BMC
Medical Research Methodology

RESEARCH ARTICLE   **Open Access**

Validation of death prediction after breast cancer relapses using joint models

Audrey Mauguen[1,2*], Bernard Rachet[2], Simone Mathoulin-Pélissier[3,4], Gill M Lawrence[5], Sabine Siesling[6], Gaëtan MacGrogan[3], Alexandre Laurent[1] and Virginie Rondeau[1]

Abstract

Background: Cancer relapses may be useful to predict the risk of death. To take into account relapse information, the Landmark approach is popular. As an alternative, we propose the joint frailty model for a recurrent event and a terminal event to derive dynamic predictions of the risk of death.

Methods: The proposed prediction settings can account for relapse history or not. In this work, predictions developed on a French hospital series of patients with breast cancer are externally validated on UK and Netherlands registry data. The performances in terms of prediction error and calibration are compared to those from a Landmark Cox model.

Results: The error of prediction was reduced when relapse information was taken into account. The prediction was well-calibrated, although it was developed and validated on very different populations. Joint modelling and Landmark approaches had similar performances.

Conclusions: When predicting the risk of death, accounting for relapses led to better prediction performance. Joint modelling appeared to be suitable for such prediction. Performance was similar to the landmark Cox model, while directly quantifying the correlation between relapses and death.

Keywords: Breast cancer, Joint frailty model, Landmark, Prediction, Relapse history, Survival

Background

Individual predictions are increasingly sought after to help guide treatment decisions and patient care. Accurate predictions are especially important in the context of personalised medicine, where the ultimate goal is to give personalised targeted treatment to every patient. To do so, it is important to evaluate patient prognosis, accounting for their individual characteristics. In recent years, prognosis research in cancer has focused mainly on the presence of biomarkers that can be targeted by treatments. Less focus has been given to the impact of relapses that patients may experience, such as loco-regional relapses or distant metastases in cancer patients, which may explain a large part of the risk of death, despite adequate methods of

analysis currently available. The impact of these relapses on the risk of death may vary according to the type of cancer. Relapses can be considered as a surrogate for patient frailty or for disease aggressiveness. It is therefore of interest to investigate how these events can be used to predict patient survival.

Relapses are recurrent events, evolving during patient follow-up. Thus it is of interest to study the impact of recurrences on death. However, relapses and death are non-independent events, sometimes called semi-competing risks, and relapses cannot be included in a survival model as a standard time-dependent covariate to study the risk of death [1]. The Landmark approach [2,3] resolves the problem by updating the population of interest. At each chosen prediction time, the model is estimated again on patients still alive. Thus, relapse history, observed before the prediction time, can be resumed as a baseline covariate, such as the number of previous relapses. This method offers the advantage of a simple and robust model. However, to do some dynamic predictions using several prediction times, several models have to be

*Correspondence: audrey.mauguen@isped.u-bordeaux2.fr
[1]Biostatistic unit, INSERM U897, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France
[2]Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London, UK
Full list of author information is available at the end of the article

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 2 of 12

run on sub-populations of alive patients. Moreover, summarising the relapse process in a single variable may result in a loss of information.

Alternatively, joint modelling can be used to study recurrent events [4,5]. The hazard of the recurrent event and the hazard of the terminal event (death) are jointly modelled. Such models allow us to fully consider the correlation between the two processes using a shared random effect (frailty). Dynamic predictions can then be derived, accounting for all previous events. Once the parameters are estimated, predictions can be updated without running the model again. A recent work investigated the impact of relapses on the risk of death in breast cancer using a joint frailty model [6]. The proposed method was shown to be an adequate framework for prediction, and the model seemed to give a satisfying performance on the sample used to develop the model (development sample). However, the high number of parameters to be estimated may be a concern in prediction context, especially for generalisability purpose. It is therefore essential to validate the proposed prediction on independent data [7-9].

Here, our main goal is to validate a method of predictions from a joint model for recurrent events and death as accurate predictions. We present the results of the external validation of the previously developed prediction based on relapses in addition to well-known prognostic factors [6]. In order to assess whether the high number of parameters is a concern, we compare this new prediction performances with those of a Landmark Cox model. As in the development step, we apply the proposed prediction on breast cancer cases, here from two population-based registries, in West Midlands (England) and the Netherlands.

Section "Methods" of this paper explains the prediction probabilities within the framework of a Landmark Cox model and a joint frailty model, as well as the tools to validate them. The validation of the prediction on the West Midlands and Netherlands registry datasets is developed in section "Results". Finally, sections "Discussion" and "Conclusion" contain concluding remarks.

Methods

We are interested in the prediction of the risk of death between a prediction time s and a prediction horizon $s+w$ considering all the information available at time s. The information includes some baseline covariates, but also history of recurrent events (loco-regional relapse or distant metastasis) until time s. In this context, the predicted risk of death can be updated after each new recurrence.

Prediction of death in the joint modelling framework

The joint frailty model for a recurrent event and a terminal event is defined as follows [5]: for subject i ($i =$

$1, \ldots, N$), let X_{ij} be the jth recurrent time ($j = 1, \ldots, n_i$) measured from the study origin (calendar time), D_i be the death time and C_i be the independent censoring time. Note that the censoring time can happen to be a recurrence time. $T_{ij}^R = min(X_{ij}, C_i, D_i)$ corresponds to each follow-up time and δ_{ij}^R is a binary indicator for recurrent events which is 0 if the observation is censored or if the subject died, and 1 if X_{ij} is observed $\left(\delta_{ij}^R = I \left[T_{ij}^R = X_{ij} \right] \right.$ where $I[.]$ denotes indicator function). Similarly, we note T_i^D as the last follow-up time for subject i, which is either a time of censoring or a time of death $\left(T_i^D = min(C_i, D_i) \right)$ and $\delta_i^D = I \left[T_i^D = D_i \right]$. We actually observe the sequence $\left(T_{ij}^R, \delta_{ij}^R, T_i^D, \delta_i^D \right)$. Finally, we denote by Z_{ij}^R and Z_i^D the vectors of covariates associated with the hazard of recurrent events and death, respectively. However, a patient is considered at risk of a jth recurrence only after the $(j-1)^{st}$ recurrence. The joint model is then written as:

$$\begin{cases} \lambda_{ij}^R(t|u_i) = u_i \lambda_0^R(t) \exp\left(\beta_1' Z_{ij}^R \right) = u_i \lambda_{ij}^R(t) \\ \lambda_i^D(t|u_i) = u_i^\alpha \lambda_0^D(t) \exp\left(\beta_2' Z_i^D \right) = u_i^\alpha \lambda_i^D(t) \end{cases} \quad (1)$$

