataset that has missing data (18). For doing MI method, mechanism of missing data should be MCAR or MAR (18) that in the current study, this condition was met. The mice (20) package was used for doing MI. Missing data imputation of continuous and categorical variables is predictive mean matching and polytomous logistic regression, respectively.

Model development

The algorithms of C5.0 and RIPPER were applied to classify survival status of datasets. The algorithms are belonged to decision tree and rule learner families, respectively. C5.0 is one of the most well-known algorithms of decision trees that as an all-purpose classifier, does well for most types of problems (6). The C50 package was used to run algorithm. RIPPER is an algorithm to generate rules that are easy-to-understand and human-readable and generally produces a simpler model than a comparable decision tree (23). RIPPER was applied by RWeka package (24). Both algorithms were run on 80% of data as training data and 20% of them were kept as test data to evaluate algorithm's performance.

Model evaluation

The performance of C5 algorithm was evaluated by criteria, such as accuracy, sensitivity, specificity, kappa statistic and the area under the ROC curve (AUC). They are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{Specificity} = \frac{TN}{FP + TN} = \frac{\text{true negative}}{\text{false positive} + \text{true negative}}$$

$$k = \frac{P_r(a) - P_r(e)}{1 - P_r(e)} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

$$AUC = \frac{\text{average sensitivity}}{\text{all possible specificities}}$$

The clinical accuracy of extracted rules from RIPPER algorithm was evaluated by the center oncologists.

Results

Overall, the results presented below showed that after imputation of missing data, the performance of C5.0 was increased in all evaluation criteria. The Stage and PR were the first appearing variables in the most rules extracted by RIPPER in all six datasets. Table 2 summarizes the performance of C5.0 in all datasets.

The averages of evaluation criteria for imputed datasets (one to five) are 84.42%, 92.21%, 64%, 59.06% and 0.8369 for accuracy, sensitivity, specificity, Kappa statistic and AUC, respectively.

The RIPPER extracted nine rules from all datasets, except the forth imputed dataset that had 18 rules. The first three rules from each dataset have been showed in Table 3.

Table 2: The performance of C5.0

Evaluation criteria	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*Kappa Statistic (%)*	*AUC*
LD* dataset	79.03	88.17	51.61	41.57	0.7910
MI** dataset 1	84.66	90.68	68.89	60.82	0.8667
MI dataset 2	81.6	88.98	62.22	52.65	0.8142
MI dataset 3	85.28	91.53	68.89	62.12	0.8456
MI dataset 4	85.28	94.92	60	59.85	0.8093
MI dataset 5	85.28	94.92	60	59.58	0.8491

* Listwise Deletion: The dataset that missing data of it have been removed. **Multiple Imputations: The dataset that missing data of it have been imputed

Table 3: The most important rules extracted from datasets

Datasets	*Rule*	*Right classified records*	*Wrong classified records*
LD dataset	1. If cancer is *Stage* IV and *Progesterone Receptor* testing is not ordered by oncologist and *Age* of patient is less than or equal to 42, then the patient is *not* survived	19	0
	2. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 4.5, then the patient is *not* survived	48	6
	3. If *Positive regional lymph nodes* is greater than or equal to 6 and cancer is *Stage* IV, then the patient is *not* survived	21	6
MI dataset 1	1. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 4 and *Her2* testing is ordered by oncologist, then the patient is *not* survived	52	2
	2. If cancer is *Stage* IV and *Positive regional lymph nodes* is greater than or equal to 5, then the patient is *not* survived	61	15
	3. If *Age* of patient is between 54 and 67 and *Her2* testing is ordered by oncologist and *Positive regional lymph nodes* is greater than or equal to 4 and *Behavior* of cancer is malignant, then the patient is *not* survived	19	2
MI dataset 2	1. If cancer is *Stage* IV and *Progesterone Receptor* testing is not ordered by oncologist, then the patient is *not* survived	99	20
	2. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 5 and *Her2* testing is ordered by oncologist, then the patient is *not* survived	28	1
	3. If *Positive regional lymph nodes* is greater than or equal to 9 and *Age* of patient is greater than or equal to 48 and *Removed regional lymph nodes* is between 17 and 24, then patient is *not* survived	10	0
MI dataset 3	1. If cancer is *Stage* IV and *Progesterone Receptor* testing is not ordered by oncologist, then patient is *not* survived	97	18
	2. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 4 and *Her2* testing is ordered by oncologist, then patient is *not* survived	29	1
	3. If *Age* is greater than or equal to 54 and *Her2* testing is ordered by oncologist and *Positive regional lymph nodes* is greater than or equal to 4, then patient is *not* survived	22	6
MI dataset 4	1. If cancer is *Stage* IV and *Progesterone Receptor* testing is not ordered by oncologist and *Her2* testing is ordered by oncologist, then patient is *not* survived	48	3
	2. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 5, then patient is *not* survived	60	12
	3. If *Positive regional lymph nodes* is greater than or equal to 9 and *Tumor size* is less than or equal to 4 and *Age* is greater than or equal to 47, then patient is *not* survived	20	4
MI dataset 5	1. If cancer is *Stage* IV and *Surgery* is not recommended, then patient is *not* survived	42	3
	2. If cancer is *Stage* IV and *Tumor size* is greater than or equal to 5 and *Her2* testing is ordered by oncologist, then patient is *not* survived	22	2
	3. If cancer is *Stage* IV and *Positive regional lymph nodes* is greater than or equal to 5, then patient is *not* survived	56	13

Only nine variables from 15 variables have constituted first three rules. Stage was the most frequent variable (14 times) appeared in the first three rules. Stage, Tumor size and Her2 are variables that often appear (four times) together in the first three rules that have been generated from imputed datasets.

