Journal of Biomedical Informatics 42 (2009) 1–10

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Probabilistic classifiers and automated cancer registration: An exploratory application

Sandro Tognazzo [a,*], Bovo Emanuela [a], Fiore Anna Rita [a], Guzzinati Stefano [a], Monetti Daniele [a], Stocco Cramen Fiorella [a], Zambon Paola [b]

[a] Venetian Tumour Registry, Registro Tumori del Veneto, Istituto Oncologico Veneto—IRCCS, 64, 35128 Padua, Italy
[b] Department of Oncology, University of Padua, Padua, Italy

ARTICLE INFO

Article history:
Received 1 August 2007
Available online 21 June 2008

Keywords:
Automated cancer registration
Cancer Registry
Probabilistic classifiers
Random forest

ABSTRACT

A test of the performance of two probabilistic classifiers (random forests and multinomial logit models) in automatically defining cancer cases has been carried out on 5608 subjects, registered by the Venetian Tumour Registry (RTV) during the years 1987–1996 and manually checked for possible second cancers that occurred during the 1997–1999 period.

An eightfold cross-validation was performed to estimate the classification error; 63 predictive variables were entered into the model fitting. The random forest allows to automatically classify 45% of subjects with a classification error lower than 5%, while the corresponding error is 31% for the multilogit model. The performance of the former classifier is appealing, indicating a potential drop of manually checked cases from 1750 to 960 per incidence year with a moderate error rate. This result suggests to refine the approach and extend it to other categories of manually treated cases.

1. Introduction

In most cancer registries, the evaluation of available diagnostic evidence and the decision about whether to register a case or not is carried out exclusively by registration technicians. Only in a relatively small number of registries, a substantial part of cases are accepted in an automatic fashion [1] and manual evaluation is restricted to those cases whose evidence is not concordant or is insufficient. The main reason for adopting automatic decision procedures is that they allow for a reduction of the unitary cost (i.e. cost per registered case) of a registration system.

Each registry has implemented its own set of automatic acceptance rules but, in almost all instances, the share of automatically registered cases is far from 100%. Therefore, the question arises about whether and how feasible it is to further reduce the burden of manual registration.

In the Venetian Tumour Registry (Registro Tumori del Veneto, RTV), which has relied on automatic evaluation programmes since its beginning [2,3], the percentage of automatic acceptance on the total number of registered cases is around 55%. Approximately, 8000 subjects per registration year must be manually evaluated, on a population base of two million residents.

Three types of rejection account for most of manually defined cases:

1) About 30% are rejected because referred only by hospital discharge records, without confirmation by pathology records and/or death certificates;
2) In 27% of instances, the programme is unable to choose among a series of disagreeing diagnoses;
3) In another 22% of cases, the programme does not state whether a second cancer really occurred or not, since recent diagnoses are discordant with a previously registered cancer.

The remaining occurrences sum up to 21%, and include various types of situations, like childhood cases and unlikely or rare tumours.

In general, cases are manually evaluated when the available diagnoses do not agree or their diagnostic base is regarded as "weak", i.e. not precise or not very reliable. An experienced cancer registrar may very often draw a conclusion about the case by simply using the same information submitted to the decision programme, otherwise further information sources (clinical records, diagnostic records in verbal form) are examined. Informative items, already available but not exploited by the decision programme, like surgical interventions and therapies, can often be decisive to define the case. To embed the solving rules applied by the registrar into a computer programme, however, may be quite

* Corresponding author. Fax: +39 049 8215983.
E-mail address: sandro.tognazzo@ioveneto.it (S. Tognazzo)

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.06.002

difficult, since the situations are more complex than those solved by the existing programmes.

As an alternative to algorithms based on deterministic rules, we could think of applying probabilistic models, commonly used for forecasting in several contexts but poorly considered in automatic cancer registration. Case definition can be seen as a classification problem in a "supervised learning" frame [4]. In simpler words, we may classify a subject as an incident cancer case rather than a "false positive", based on the available diagnostic evidence, and other relevant characteristics (age, sex, treatment) using a convenient statistical model, fitted on a set of subjects already manually classified and referred to as "training" set. The resulting model is used to classify a second set of subjects, always with known classification, and called "test" set, in order to determine the classification error. If such an error is sufficiently low, the model can be applied to new candidates.

To reduce the probability of underestimating the classification error, more than one test set is generally used. This is usually done by deriving several training and test sets from the available data ("learning" set), using cross-validation or bootstrap sampling.

The present paper deals with an exploratory application of such an approach to the third category of manually evaluated subjects, formerly outlined with regard to RTV.

2. Methods

A summary of data processing steps is shown in Fig. 1.

2.1. Learning set and outcome variable

The exercise was carried out on a set of 5608 subjects, registered as cancer cases during the period 1987–1996. All were scrutinized by RTV registration personnel to update the incidence data for the 1997–1999 period, for new diagnoses that were discordant with previously registered tumours.

Usually, one of these diagnoses is the most likely to be registered as a new cancer. Such site was individuated, by ordering discordant sites by combination of diagnostic sources, base of diagnosis and number of diagnoses reporting the site. For example, a diagnosis referred both from hospital discharge and pathology records has higher evidence than one based on a pathology source alone, which in turn has higher evidence than one reported exclusively by a hospital discharge record.

The outcome variable was categorized into the following four levels:

1. PREV: case confirmed as prevalent with no relevant modifications (3048 subjects, 54.4%);
2. NEW1: case confirmed and adding a new cancer (except for non-melanotic skin cancers), corresponding to the recent diagnosis with strongest evidence (1905 subjects, 34%);
3. NEW2: case confirmed and adding a new cancer (except for non-melanotic skin cancers), not corresponding to the recent diagnosis with strongest evidence, or adding more than one cancer (266 subjects, 4.8%);
4. MOD: case whose registered cancer was substantially modified or not confirmed, irrespective of the possible recognition of a further cancer (389 subjects, 6.9%).

The analysis was focused on assessing to what extent the automatic identification of the first two outcomes allows acceptable error rates.

To this aim, we chose to use multinomial logistic models [5] and random forests [6], since, they allow to identify those predictive variables playing a major role in classification; an issue of great interest, particularly in an exploratory phase, which other methods, like discriminant analysis and neural networks, are not well suited to.