where $\lambda_0^R(.)$ is the baseline hazard of a recurrent event, irrespective of event rank, and $\lambda_0^D(.)$ the baseline hazard of death. The effects of explanatory variables β_1 and β_2 are assumed to be different for the hazard of recurrent events and the risk of death. The two processes are linked by the patient-specific frailty effect u_i. The frailty terms are independent and identically distributed following a gamma distribution with variance θ and, without loss of generality, a mean equal to 1. That is:

$$u_i \sim \text{Gamma}\left(\frac{1}{\theta}; \frac{1}{\theta} \right) \quad \text{and} \quad g(u_i) = \frac{u^{1/\theta-1}\exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}$$

$$(2)$$

The baseline hazard functions ($\lambda_0^R(.)$ for recurrent events and $\lambda_0^D(.)$ for death) are approximated using cubic splines, but alternative flexible functions such as fractional polynomials could be used. Splines are piecewise polynomials that are constrained to smoothly joint to fit curves. We used $\lambda_0^R(.) = \sum_{i=1}^m \eta_i^R M_i(.)$ and $\lambda_0^D(.) = \sum_{i=1}^m \eta_i^D M_i(.)$ where $M_i(.)$ is the common basis splines, and η^R and η^D the two vectors of splines coefficients. The parameters of the model $\xi = (\eta^R, \eta^D, \beta_1, \beta_2, \alpha, \theta)$ are estimated using penalized maximum likelihood estimators. For more details on the inference method, please see [5]. To estimate this model, all the available information is used, from the origin to the end of follow-up of all patients.

Using the joint modelling framework, we are interested in two prediction settings previously defined [6]. The first prediction of the risk of death P^{Rec} is calculated based on all relapses information. In this setting, the J relapses occurring before the prediction time s are considered ($J \leq$

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 3 of 12

n_i). We consider the patient history $\mathcal{H}_i^J(s) = \{N_i^R(s) = J, X_{i1} < \ldots < X_{iJ} \leq s\}$, with $X_{i0} = 0$ and $X_{i(J+1)} > s$, to define the conditional probability of death P^{Rec} as follows:

$$P^{Rec}(s, s + w; \xi)$$
$$= P\left(D_i \leq s + w | D_i > s, \mathcal{H}_i^J(s), Z_{s,ij}^R, Z_{s,i}^D, \xi\right)$$
$$= \frac{\int_0^\infty \left[S_i^D\left(s | Z_{s,i}^D, u_i, \xi\right) - S_i^D\left(s + w | Z_{s,i}^D, u_i, \xi\right)\right] (u_i)^J S_{i(J+1)}^R\left(s | Z_{s,ij}^R, u_i, \xi\right) g(u_i) \mathrm{d}u_i}{\int_0^\infty S_i^D\left(s | Z_{s,i}^D, u_i, \xi\right) (u_i)^J S_{i(J+1)}^R\left(s | Z_{s,ij}^R, u_i, \xi\right) g(u_i) \mathrm{d}u_i}$$

(3)

where $Z_{s,ij}^R$ and $Z_{s,i}^D$ are the values of the covariates at time s, and $g(u_i)$ is the density of the gamma distribution defined in equation (2).

The second setting P^{Ign} also uses the joint modelling framework. However, while recurrences information is used in model estimation, the information about previous recurrences is not considered in the prediction, and it can be missing. It is defined by:

$$P^{Ign}(s, s + w; \xi)$$
$$= P\left(D_i \leq s + w | D_i > s, Z_{s,i}^D, \xi\right)$$
$$= \frac{\int_0^\infty \left[S_i^D\left(s | Z_{s,i}^D, u_i, \xi\right) - S_i^D\left(s + w | Z_{s,i}^D, u_i, \xi\right)\right] g(u_i) \mathrm{d}u_i}{\int_0^\infty S_i^D\left(s | Z_{s,i}^D, u_i, \xi\right) g(u_i) \mathrm{d}u_i}$$

(4)

Both settings are dynamic in the sense that the prediction can be updated by changing the prediction time s, thus the quantity of available information, and/or the prediction window w. The first setting considers the individual relapse history, whereas the second ignores it.

Prediction of death using a Landmark Cox model

The Landmark approach involves fixing a prediction time s and fitting the model on the sub-group of patients still at risk of death at this time, that is, patients alive and not lost to follow-up [2]. Thus, the number of relapses occurring before time s can be treated as a baseline covariate, and the recurrences occurring after s are ignored. This covariate can be updated when another Landmark time s is considered, and a new model is fitted. With this approach, a robust model can be used, requiring few parameters, and the time-dependent effects are easily dealt with. However, only a sub-group of alive patients is included to fit the model, which may result in a loss of information in the parameter estimation.

We have D_i the death time and C_i the independent censoring time. Let $\lambda_{s,i}^D\left(.|Z_{s,i}^D\right)$ denote the hazard function of death conditional on being alive at the Landmark time s, $\lambda_{s,0}^D(.)$ be the conditional baseline hazard function, $Z_{s,i}^D$ be the covariate vector at time s and β_s be their effect estimated at time s. The Landmark Cox model is then written as follows:

$$\lambda_{s,i}^D\left(t | Z_{s,i}^D\right) = \lambda_{s,0}^D(t) \exp\left(\beta_s' Z_{s,i}^D\right), \text{ for } t \geq s \quad (5)$$

This model is estimated with the information available at the Landmark time s. The prognostic factors of interest $Z_{s,i}^D$ may include information about previous recurrent events, for example, their frequency and timing. The baseline hazard of death $\lambda_{s,0}^D(.)$ is estimated using splines with parameters η_s^D. The parameters of the model $\xi_s^{LM} = (\eta_s^D, \beta_s)$ are estimated using penalized maximum likelihood estimators as in [5]. The corresponding prediction of the risk of death is written as:

$$P^{LM}\left(s, s + w; \xi_s^{LM}\right)$$
$$= P\left(D_i \leq s + w | D_i > s, Z_{s,i}^D, \xi_s^{LM}\right)$$
$$= \frac{S_i^D\left(s | Z_{s,i}^D, \xi_s^{LM}\right) - S_i^D\left(s + w | Z_{s,i}^D, \xi_s^{LM}\right)}{S_i^D\left(s | Z_{s,i}^D, \xi_s^{LM}\right)}$$

(6)

where $S_i^D(.|Z_{s,i}^D, \xi_s^{LM})$ is the survival function conditional on being alive at time s associated to the hazard of death.