The extracted rules generally define that combination of some variables with each other lead to patient's death (Table 3).

Discussion

The main goals of this study generally were com-

pare breast cancer survival prediction models and extract rules developed from imputed and non-imputed datasets. The results showed that performance of developed model from imputed dataset was better than model established from non-imputed dataset. However, the extracted rules had not any significant differences clinically.

The present study suggested that MI method was better than LD method to increase the performance of C5 algorithm to predict breast cancer survival. The average scores of all evaluation criteria in MI datasets were higher than scores of LD dataset. The study of Jerez et al (16) about prediction of breast cancer relapse showed that MI as a subset of imputation statistical methods outperformed LD method. They applied MI method in three different packages that one of them was MICE. AUC as only reported criterion to measure algorithm's performance in their study was lower than our AUC (0.8369 vs. 0.7250). AUC of LD dataset in our study was also higher than their AUC (0.7910 vs. 0.7151). AUC measured the overall performance of a model and adjustments between sensitivity and specificity. The value of AUC nearer to 1, the overall performance was better. The highest AUC of their study was belonged to k-nearest neighbor (KNN) as an imputation method based on machine learning technique and was lower than our AUC (0.7910 vs. 0.7345).

Khalkhali et al (17) have conducted a research based on a different version of the dataset that has been analyzed in our study. They predicted breast cancer survival by classification and regression trees (CART) algorithm and imputed missing data by MI method in different software. The results showed that C5 applied on MI dataset imputed by mice package outperformed CART applied on MI dataset imputed by IBM SPSS Statistics 22 in accuracy (84.42 vs. 80.3) and specificity (64 vs. 53). Sensitivity of CART was higher (93.5 vs. 92.21) and kappa statistic and AUC had not been reported. Kappa statistic unlike accuracy that involves chancy correct predictions of the algorithm ignores them and adjusts accuracy (23).The specificity score of our study apparently seemed to be higher than their study's score. Lotfnezhad et al (8) applied MI method to SEER dataset and obtained 96.7, 97.7 and 95.6 for accuracy, sensitivity and specificity, respectively. These scores belonged to support vector machine algorithm were better than their counterparts were in the current study. Nevertheless, the subtle difference between sensitivity and specificity and lack of AUC score may overshadow this superiority. The mentioned discussed issue about paper of Lotfnezhad et al (8) seemed that be true about study of Ahmad et al (14) that have used EM method rather than MI. Pedro et al (15) used three methods to deal with missing data and predicted survival by four algorithms. MI was not among these methods. KNN algorithm applied to the dataset that had been handled with KNN missing data method, achieved AUC: 0.7845, accuracy: 81.73% and specificity: 70.46% and had the best performance. The highest sensitivity (88.38%) was belonged to SVM applied to EM method dataset. The scores of all evaluation criteria, except specificity were lower than our study. KNN missing data handling method outperforms other methods (15, 16). The mentioned evaluation criteria scores were not perfect representatives to compare results of papers. The factors such as: number of records and variables constituting the dataset, extent of correlation between variables, the type of class variable (binary or multi), the equilibrium of the dataset and the percentage and distribution of missing data are also important (25).

The extracted rules by RIPPER were clear clinically and presented previously known knowledge. Generally, all rules generated from LD and MI datasets predicted that combination some of predictors with each other lead to death of the patients. Important rules produced from MI datasets included four variables that were not in the LD dataset. Her2 was most frequent among them. It is assumed that the high percentage of missing data in Her2 was the main reason. Stage IV was the core variable in the most important rules (first and second) of LD and MI datasets. The five year breast cancer survival rate for stage IV is between 20 and 25 percent (26, 27). The rules of our study were consistent with these sta-

tistics, but based on opinions of the OTRC oncologists did not present any new knowledge clinically. The oncologists stated that because a patient with stage IV breast cancer does not have much chance to live, knowing other predictors constituting the rules do not help much to learn new knowledge. Although these comments seemed to be true, it should be considered that the rules are extracted from medical records (paper or electronic) documented by the clinicians. The more specific data documented, more specific knowledge extracted. The paper of Khalkali et al (17) also contained the rules extracted by CART algorithm. The most of their rules defined the conditions that predicted survival of patients. However, clinicians are more interested to know what conditions threatened survival of patients. A few rules (three) that had predicted death of patients were consistent with our rules.

Although this study has done on nearly small dataset and missing data have imputed by a statistical method, these did not affect our results in term of algorithm performance.