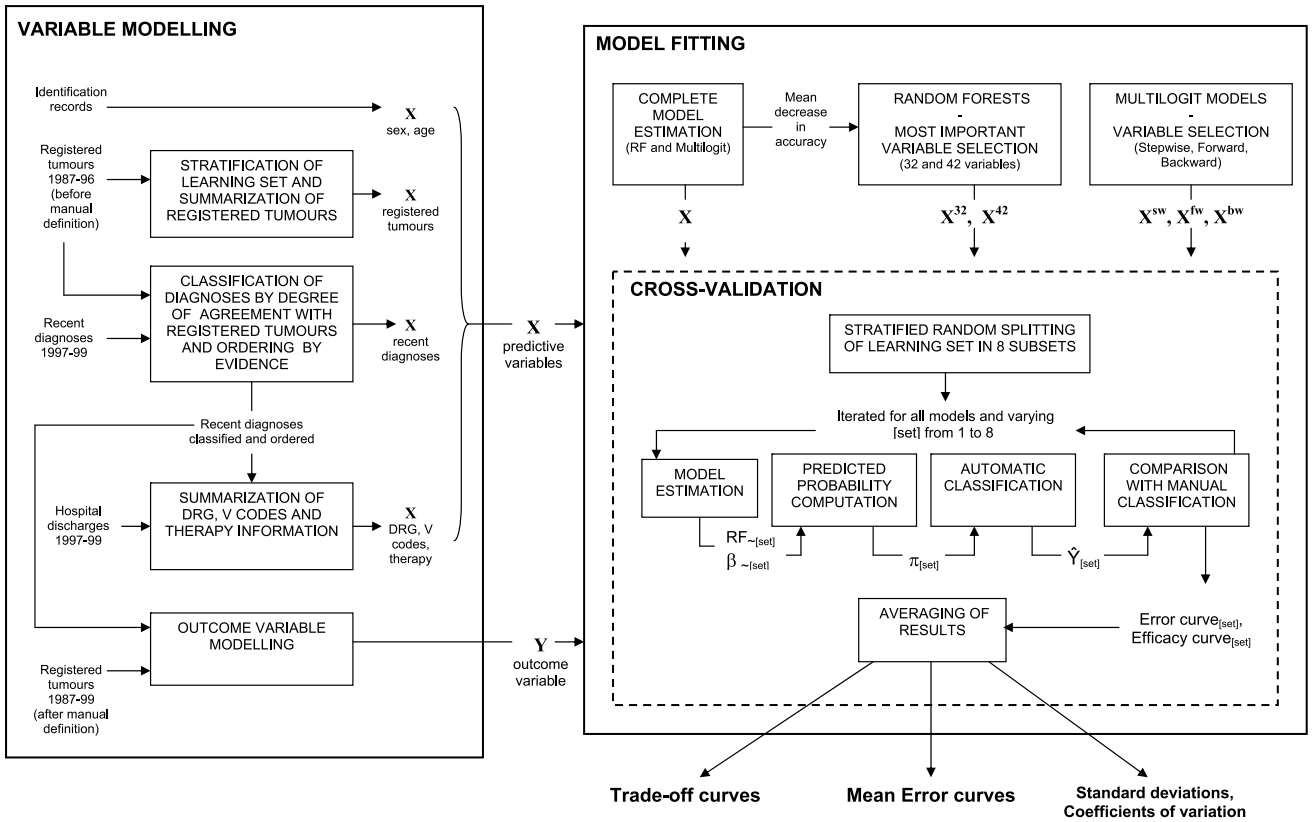

Fig. 1. Summary of variable modelling and classifier fitting.

2.2. Predictive variables

The predictive variables entered into the classification models are listed in Table 1. The following definitions and terms apply to several variables:

- Tumour sites are coded following the ICD-9 nomenclature [7];
- Histological groups are those groups of malignant neoplasms considered to be histologically 'different' for the purpose of defining multiple cancers [8];
- Diagnostic base is coded using five ordinal levels: 0 = only death certificate, 1 = clinical, 2 = cytological, 3 = histology of metastasis, 4 = histology or autopsy of primary site;
- The combination of sources reporting a diagnosis is coded using six ordinal levels: 1 = DC, 2 = H, 3 = P, 4 = H + DC, 5 = H + P or DC + P, 6 = H + P + DC (where H, P, DC indicate, respectively, hospital discharge records, pathology records, and death certificates);
- Indicator variables assume only values 1 or 0, depending on the existence of the named characteristic.

The items related to the recent diagnoses are grouped according to a categorization, applied by the evaluation programme in use [2,3], and based on the level of agreement with the registered tumours. In short, a diagnosis may be:

- concordant, when the site is the same as a registered cancer;
- compatible, when reporting:
  - an ill-defined or unknown primary cancer site or a metastasis;
  - a non melanotic skin cancer or a non malignant tumour;
  - some primary sites different from a neighbouring registered site;
  - discordant, when reporting a primary site different from all registered sites, and neighbouring with none of them.

Three particular groups of informative items were drawn from hospital discharge records:

- diagnosis-related group of hospitalization (DRG) mentioning a tumour;
- codes of the ICD-9 supplementary classification of factors influencing health status (V codes) mentioning a tumour;
- chemotherapy and radiotherapy treatment, derived from DRG, V codes or medical procedures ICD-9 CM codes.

These items were associated to the registered cancer or to a recent diagnosis, according to the ICD-9 site reported in the same discharge record.

2.3. Multinomial logistic models

To reduce computational problems caused by sparse data, a multinomial logistic analysis was carried out considering only three classes, by grouping the two less frequent outcomes (NEW2 and MOD). On the whole learning set, logistic regressions were fitted for the logits of the aggregate class versus NEW1 and of class PREV versus NEW, using the LOGISTIC procedure of the SAS package [9]; in addition to the complete model (63 variables), another three models were fitted, using only the relevant predictive variables detected by stepwise (31 variables), backward (37 variables) and forward (36 variables) selection.

The importance measure used to rank variables is Wald χ^2 [5,10], a statistic usually used to test the significance of a specific variable. The partial effect of a variable, on the target outcome probabilities, is expressed by the regression coefficient of the logit of PREV versus NEW1.

2.4. Random forest classifiers

Considering all the four outcomes, a random forest of 500 classification trees was "grown" on the learning set and the predictive variables were ranked by a measure of importance, called "mean decrease in accuracy"; another two forests of the same size were then grown, using only the 32 and 42 most important variables. The "random Forest" package for R environment [11] was used for calculations.