External validation of the prediction

In order to make predictions using the three proposed settings P^{LM}, P^{Rec} and P^{Ign} on new patients, the model parameters are estimated on the development sample. Based on these estimators, predictions for new patients are obtained by replacing the patient level information, $Z_{s,i}^D$ in equation (6) or J, $Z_{s,ij}^R$ and $Z_{s,i}^D$ in equations (3) and (4), with actual information on the new patient.

The quality of fit of the two models can be compared on the development data using an approximate likelihood cross-validation criterion as in [10].

Prediction error

To estimate if the predictions are accurate, error of prediction curves are used, based on the Brier score. The Brier score aims to measure how far the prediction is from the actual outcome of the patients. We used a weighted estimator of the Brier score to account for right censoring using the Inverse Probability of Censoring Weights (IPCW) [11].

Let N_s be the number of patients alive and uncensored at prediction time s, that is, patients for whom the prediction can be made. Given T_i^D the possibly right-censored survival time, δ_i^D the corresponding event indicator ($\delta_i^D = 1$ if the observed time is a death time, 0 otherwise). We denote $\widehat{G}_N(.)$ the Kaplan-Meier estimate of the censoring distribution on the sample. Using the generic term $\hat{P}(s, s + w; \hat{\xi})$ which can be one of the three prediction probability settings previously described, the error of prediction is defined by:

$$Err_{s,w} = \frac{1}{N_s} \sum_{i=1}^{N_s} \left[I\left(T_i^D > s + w\right) - \left(1 - \hat{P}\left(s, s + w; \hat{\xi}\right)\right)\right]^2$$
$$\times \hat{h}_i\left(s + w, \hat{G}_N(.)\right)$$

with $\hat{h}_i(s + w, \hat{G}_N(.))$ being a weight that accounts for right censoring:

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 4 of 12

$$\hat{h}_i\left(s+w, \hat{G}_N\left(.\right)\right) = \frac{I\left(T_i^D \le s+w\right)\delta_i^D}{\hat{G}_N\left(T_i^D\right)/\hat{G}_N\left(s\right)} + \frac{I\left(T_i^D > s+w\right)}{\hat{G}_N\left(s+w\right)/\hat{G}_N\left(s\right)}$$

The performance of the models is compared using a measure of explained residual variation defined as follows [12]:

$$R^2 = 1 - \frac{Err_{s,w}}{Err_{s,w;KM}}$$

where $Err_{s,w}$ is the error of one of the predictions (P^{LM}, P^{Rec} or P^{Ign}) described as above and $Err_{s,w;KM}$ is the error of prediction using the Kaplan-Meier estimate at $s+w$ in the entire set of patients. It can be interpreted as how much the prediction error is decreased using the model-based prediction as compared to an average prediction estimated by Kaplan-Meier.

The proposed error of prediction is calculated in two different ways: either s is fixed and w varies, or s varies while w is fixed.

Calibration plot

Another indicator of the accuracy of the prediction tool proposed is the calibration [13]. A well-calibrated prediction means that, among 100 patients with a predicted event risk of p%, p of them will actually experience the event. This can be computed only for a binary endpoint, meaning that we must choose a time of prediction. We set it at $s+w = 10$ years.

The calibration is illustrated using a calibration plot. The predicted risks of death are grouped according to the deciles of their distribution. For each decile, the observed proportion of an event is plotted against the mean predicted value, along with the 95% confidence interval for the observed proportion. For a well-calibrated prediction, all points should fall very close to the first bisector.

On the calibration plot the histogram of the predicted values is also represented, showing how they are distributed between 0 and 1.

Software

Joint model and predictions were computed using the R functions *frailtyPenal* and *prediction* from the R package *frailtypack* [14]. The Brier score were computed using the R package *pec* [15]. The calibration plots were drawn using R software (code available as Additional file 1).

Ethical approval and availability of supporting data

For the French hospital series, ethical approval from the national ethics committee (Commission Nationale de l'Informatique et des Libertés) was obtained for this study, which allowed the use of data recorded in this clinical and pathological database. In this comprehensive cancer center, each patient was informed that medical data can be use in observational research. The procedure follows the French law for medical research. We hold statutory and ethical approvals to analyse the data from West Midlands Cancer Intelligence Unit and Comprehensive Cancer Centre The Netherlands registry data.

All data used are confidential. Researchers may access the data by sending a formal request to the appropriate institution (Institut Bergonié for the French series, West Midlands Cancer Intelligence Unit for the UK registry data and Comprehensive Cancer Centre The Netherlands for the Dutch registry data).

Results

Population comparison

The first validation sample consisted of all breast cancer cases diagnosed in West Midlands, England, in 1996 and followed until 2012. The second validation sample consisted of cases from the Netherlands Cancer Registry, South Netherlands region excluded, diagnosed between 2003 and 2006 and followed until the end of 2012. Because the vital status of the patients is ascertained in both registries through a passive approach, no lost to follow-up is assumed. The development cohort consisted of 1067 patients operated in a comprehensive cancer centre between 1989 and 1993, and with a median follow-up of 14 years. Thus, the two validation populations differ from the development population in terms of country and selection of population (general population in West Midlands and Netherlands; hospital-based patients in France) and inclusion period (1996 and 2003-2006 vs. 1989-1993).

In the two validation samples, a high rate of missing data was observed. Out of the 3168 cases recorded in the year 1996 in West Midlands, 1196 (38%) had non-missing values for all of the five studied prognostic factors. In the Dutch sample, information about peritumoural vascular invasion was not recorded. Of the 41,676 recorded patients, 31,075 (75%) had non-missing values for the four remaining factors. In our validation sample, we included only patients with complete information in the two datasets. This decision is discussed in the last part of the paper.

Table 1 compares the repartition of the prognosis factors in the three samples, as well as the number of relapses per patient, and the overall survival. Patients in both validation samples had more severe disease, i.e., more peritumoural vascular involvement (38.5% in West Midlands vs. 26.7%), a tumour size greater than 20 mm (46.8% in West Midlands and 39.8% in Netherlands vs. 22.7%) and grade III disease (37.1% in West Midlands and 33.8% in Netherlands vs. 24.6%), despite a similar age. As a result, overall survival in both West Midlands and Netherlands was lower than in the development cohort.