Conclusion

MI method was used to handle missing data of a real breast cancer dataset and the prerequisite conditions of this method were analyzed thoroughly to avoid probable biases. It appears that such preprocessing methods can increase the quality of predictive models. Using of MI method lead to extract rules that were created by more variables and consequently contained more knowledge. Despite this, they were not considered seriously by oncologists. If the dataset of current study had more records and genome variables then the extracted rules might contain new information.

Future work should benefit greatly by using a method from machine learning to handle missing data and a bigger dataset that has preferably genome variables.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study was supported and funded by Urmia University of Medical Sciences (Grant No. 92-01-52-1140). The authors acknowledge the Vice Chancellor of Research and Technology, Urmia University of Medical Sciences, which approved this project, and thank all the staffs and directors in Omid Treatment and Research Center.

Conflict of interest

The authors declare that there is no conflict of interests.

References

1. Jazayeri SB, Saadat S, Ramezani R, Kaviani A (2015). Incidence of primary breast cancer in Iran: Ten-year national cancer registry data report. *Cancer Epidemiol*, 39(4):519-27.
2. Ferlay J, Soerjomataram I, Dikshit R, et al (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer,* 136(5):E359-86.
3. Sharifian A, Pourhoseingholi MA, Emadedin M, et al (2015). Burden of Breast Cancer in Iranian Women is Increasing. *Asian Pac J Cancer Prev,* 16(12):5049-52.
4. Rahimzadeh M, Pourhoseingholi MA, Kavehie B (2016). Survival rates for breast cancer in iranian patients: a meta-analysis. *Asian Pac J Cancer Prev,* 17(10): 4615–4621.
5. Kate RJ, Nadig R (2017). Stage-specific predictive models for breast cancer survivability. *Int J Med Inform,* 97:304-311.
6. Delen D, Walker G, Kadam A (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med,*34(2):113-27.
7. Park K, Ali A, Kim D, et al (2013). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence,*26(9):2194-2205.

8. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, Sadoughi F (2015). Prediction of breast cancer survival through knowledge discovery in databases. *Glob J Health Sci,*7(4):392-8.
9. Jerez JM, Franco L, Alba E, et al (2005). Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Res Treat,*94(3):265-72.
10. Thongkam J, Xu GD, Zhang YC, Huang FC (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications,*36(10):12200-12209.
11. Han J, Kamber M, Pei J (2011). *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers Inc,USA, pp.: 100-115.
12. Dehghan M, Dehghan D, Sheikhrabori A, et al (2013). Quality improvement in clinical documentation: does clinical governance work? *J Multidiscip Healthc,*6:441-50.
13. Saravi BM, Asgari Z, Siamian H, et al (2016). Documentation of Medical Records in Hospitals of Mazandaran University of Medical Sciences in 2014: a Quantitative Study. *Acta Inform Med,*24(3):202-6.
14. Ahmad L, Eshlaghy A, Poorebrahimi A, et al (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform,*4(2):3.
15. Garcia-Laencina PJ, Abreu PH, Abreu MH, Afonoso N (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Comput Biol Med,*59:125-133.
16. Jerez JM, Molina I, García-Laencina PJ, et al (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med,* 50(2):105-15.
17. Khalkhali HR, Lotfnezhad Afshar H, Esnaashari O, Jabbari N (2016). Applying Data Mining Techniques to Extract Hidden Patterns about Breast Cancer Survival in an Iranian Cohort Study. *J Res Health Sci,* 16(1):31-5.
18. Sterne JAC, White IR, Carlin JB, et al (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ,*338:b2393.
19. Horton NJ, Kleinman KP (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat,*61(1):79-90.
20. Buuren. Sv, Groothuis-Oudshoorn. K (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,*45(3):1-67.
21. Bonadonna G, Gabriel N H, Pinuccia V (2006). *Textbook of Breast Cancer: A Clinical Guide to Therapy*. 3rd ed. Informa HealthCare, UK, pp.: 85-93.
22. Kuhn. M, Weston. S, Coulter. N, Culp. M. C50: C5.0 Decision Trees and Rule-Based Models (2015). R package version 0.1.0-24. https://CRAN.R-project.org/package=C50
23. Lantz B (2015). *Machine Learning with R*. 2nd ed. Packt, UK, pp.: 123-32.
24. Hornik. K, Buchta. C, Zeileis (2009). A. Open-Source Machine Learning: R Meets Weka. *Computational Statistics,*24(2):225–232.
25. Zhang Y, Xin Y, Li Q, et al (2017). Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *BioMedical Engineering OnLine,*16(1):125.
26. Hölzel D, Eckel R, Bauerfeind I, et al (2017). Survival of de novo stage IV breast cancer patients over three decades. *J Cancer Res Clin Oncol,*143(3):509-519.
27. Macià F, Porta M, Murta-Nascimento C, et al (2012). Factors affecting 5- and 10-year survival of women with breast cancer: An analysis based on a public general hospital in Barcelona. *Cancer Epidemiol,*36(6):554-9.