The mean decrease in accuracy is the average, over all trees, of the difference between the number of correct classifications obtained using the actual values of a variable and the number obtained using random permutations of them. On average, this difference is expected to be high when a variable has a strong discriminating power, because the classifier fails more often when the true values are substituted by random ones.

The effect of a variable on each target outcome probability is described by a partial dependence function, which associates, to each value of the variable concerned, the average over the subjects exhibiting that value of the logits of the predicted probabilities for the specific outcome.

2.5. Cross-validation

To estimate the classification error and the proportion of cases automatically assigned to the classes PREV and NEW1, an eightfold cross-validation has been performed.

Firstly, the learning set was partitioned according to the registered cancer site in fourteen strata, plus a further stratum for multiple cancers; subsequently, each stratum was randomly split into eight subsets of equal size (in total 701 subjects per set). Thus, each fold should be representative of the whole learning set, since each registered cancer site appears with the same proportion. The strata list includes: oral cavity, colon and rectum, other digestive organs, larynx, lung, breast, soft tissues and melanoma, female genital organs, prostate, urinary organs, lymphomas, leukemias and myelomas, other specified site, ill-defined or unknown site.

Secondly, for each subset, the following steps were iterated:

a. The cases falling into the other seven sets were used as a training set and all the models, mentioned in the previous paragraphs, were fitted;
b. The fitted models were applied to the subjects belonging to the current fold, used as a test set, in order to calculate the predicted probabilities of the two target outcomes;
c. Each case was then assigned to one of such outcomes, if the corresponding probability was greater or equal to a given threshold *P*. Such an attribution was compared with the manual outcome, to individuate wrong classifications. Varying *P* from 0.50 to 0.99, two curves were determined for each model:
   - The error curve, which shows the variation in the percentage of individuals wrongly classified over the classified ones (error rate);
   - The efficacy curve, which shows the variation in the proportion of individuals classified in the test set (classification rate).

Thirdly, the two curves were averaged over the eight subsets, to obtain the mean error and mean efficacy curves, as well as the curves of their standard errors and coefficients of variation (ratio of the standard error to the mean).

Finally, a "trade-off" curve was obtained for each model, by matching the mean error rate and the mean classification rate hav-

Table 1
Predictive variables used for classification

Variable group	Description	Label
Demographic data	Gender	SEX
	Age class at the incidence date of the earliest tumour with worst behaviour registered	AGE
Registered tumours	Number of further primary cancers, not melanotic skin cancers excluded	NPRIM2
	Number of not malignant tumours and not melanotic skin cancers	NIGN
	Site or group of sites of the first primary cancer	CLASTUM1
	Site or group of sites of the second primary cancer	CLASTUM2
	Histological group of the first primary cancer	ISTGRUP1
	Histological group of the second primary cancer	ISTGRUP2
	Highest diagnostic base among primary cancers	BASE
Recent diagnoses concordant with a registered cancer	Number of concordant diagnoses	NUM_CON
	Indicator of difference in histological group with respect to the registered cancer	ISTDIF_CON
	Highest diagnostic base among concordant diagnoses	BASE_CON
	Sources of concordant diagnoses	SOUR_CON
Recent diagnoses not concordant with registered cancers	Proportion of recent diagnoses not concordant	PERC_NOCON
	Indicator of difference in histological group with respect to the registered cancer	ISTDIFF
	Time interval between the last registered cancer and the most discordant diagnosis	DATDIF_D
Recent diagnoses not concordant but "compatible" with registered cancers	Number of diagnoses referring cancer of oral cavity or larinx	C10
	Number of diagnoses referring cancer of digestive organs	C11_14
	Number of diagnoses referring cancer of respiratory organs	C15_16
	Number of diagnoses referring cancer of female genital organs	C17_18
	Number of diagnoses referring cancer of prostate or urinary organs	C19
	Number of diagnoses referring lymphoma or leukaemia	C20_22
	Highest diagnostic base among "compatible" diagnoses	BASE_CMP
	Sources of "compatible" diagnoses	SOUR_ CMP
Recent diagnoses reporting metastasis, ill-defined or unknown primary cancer site	Number of diagnoses reporting ill-defined or unknown primary cancer site	D195_199
	Number of diagnoses reporting metastasis of lymphnodes	D196
	Number of diagnoses reporting metastasis of respiratory or digestive organs	D197
	Number of diagnoses reporting metastasis of other organs	D198
	Highest diagnostic base	BASE_MU
	Sources of diagnoses	SOUR_ MU
Recent diagnoses reporting a well-defined primary cancer different from all registered cancers	Number of discordant diagnoses	NDIA_D
	Discordant site with highest evidence	ICD9_DIS1
	Diagnostic base of the site with highest evidence	BASE_DIS1
	Sources of the discordant site with highest evidence	SOUR_ DIS1
	Discordant sites with second highest evidence	ICD9_DIS2
	Diagnostic base of the site with second highest evidence	BASE_DIS2
	Sources of the discordant site with second highest evidence	SOUR_ DIS2
	Discordant site with third highest evidence	ICD9_DIS3
Recent diagnoses reporting not melanotic skin cancer or not malignant tumours	Number of diagnoses reporting not melanotic skin cancer	D173
	Number of diagnoses reporting benign tumour	D210_229
	Number of diagnoses reporting carcinoma "in situ", uncertain, unknown behaviour tumour	D230_239
	Highest diagnostic base	BASE_NM
	Sources of diagnoses	SOUR_NM
DRG codes associated with a concordant diagnosis	Number of hospitalizations reporting a medical DRG mentioning tumour	DRG_M_CON
	Indicator of surgical DRG mentioning tumour	DRG_C_CON
	Number of hospitalizations reporting DRG not compatible with the associated diagnosis	DRG_NOCMP_CON
DRG codes associated with the discordant cancer site with highest evidence	Number of hospitalizations reporting a medical DRG mentioning tumour	DRG_M_DIS1
	Indicator of surgical DRG mentioning tumour	DRG_C_DIS1
	Indicator of DRG not compatible with the associated diagnosis	DRG_NOCMP_DIS1
DRG codes associated with other discordant cancer sites	Indicator of surgical DRG mentioning tumour	DRG_DIS2
	Indicator of DRG not compatible with the associated diagnosis	DRG_NOCMP_DIS2
DRG or ICD-IX V codes reporting anamnesis of cancer	Indicator of DRG reporting anamnesis of cancer	DRG_ANAM
	Frequence of V codes compatible with a registered cancer	V_CMP
	Frequence of V codes compatible with the discordant site having highest evidence	V_DIS1
	Indicator of V codes compatible with a discordant site having lower evidence	V_DIS2
	Indicator of V codes compatible with not compatible with any recent diagnosis or registered cancer	V_ALT
	Indicator of V codes compatible with reporting execution of diagnostic procedures to ascertain cancers	V_OSS
Chemiotherapy	Indicator of association with a registered cancer	CHT_CON
	Indicator of association with the discordant site having highest evidence	CHT_DIS1
	Indicator of association with another discordant diagnosis or no association	CHT_DIS2
Radiotherapy	Indicator of association with a registered cancer	RT_CON
	Indicator of association with the discordant site having highest evidence	RT_DIS1
	Indicator of association with another discordant diagnosis or no association	RT_DIS2