The number of relapses per patient also varied. There were up to four relapses recorded in the West Midlands

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 5 of 12

Table 1 Description of the three samples used to develop (n=1067) and validate the model (n=3168 and n=31,075)

Variable	French cohort (1989-1993)		West Midlands (1996)		Netherlands (2003-2006)	
	N=1067	%	N=1196	%	N=31075	%
Age						
Age ≤40	82	7.7	73	6.1	2126	6.8
Age [40-55]	391	36.6	456	38.1	10681	34.4
Age >55	594	55.7	667	55.8	18268	58.8
Peritumoural vascular involvement	285	26.7	460	38.5	-	-
Tumour size > 20 mm	242	22.7	560	46.8	12365	39.8
Nodal involvement	451	42.3	496	41.5	12588	40.5
Grade						
Grade I	316	29.6	226	18.9	6565	21.1
Grade II	488	45.7	526	44.0	13993	45.0
Grade III	263	24.6	444	37.1	10517	33.8
Number of recurrent events						
None	705		895		27231	
1	301		240		3834	
2	57		49		10	
3	4		10		0	
4	0		2		0	
5-year survival	89.1	(87.3-91.0)	76.6	(74.2-79.0)	85.5	(85.1-85.9)
10-year survival	77.1	(74.6-79.7)	63.1	(60.5-65.9)	-	
15-year survival	65.4	(62.2-68.2)	51.6	(48.8-54.5)	-	

sample compared with a maximum of two in the Dutch sample and three in the development cohort. In the West Midlands registry, relapses were not collected but retrieved from the treatment information with an algorithm that uses the treatment type and time interval between successive treatments [16]. In the Dutch sample, relapse data was obtained directly from patient files; both clinically and pathologically confirmed relapses were recorded. The recording was limited to relapses occurring during the first five years after diagnosis and, in some regions, to the first relapse of each type (local relapse, regional relapse or distant metastasis). In the French cohort, relapses (loco-regional recurrence or distant metastasis) were recorded following a clinical examination. That resulted in 75% of the patients without registered relapse in the West Midlands, 88% in the Dutch sample, and 66% in the French cohort.

Validation of the prediction
Models
The results of the joint frailty model and the Landmark Cox model (thereafter called Landmark model), estimated on the French data, are shown in Table 2. The prognostic factors kept for prediction were those associated with the risk of recurrent events or with the risk of death in

the joint model. The joint model estimation (first and second columns) showed that the risk of recurrent event was higher in younger women (age ≤ 40), in women with peritumoural vascular involvement, in women with larger tumour (tumour size >20 mm), in women with nodal involvement, and in women with grade other than I. The risk of death was decreased for women between 40 and 55-year old, and increased in women with peritumoural vascular involvement, with large tumour size, with nodal involvement, and with grade other than I. The variance of the random effects θ differed significantly from zero, meaning that there exists an heterogeneity between patients regarding their risk of recurrence and death, and that this heterogeneity could not be totally explained by the measured prognostic factors. The $\alpha > 0$ showed that the patients at higher risk of relapses were also at higher risk of death. Results of the Landmark model are shown in the last column. The main difference is that only the effect of covariates on the risk of death are estimated, and the estimations were adjusted parametrically on the number of relapses. As a consequence, the Landmark model provided lower estimated effects than with the joint model. After adjustment on the number of relapses, only the presence of a nodal involvement was significantly associated with a higher risk of death, and both age lower than or

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 6 of 12

Table 2 Joint and Landmark Cox models estimations on the French cohort (n=1067 patients, 427 recurrent events)

	Joint model				Cox landmark	
	For recurrent events		For death		For death	
Variable	HR	(95% CI)	HR	(95% CI)	HR	(95% CI)
Age						
$[40-55]$ vs. > 55	1.17	(0.91-1.51)	0.31	(0.16-0.60)	0.56	(0.41-0.76)
≤ 40 vs. > 55	2.41	(1.73-3.37)	1.57	(0.73-3.38)	0.54	(0.31-0.92)
Peritumoural vascular involvement	1.61	(1.26-2.06)	4.74	(2.54-8.85)	1.04	(0.76-1.43)
Tumour size (> 20 mm vs. $\leq$ 20 mm)	1.95	(1.52-2.50)	6.21	(2.99-12.86)	1.20	(0.88-1.65)
Nodal involvement	1.84	(1.44-2.36)	4.89	(2.47-9.67)	1.95	(1.45-2.60)
Grade						
II vs. I	2.18	(1.57-3.01)	7.48	(2.71-20.66)	1.07	(0.75-1.52)
III vs. I	3.09	(2.16-4.41)	44.33	(15.61-125.93)	1.25	(0.83-1.88)
Recurrences before $t = 5$ years						
One previous recurrence					7.18	(5.25-9.83)
Two previous recurrences					6.94	(3.05-15.83)
$\theta = var(u_i)$	1.07	(se=0.06)				
α	4.45	(se=0.33)				
LCV	1.19				0.93	

HR: Hazard ratio; CI: Confidence interval; LCV: Likelihood cross-validation criterion; Cox Landmark at time t = 5 years.

equal to 40, and age between 40 and 55 years were associated with a lower risk of death. The Landmark model also showed an important effect of the number of previous relapses. The likelihood cross-validation criterion was lower for the Landmark model (0.93 *versus* 1.19), suggesting that this model fitted the data better than the joint model. However, a better fit does not necessarily result in better prediction [17].
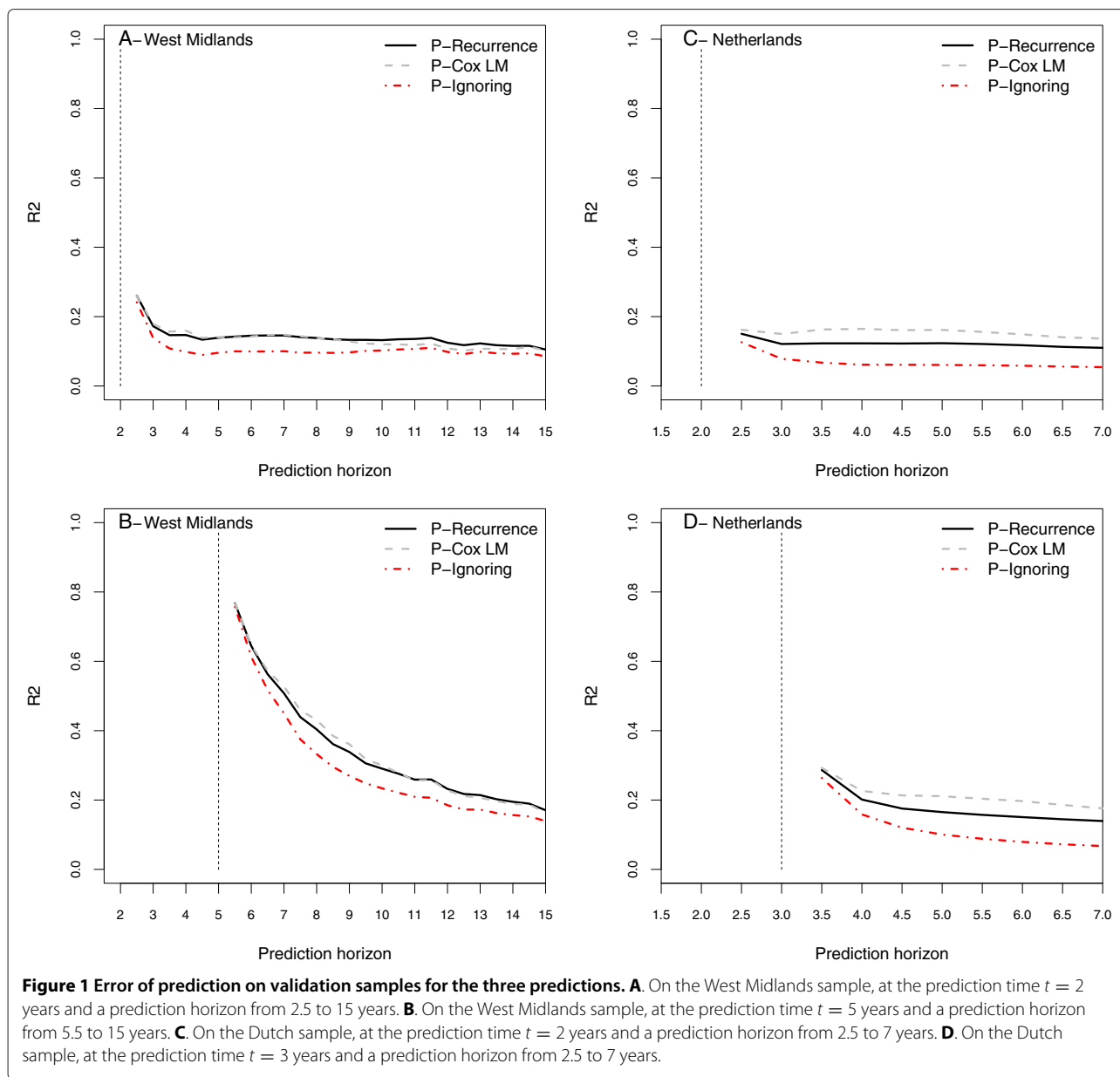
Prediction error

Overall, all of the studied prediction settings gave better results than the Kaplan-Meier, with a higher R^2 for both predictions accounting for relapses (Figure 1). In the West Midlands, when the time of prediction s is at five years (Figure 1B), R^2 was as high as 80% for early predictions and regularly decreased with increasing prediction horizon (30% at 10 years). The gain in the prediction error diminished with the prediction horizon, being around 50% at seven years and ending around 17% at 15 years, showing that the information from the model had a higher impact on short-term prediction. R^2 was still around 20% at a horizon of 15 years, but very similar for the three settings. R^2 was low, under 20%, for shorter prediction time ($s = 2$ years; see Figure 1A). This illustrates the fact that the information gathered up to two years was not enough to obtain good prediction, especially considering relapses. In the Dutch sample (Figures 1C and D), the limited follow-up prevented us from studying prediction times longer than three years. The results were very similar to those in the West Midlands sample at $s = 2$ years,

but the difference between the three settings was larger. Results were better for the prediction from the Landmark model, and worse for the prediction ignoring the relapse information.

When holding the prediction window at $w = 2$ or $w = 5$ years, results were similar (Figures 2A and B, respectively). The setting ignoring relapses always gave lower performance, while the performances of the two other settings were very similar. The more information that was collected and used, the more accurate the prediction was, as shown by the curves increasing with time of prediction t, for both window times. As expected, the entire curve was higher (i.e., lower error of prediction) when the prediction was made for a shorter window (two years as compared to five years).

Calibration

All three settings gave good calibration, with points around the first bisector (Figure 3). Interestingly, both prediction approaches accounting for relapses identified a group of patients with high risk of death in both samples. For these high-risk patients, the mean predicted risk was somewhat lower than the observed risk (40% in the West Midlands vs. 50% in the Netherlands using P^{Rec}). The histograms show that predicted values were lower overall for the Landmark approach (rarely exceeding 20%) whereas both predictions from the joint model gave higher risks. This may explain why the observed probability of death seemed underestimated with the Landmark approach in the West Midlands.

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 7 of 12

Figure 1 Error of prediction on validation samples for the three predictions. A. On the West Midlands sample, at the prediction time $t = 2$ years and a prediction horizon from 2.5 to 15 years. **B**. On the West Midlands sample, at the prediction time $t = 5$ years and a prediction horizon from 5.5 to 15 years. **C**. On the Dutch sample, at the prediction time $t = 2$ years and a prediction horizon from 2.5 to 7 years. **D**. On the Dutch sample, at the prediction time $t = 3$ years and a prediction horizon from 2.5 to 7 years.

Additional validation by subgroups

The validation samples differed in many aspects from the development sample. Thus, even if good results were observed for the proposed prediction both in terms of prediction error and of calibration, it is crucial to check the accuracy of the prediction on a more similar sample. For these subgroups analyses, the model is still developed on the whole French cohort; only the prediction is made on a subgroup of the English sample. Here we selected a subsample of operated patients, as it was the main selection criterion in the development sample.

Similar results were observed with the 602 operated patients from the West Midlands (Figure 4). Large

confidence intervals were observed due to the reduced number of subjects included in this analysis (n=417 patients alive at five years). Calibration was not better than as observed on the entire West Midlands data.