ing the same threshold *P*. This curve synthesizes the information needed to assess the usefulness of a classifier. In fact, it indicates the cost, in terms of error, we must accept for automatically classifying a given share of subjects (and achieving a corresponding reduction in registrars work).

3. Results

In Fig. 2 the mean error curves associated to the estimated models are shown. Random forests always classify with a markedly lower error than the multilogit models; for example, choosing a probability greater or equal to 0.85, the error for the first class of models is around 3.7%, while that of the second ranges between 7.4% and 8.1%.

For all models, the variability of error rates is not very high; the coefficients of variation are always lower than 20%, when errors are higher than 2%, and decrease to less than 10% at a 3.5% rate.

Considering as acceptable a mean classification error lower than 5%, the trade-off curves of Fig. 3 show that we could automatically decide up to 32% of cases, using multilogit models, and at least 45%, using random forests (RF); thus, the latter classifiers clearly give a better performance.

Defining as "best" classifier the model which exhibits the highest classification rate, for a given error rate, we found that the optimal model is the RF including:

- all the 63 variables (complete model), for error rates lower than 4%;
- only the 42 most important variables, for error rates between 4% and 5%.

Actually, it is quite possible that the complete model over-fits the data and, therefore, extremely low errors (2% or less) are unrealistic, when applying the classifier to a different data set. Moreover, in most instances, the error rates of the more parsimonious model differ from those of the complete model less than the standard error, when the *P* threshold is less or equal to 0.90 (see Fig. 2). These arguments suggest to choose the more parsimonious model,

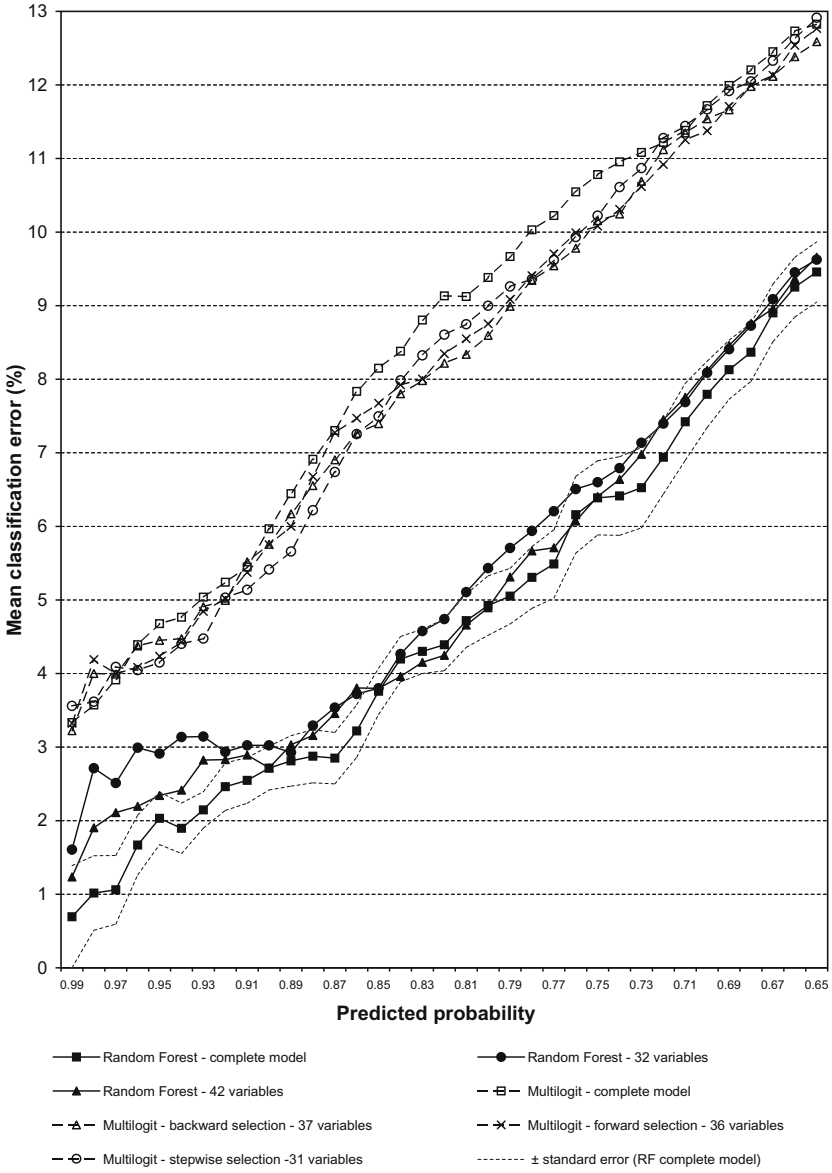

Fig. 2. Error curves.

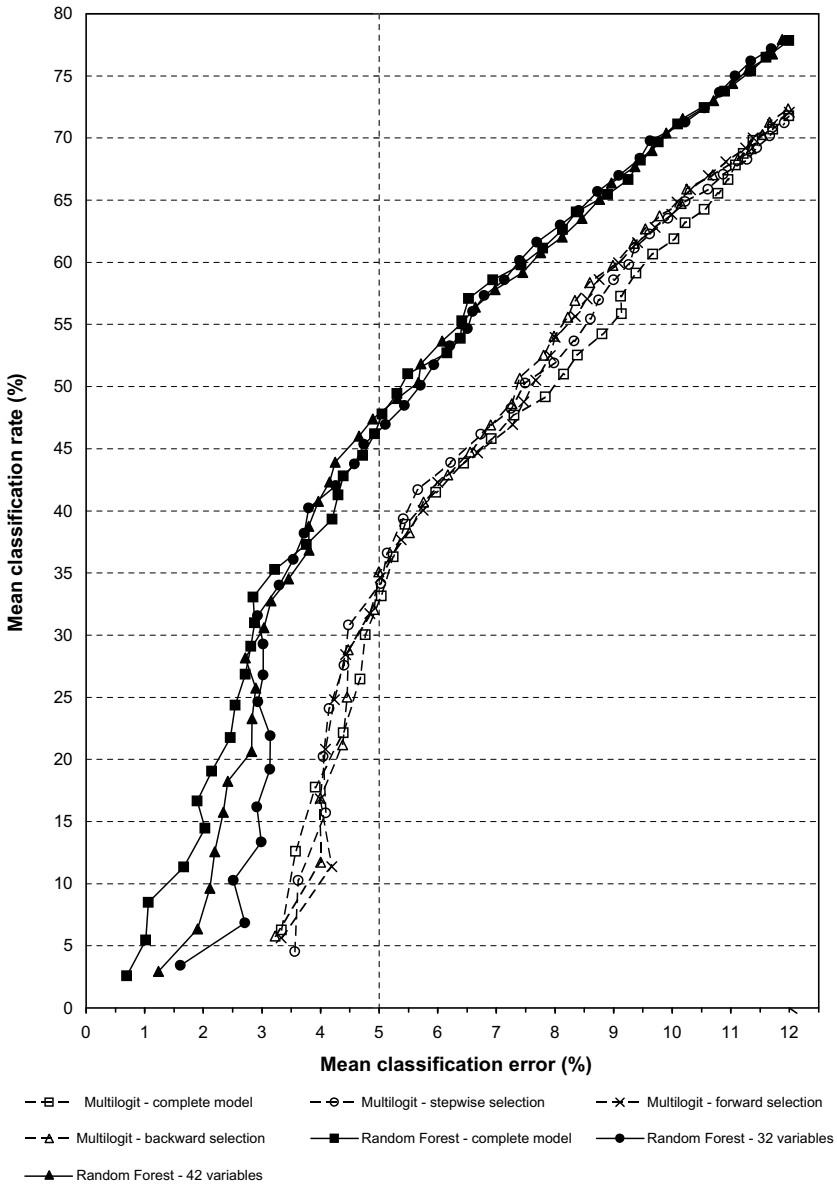

Fig. 3. Trade-off curves between classification error and classification rate.

for practical use, and to regard a rate around 3% as a likely lower bound for the classification error.

3.1. Scrutiny of misclassified subjects

A mean error lower than 5%, for RF models, occurs at a 0.82 predicted probability (Fig. 2) and corresponds to 114 misclassifications, on the whole learning set.

From the scrutiny of these subjects, two relevant findings have emerged:

(a) Thirty-four cases (30%) had correctly classified recent diagnoses and the majority of them (29 subjects) fall into the category labelled as MOD (which accounts for 33 misclassifications). This means that the registrar has modified the information about the already registered cancer, but he/she has taken the same decision as the classifier, regarding the acceptance of a second tumour.

(b) The classifier regularly fails to detect second cancers on the same site but with a different morphology. In this rare situation, the recent diagnosis appears concordant with the registered cancer to all respects, except for the histological group. This kind of error accounts only for 9% of the misclassifications, but it is systematic and, likely, beyond the "learning" ability of any model.

3.2. Importance and role of predictive variables

Table 2 lists the 31 most important predictive variables of each model and their rank. Seven of them are in the first ten ranks in both models, so exhibiting prominent importance:

- three variables relate to the diagnoses suggesting a second cancer: absolute frequency, diagnostic sources and site (NDIA_D, SOURCES_DIS1, ICD9_DIS1);

Table 2
Most important predictive variables

Variable group	Variable label	Random forest		Multinomial logit model		
		Rank	Mean decrease in accuracy × 1000	Rank	Partial effect measure (Wald X^2)	Degrees of freedom
Demographic data	SEX	27	2.10	28	9.41	2
	AGE	22	4.48	Not significant		
Registered tumours	NIGN	>31		25	10.96	2
	NPRIM2	>31		30	8.12	2
	CLASTUM1	**5**	**24.39**	**3**	**207.06**	**28**
	ISTGRUP1	**8**	**12.89**	**8**	**55.63**	**12**
	BASE	15	9.02	**6**	**88.78**	**2**
Recent diagnoses concordant with a registered cancer	NUM_CON	11	10.95	22	14.02	22
	ISTDIF_CON	>31		18	18.44	2
	BASE_CON	17	6.42	24	11.51	2
	SOUR_CON	13	10.42	**10**	**40.94**	2
Recent diagnoses not concordant with registered cancers	**PERC_NOCON**	**6**	**14.13**	19	17.16	2
	ISTDIFF	**9**	**12.49**	**5**	**112.16**	2
	DATDIF_D	12	10.95	11	37.18	2
Recent diagnoses not concordant but "compatible" with registered cancers	C10	>31		14	27.66	2
	C11_14	28	1.72	16	23.88	2
	C19	24	2.76	15	23.96	2
	C17_18		31	7.27	2	
	C20_22	31	1.26	21	14.09	2
	BASE_CMP	**10**	**11.68**	Not significant		
	SOUR_CMP	**7**	**13.31**	**9**	**55.49**	**2**
Recent diagnoses reporting metastasis, ill-defined or unknown primary cancer site	D197	23	3.04	13	33.52	2
	D198	30	1.46	29	9.38	2
	BASE_MU	21	4.48	23	11.97	2
	SOUR_MU	19	5.49	Not significant		
Recent diagnoses reporting a well-defined primary cancer different from all registered cancers	**NDIA_D**	**2**	**54.07**	**4**	**124.44**	**2**
	ICD9_DIS1	**3**	**42.96**	**2**	**338.00**	**40**
	BASE_DIS1	**4**	**32.12**	Not significant		
	S0UR_DIS1	**1**	**88.30**	**1**	**388.74**	**2**
	ICD9_DIS2	14	9.49	**7**	**84.36**	**18**
	BASE_DIS2	18	6.05	Not significant		
	S0UR_DIS2	16	6.84	12	36.82	2
	ICD9_DIS3	29	1.56	Not significant		
Recent diagnoses reporting not malignant tumours	D230_239	>31		26	10.26	2
DRG codes associated with a concordant diagnosis	DRG_M_CON	25	2.42	Not significant		
	DRG_NOCMP_CON	>31		Not significant		
DRG or V codes associated with a discordant cancer site	DRG_M_DIS1	20	5.37	Not significant		
	DRG_C_DIS1	26	2.25	17	21.48	2
	DRG_N0CMP_DIS2	>31		20	14.37	2
	V_DIS1	>31		27	10.49	2

Variables marked in bold are in the first ten ranks.