A second subgroup analysis was performed comparing the performance of the proposed predictions between subjects who relapsed at least once before the prediction time of five years and those who did not (Figure 5). As expected, in the subgroup of patients without relapse, the results were very similar to those in the entire sample. However, no high-risk subject groups were identified, as observed in the main analysis. In the subsample with relapses, the prediction ignoring the relapses underestimated the observed probabilities of events (all the points
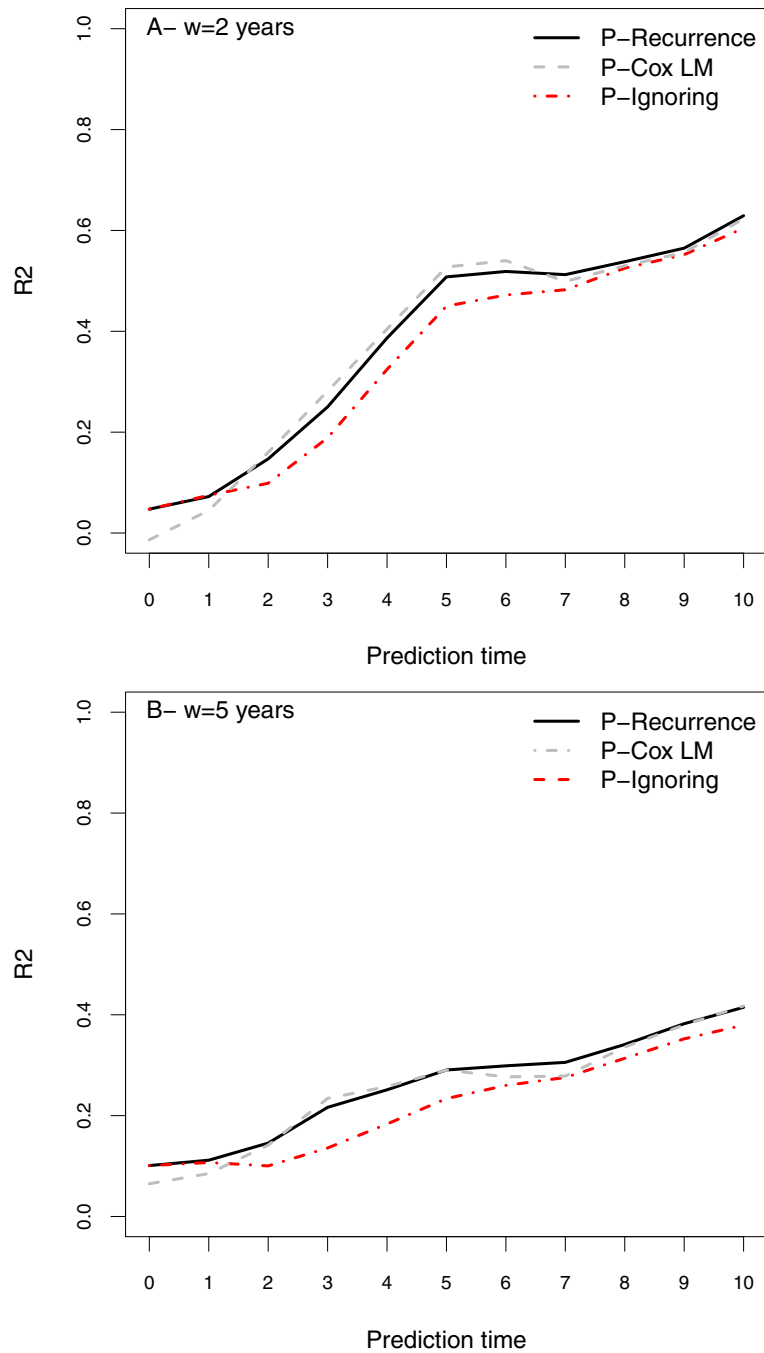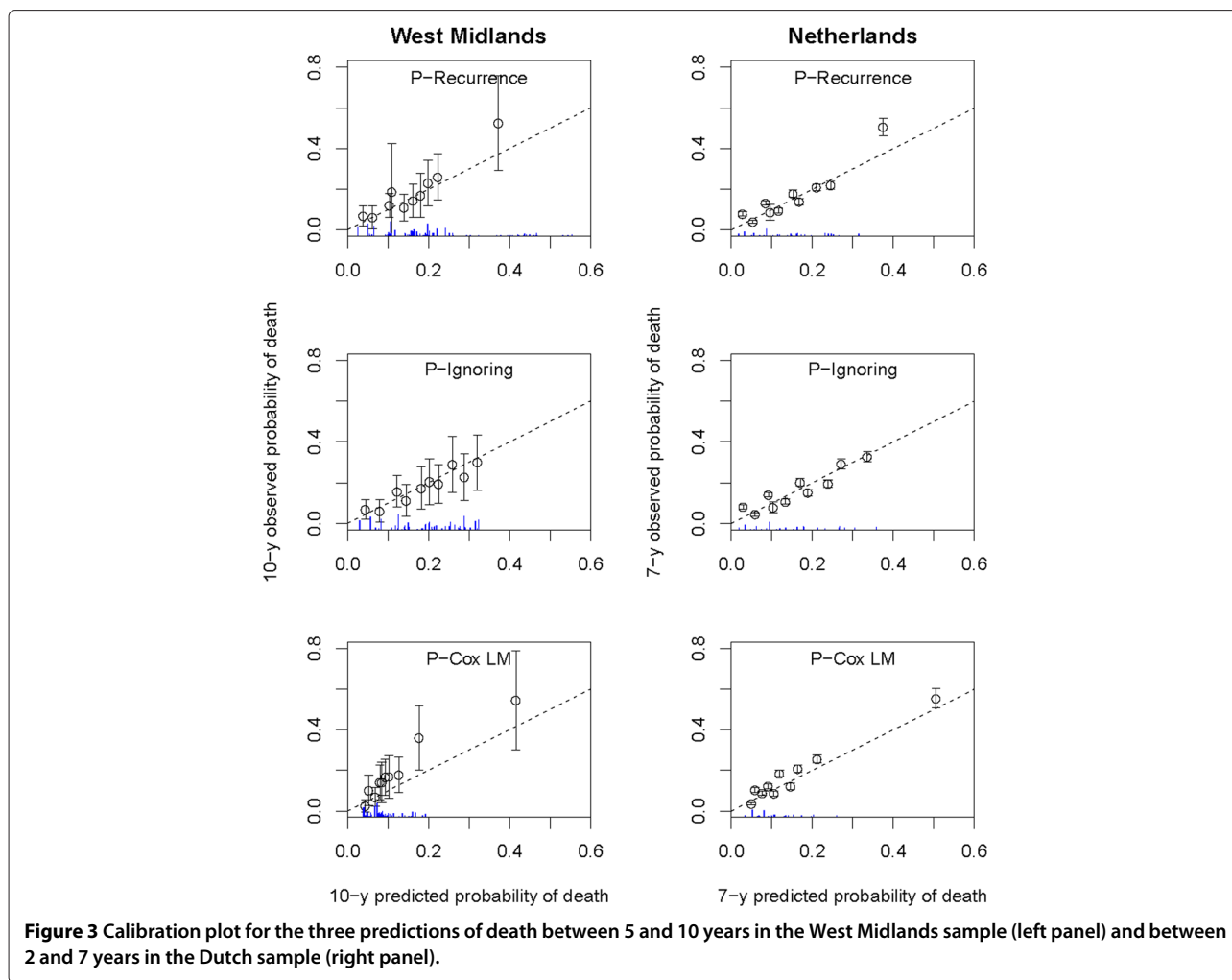
Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 8 of 12

Figure 2 Error of prediction on the West Midlands sample when the prediction time *t* is increasing from 0 to 10 and the window of prediction is set at A. 2 years and **B**. 5 years.

are above the line) and had a very low R^2, negative after 7.5 years.