- two items relate to the registered cancer: site and histological group (CLASTUM1, ISTGRUP1);
- two variables concern the similarity of new diagnoses to the registered cancer: indicator of histological difference and diagnostic sources indicating a "compatible" site (ISTDIFF, SOURCE_CMP).

The influence of a variable on the classification outcome is reflected by the sign and magnitude of regression coefficients (β), for the multilogit model, and by the trend of partial dependence functions (p.d.f.), for the random forest.

When higher values of the variable imply a higher probability of class PREV against NEW1, the corresponding β is positive and the p.d.f. of class PREV increases, while that of NEW1 decreases. Instead, a negative β and opposite trends of the p.d.f. reflect increasing probabilities of class NEW1 against PREV.

Due to space limitations, results are reported only for a few variables.

The probability of a diagnosis being accepted as a second cancer increases when the combination of reporting sources (SOURCES_DIS1) or the frequency it occurs (NDIA_D) increase; β values for the logit of PREV over NEW1 are, respectively, −0.98 and −0.35, while the p.d.f. for class NEW1 increases and that for PREV decreases (Figs. 4 and 5). These are common-sense indications; a diagnosis is more plausible when frequently reported and referred by more sources.

Furthermore, the probability of accepting a second cancer:

- increases, when the site referred by the new diagnosis (ICD9_DIS1) is prostate or breast, since the β values for the logit of PREV over NEW1 are −1.56 and −0.75;
- decreases, when the referred site is peritoneum or bones, an ill-defined site of the digestive or respiratory apparatus, brain and nervous system, soft tissues and melanoma; β values are 4.76, 2.66, 1.34, 1.53, respectively. All of these are frequent metastatic sites.

Such indications are confirmed by the p.d.f. of Figs. 6 and 7.

Once more, these are sound results, since an error often detected in the diagnostic sources consists exactly in coding metastatic tumours as primary cancers.

Finally, the p.d.f. for BASE_DIS1 and CLASTUM1, not shown, indicate that the probability of accepting a second cancer increases when the new diagnosis is based on a histology, but lowers when the registered cancer is ill-defined or concerns digestive organs, brain, thyroid and some rare sites as bones, nose, pleura.

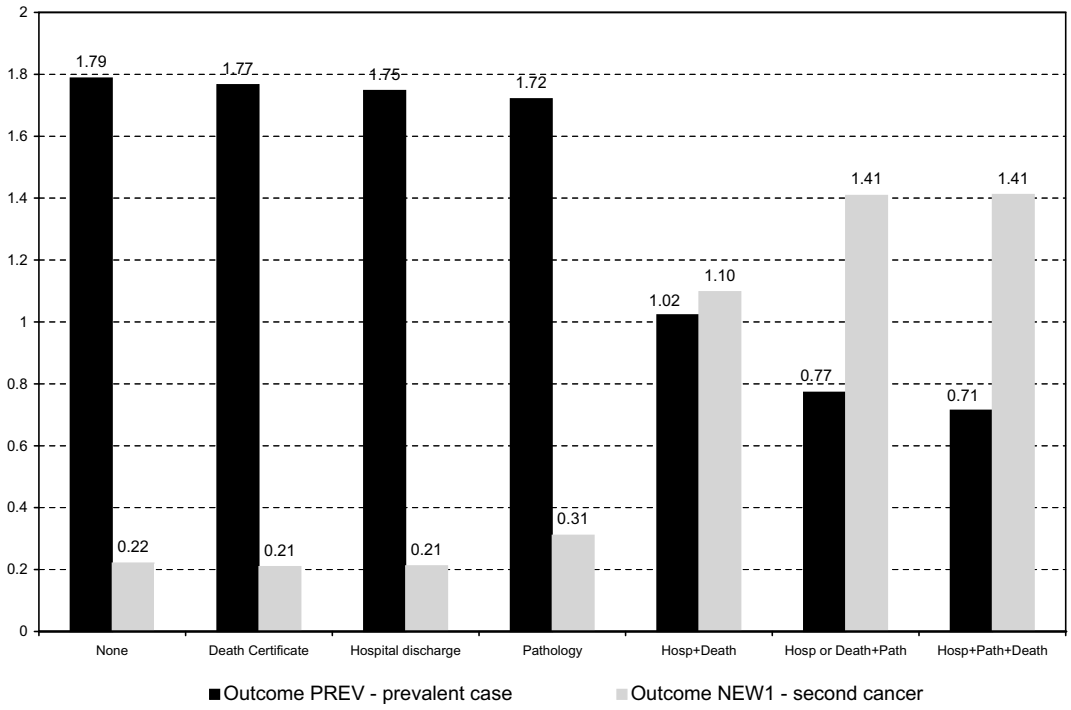

Fig. 4. Partial dependence functions—Variable SOUR_DIS1. (Combination of sources reporting the candidate second cancer).

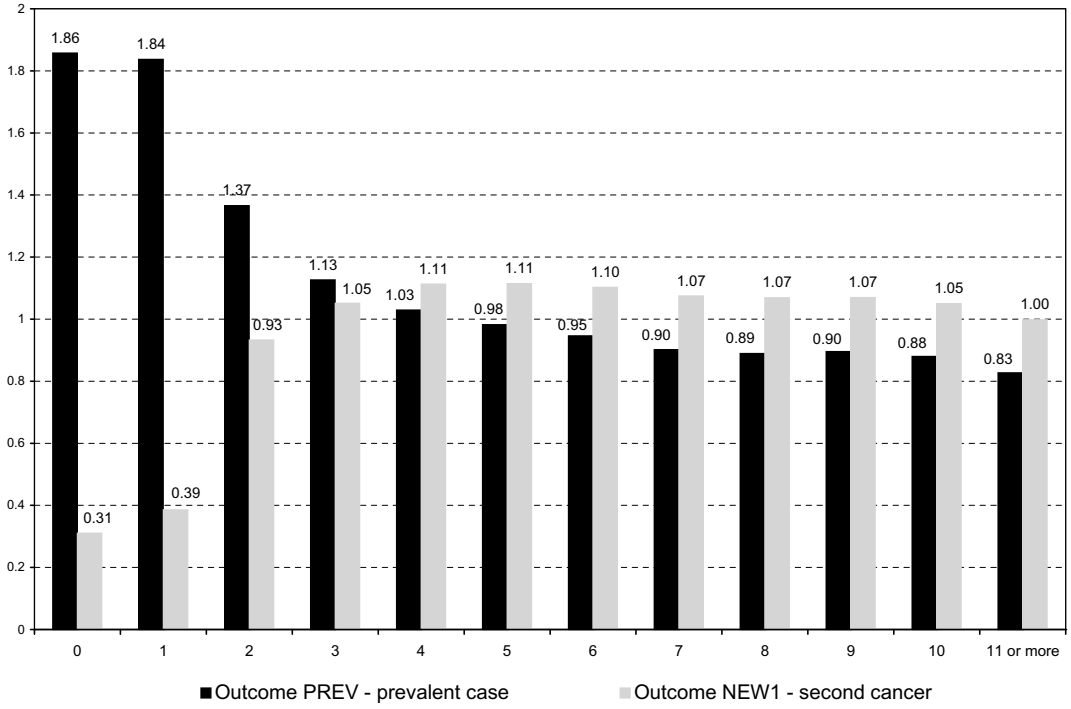

Fig. 5. Partial dependence functions—Variable NDIA_D. (Number of discordant diagnoses reported).

4. Conclusions

The present analysis was intended to answer a preliminary question: may we significantly increase the share of automatically defined cases, without lowering the data quality, using probabilistic classifiers?

Almost all studies, aimed at assessing the data quality of cancer registration [1,12–15], have shown that errors of some kind affect from 5% to 10% of registered cases; only one has reported a lower error rate [16]. In particular, a study concerning the cases automatically accepted by RTV registration system [15] gives the following figures: 1% of undetected second cancers, 2% of prevalent cases

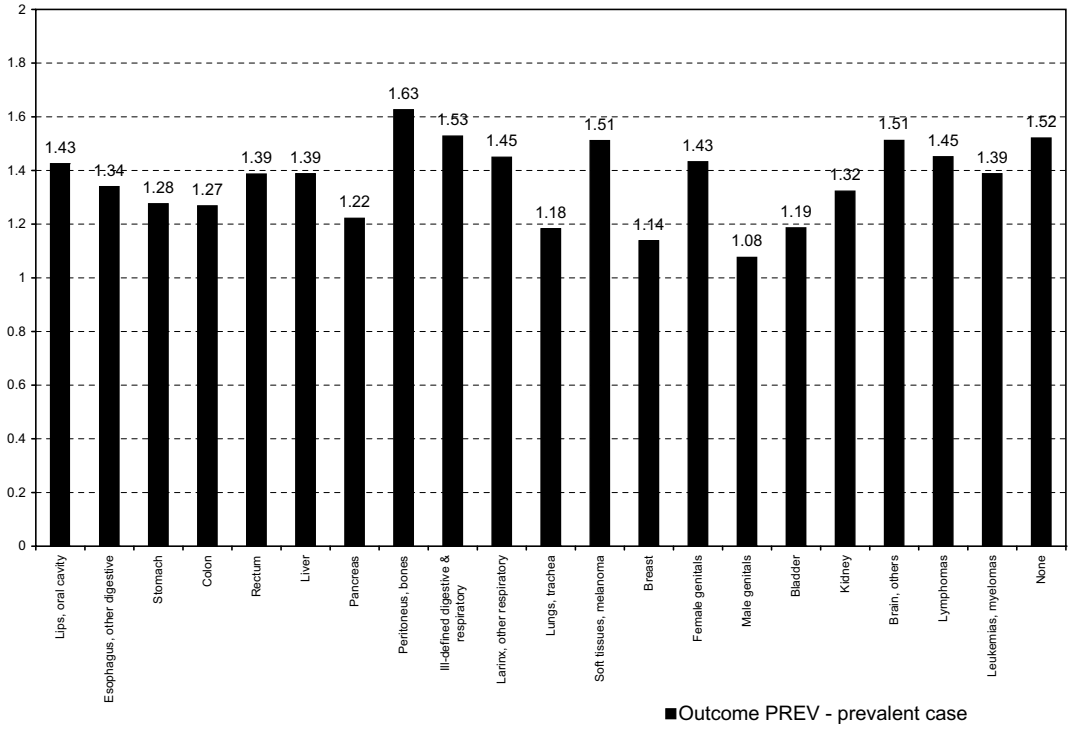

Fig. 6. Partial dependence function—Variable ICD9_DIS1. (Site of the candidate second cancer).