Discussion

To account for the recurrent events and their association with the risk of death, we used a joint frailty model for recurrent events and a terminal event. This framework appears suitable to derive such predictions. The obtained prediction gave a similar error of prediction to a Cox model in a Landmark approach. The two approaches have modelling assumptions. Both are proportional hazards models. In both models, we estimated the baseline hazards with a semi-parametric approach using splines. It was compared to a non-parametric estimation of the baseline

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 9 of 12

Figure 3 Calibration plot for the three predictions of death between 5 and 10 years in the West Midlands sample (left panel) and between 2 and 7 years in the Dutch sample (right panel).

hazard, and both approaches gave very similar results (data not shown). The landmarking approach allows the use of a simple and robust model, such as the Cox model. Moreover the covariates effects are re-estimated at each prediction times. However, only the alive patients are kept to estimate the model. By contrast, the joint model use all the sample patients and parameters are estimated once to do predictions. However, assumptions are made on the frailty structure and distribution. We can note that in the hypothesis of no association between the relapse and death processes (i.e., parameter $\alpha = 0$), the prediction of death in the joint model simplifies to an expression similar to equation (6), independent of the risk of relapses. Waiting for a more careful comparison of the two approaches, the choice between the joint and the Landmark approach should thus be mainly guided by the willingness (i) to fully describe both processes (recurrent event and death) and their correlation (joint model); or (ii) to focus only on the death (Landmark approach). A Landmark approach using a non-parametric prediction was also recently proposed

to predict a long-term outcome accounting for a short-term event [18].

Initially, the proposed prediction incorporated information about the human epidermal growth factor receptor 2 (HER-2) status and hormonal receptor status. However, considering the non-availability of this information at the general population level at the time of this data collection, we have re-estimated the model and prediction without this information. On the initial development sample, we compared the prediction error of the two joint models, with and without this biological information. The prediction error was very similar for both models (data not shown).

One perspective of such prediction can be its use in clinical trials context. Indeed, the validation of surrogate markers in several cancers, such as the progression-free survival as a surrogate for the overall survival, raises the question about how to use the progression-free survival information in practice. One of the options is to use progressions to predict the risk of death in the two arms to
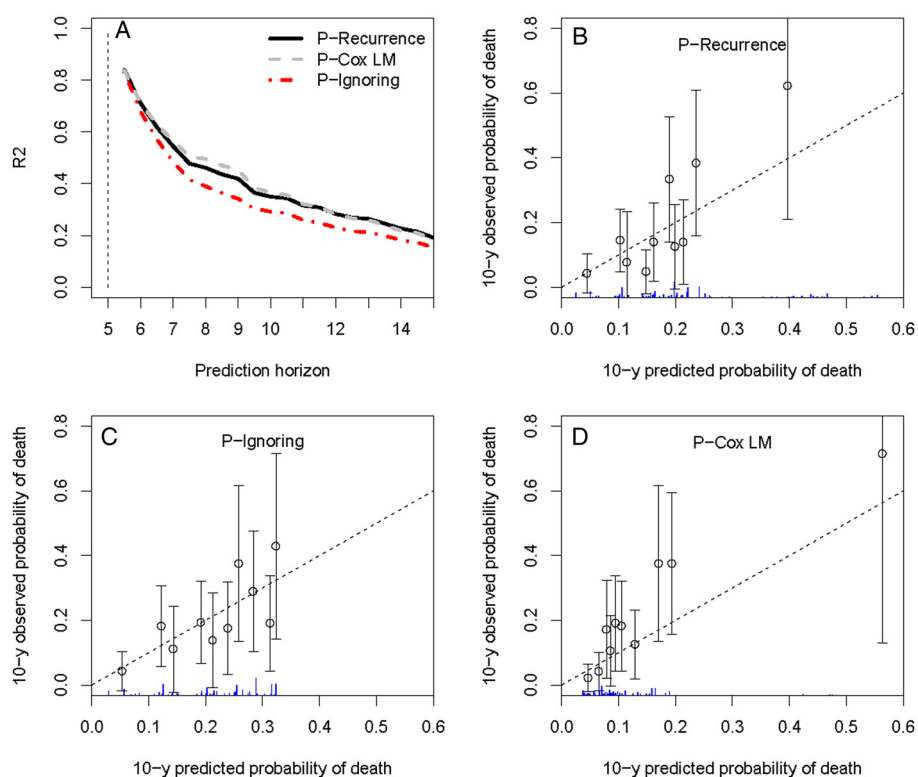
Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 10 of 12

Figure 4 Results of the prediction on the operated patients from West Midlands. A. Relative prediction error at the prediction time $t = 5$ years and a prediction horizon from 5.5 to 15 years. **B**,**C**,**D**. Calibration plots for the three predictions.