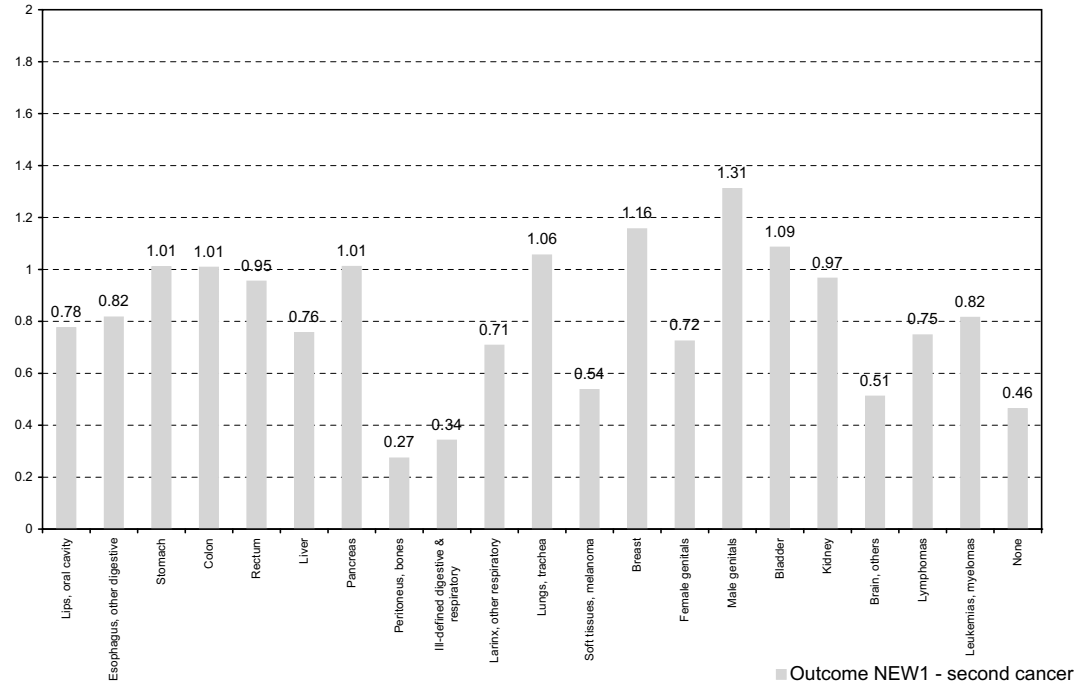

Fig. 7. Partial dependence function—Variable ICD9_DIS1. (Site of the candidate second cancer).

wrongly accepted as incident, 5.2% of cases with misclassification of the tumour site, 1.5% of "false positive" cases.

Thus, it seems reasonable to accept a 5% ceiling for the classification error. With this assumption and looking back at the trade-off curve in Fig. 3, we see that multinomial logistic models allow to reduce the share of subjects to be manually evaluated by 31% (545 cases over 1750 cases per registration year), but random forest models by 45% (790 cases). Both classifiers exhibit a significant performance, but random forests clearly appear more promising.

It has also been shown that the RF model is less prone to error than it appears at first glance, as the evaluation of recent evidence may be correct even if the previous cancer was not well registered. This circumstance has been concealed by the rough definition of the residual category of the outcome variable.

Moreover, the role played by the predictive variables and their rank of importance agree with what could be expected based on registrars' experience and rationale considerations, so the model is interpretable and not counterintuitive.

Therefore, the answer to the question seems to be affirmative and it would be worthwhile to deepen the present analysis and to extend this approach to the other categories of manually defined cases. The performance of alternative methods, such as the discriminant analysis, could also be tested.

As previously discussed, a more accurate definition of the outcome variable is surely needed; the inclusion of further predictive variables should also be considered. In particular, to extend this approach to subjects referred only by hospital discharge records (the largest category of manually defined cases in our Registry), the information concerning surgical interventions and other therapeutic and diagnostic procedures cannot be ignored.

A concern involves second cancers on the same site that may be detected only manually and should be excluded from automatic classification. This could easily be done by checking the compatibility between morphologies.

The classifier fitted on the RTV learning set is not expected to perform similarly when applied to data of other Registries, since the structural features of the diagnostic sources are generally different. However, each Registry could apply the same methodology on its own learning set, in order to identify a proper classifier.

Probably, the main concern about the classification techniques outlined relates to their reliability over time, which could be undermined by changes in the quality and coverage of the available information, as well as in diagnostic and therapeutic practices. Major changes tend to affect the share of automatically acceptable subjects, rather than the accuracy of classification. However, an upsurge of the classification error can never be excluded.

It must be pointed out that the same concern involves the automatic decision systems currently applied by RTV and other registries. A convenient way of detecting the faults of the automatic classification system is to assess the quality of registered data in a routinely way, by periodically drawing a sample of "automatic" cases and manually defining them.

References

[1] Black RJ, Simonato L, Storm HH, Démaret E. Automated data collection in cancer registration. IARC Technical Reports No. 32. Lyon: IARC; 1998.
[2] Simonato L, Zambon P, Rodella S, Giordano R, Guzzinati S, Stocco C, et al. A computerized cancer registration network in the Veneto region, north east of Italy: a pilot study. Br J Cancer 1996;73:1436–9.
[3] Bovo E, Tognazzo S, Monetti D, Andolfo A, Fiore Ar, Guzzinati S, et al. RTV-Evaluate Evidence. Società Italiana degli Autori ed Editori—Registro Pubblico speciale per i programmi per elaboratore. Registrazione SIAE: 22/01/2003 n 002525, D003356;2003.
[4] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2001.
[5] Agresti A. Categorical data analysis. New York: John Wiley; 1990.
[6] Breiman L. Random forests. Machine Learning 2001;45:5–32.
[7] World Health Organization 9th Revision Conference, 1975. Manual of the International Statistical Classification of Diseases injuries and Causes of Death. vol. 1. Geneva: WHO; 1977.
[8] Parkin DM, Chen VW, Ferlay J, Galceran J, Storm HH, Whelan S. Comparability and quality control in cancer registration (IARC Technical Reports no. 19). Lyon: IARC; 1994.
[9] SAS Institute Inc. "SAS/STAT User's Guide—The LOGISTIC Procedure" in: SAS OnlineDoc, Version 9, Cary, NC, SAS Institute Inc. 2003.
[10] Wald A. Test of statistical hypotheses concerning several parameters when the number of observation is large. Trans Am. Math Soc 1943;54: 426–82.
[11] Breiman L, Cutler A, Liaw A, Wiener M. Breiman and Cutler's random forests for classification and regression version 4.5–10. 2005; Available from http://cran.r-project.org/src/contrib/Descriptions/randomForest.html..
[12] West RR. Accuracy of cancer registration.Brit J Prev Soc Med 1976;30:187–92.
[13] Lapham R, Waugh NR. An audit of the quality of cancer registration data. Br J Cancer 1992;66:552–4.
[14] Brewster DH, Crichton J, Muir C. How accurate are scottish cancer registration data? Br J. Cancer 1994;70:954–9.
[15] Tognazzo S, Andolfo A, Bovo E, Fiore AR, Greco A, Guzzinati S, et al. Quality control of automatically defined cancer cases by the automated registration system of the Venetian Tumour Registry. Eur J Public Health 2005;15(6):657–64. Epub 2005 Jul 28.
[16] Tagliabue G, Maghini A, Fabiano S, Tittarelli A, Frassoldi E, Costa E, et al. Consistency and accuracy of diagnostic cancer codes generated by automated registration: comparison with manual registration. Population Health Metrics 2006;4:10. doi:10.1186/1478-7954-4-10.