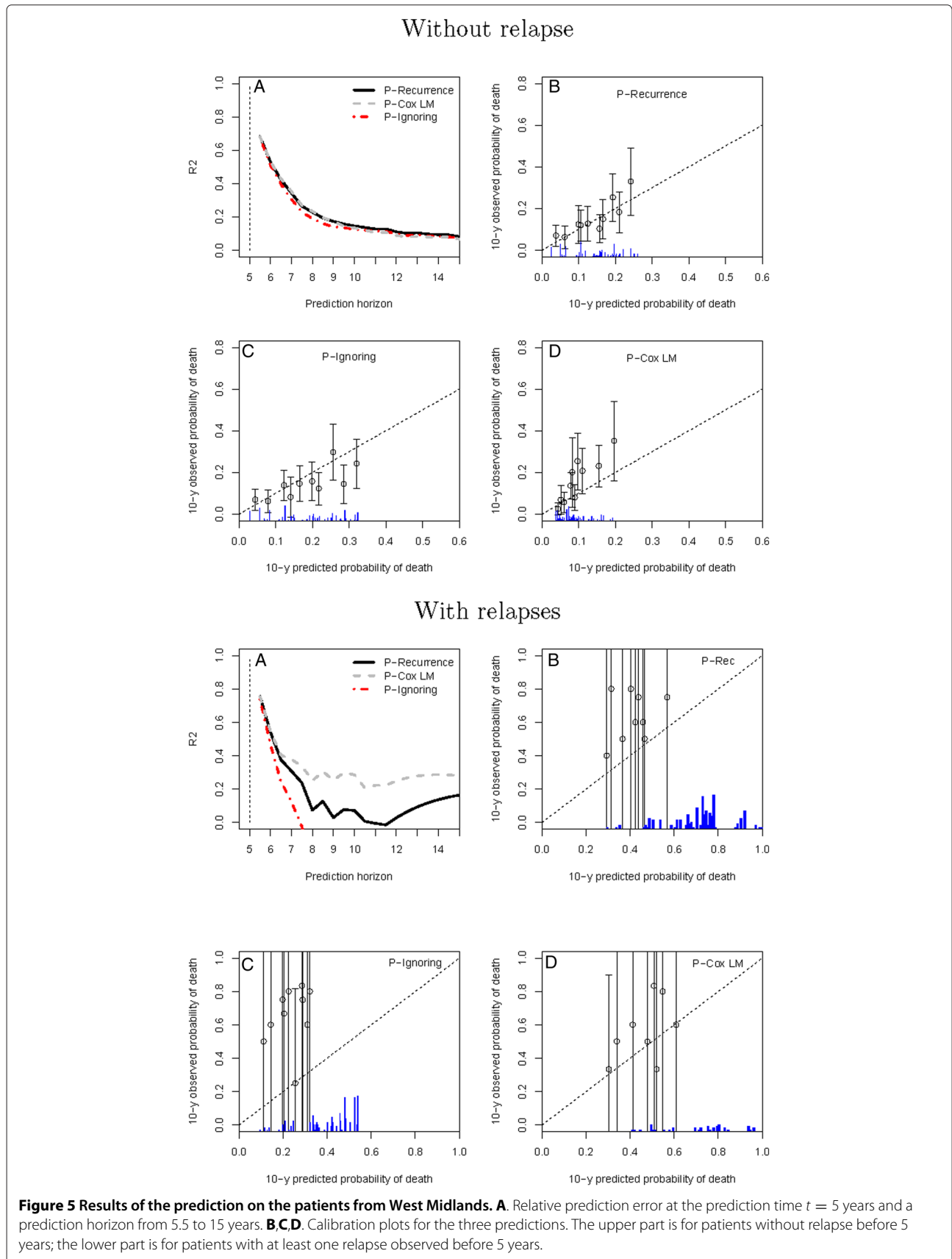be able to conclude earlier on the treatment effect [19]. In that perspective, the prediction that we validated maybe useful.

In the context of the joint modelling framework, the prediction of recurrent events can also be derived. Moreover, our analyses indicate that each prognostic factor considered separately adds very little prediction information once the baseline hazard and recurrent event processes are adequately modelled (data not shown). The covariates may be of greater interest when predicting the risk of a recurrent event. Considering relapse type differently (loco-regional relapse and metastasis) can also be of interest as they reflect various severity levels of the disease [20]. Finally, these predictions could be extended in the context of competing risks or excess mortality, where it would be possible to focus on predicting only the risk of death from cancer.

Finally, prognostic research literature is lacking when it comes to the consequences of missing data on the validation process, i.e., after the development phase. No prediction can be done if one of the covariates is missing. However, the impact of such exclusion on the validation process remains unclear. Multiple imputation has proved to be a useful approach for model estimation (e.g., [21]), and could also be used for the validation stage. However,

the benefit of such imputation to estimate model performance is uncertain. To reproduce the conditions of clinical practice, we keep in our validation only the patients with complete information. The subjects with missing data were more likely to be 55 years and older, and to have more nodal involvement, for similar stage and tumour size (data not shown). As the predictions that take into account relapse information were demonstrated to be more appropriate in predicting high risk of death, it is possible that the performances of the prediction accounting for the relapses were underestimated. In the end, the survival results of the analysed patients were in accordance with the results of the EUROCARE-4 study [22]. In this study, the age-adjusted 5-year survival was 81% for French patients diagnosed between 1990 and 1994, and 78% and 83% for the patients diagnosed between 1995 and 1999 in England and Netherlands, respectively.

Conclusion
The present work shows how recurrent events occurring in breast cancer patients may be used to obtain accurate prediction of death. The resulting calibration and error of prediction show that the estimated prognostic model is useful to predict the risk of death, in particular when enough variability in the number of recurrences

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 11 of 12

Figure 5 Results of the prediction on the patients from West Midlands. A. Relative prediction error at the prediction time $t = 5$ years and a prediction horizon from 5.5 to 15 years. **B**,**C**,**D**. Calibration plots for the three predictions. The upper part is for patients without relapse before 5 years; the lower part is for patients with at least one relapse observed before 5 years.

Mauguen *et al. BMC Medical Research Methodology* (2015) 15:27

Page 12 of 12

is observed. Good calibration was obtained, especially considering that the validation samples differed from the development sample with respect to inclusion criteria for the patients and period, country and therefore, health care system. Using different incidence years is of great interest, since the care (especially treatment and screening) of breast cancer patients evolved during the 1990s, affecting survival. It seems that finally, despite these differences, the effect of covariates and relapses remained similar and was still of interest. Therefore, the joint model for recurrent event and a terminal event gives some accurate predictions, and the specific studied model in breast cancer can be used in different populations.

Additional file

Additional file 1: R code for drawing calibration plot.

Author details
[1]Biostatistic unit, INSERM U897, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. [2]Cancer Research UK Cancer Survival Group, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT London, UK. [3]Clinical epidemiology and research, Institut Bergonié, 229 Cours de l'Argonne, 33000 Bordeaux, France. [4]INSERM CIC-EC7, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. [5]West Midlands Cancer Intelligence Unit, 5, St Philip's Place, B3 2PW Birmingham, UK. [6]Comprehensive Cancer Centre The Netherlands (IKNL), Godebaldkwartier 419 ingang Janssoenborch, 3511 Utrecht, The Netherlands.

References

1. Kalbfleisch J, Prentice R. The statistical analysis of failure time data: Hoboken: John Wiley; 2002.
2. Van Houwelingen H. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34(1):70–85. doi:10.1111/j.1467-9469.2006.00529.x.
3. Van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis, 1st edn. Boca Raton: CRC Press Inc; 2011.
4. Liu L, Wolfe R, Huang X. Shared frailty models for recurrent events and a terminal event. Biometrics. 2004;60(3):747–56. doi:10.1111/j.0006-341X.2004.00225.x.
5. Rondeau V, Mathoulin-Pélissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. Biostatistics. 2007;8(4):708–21. doi:10.1093/biostatistics/kxl043.
6. Mauguen A, Rachet B, Mathoulin-Pélissier S, MacGrogan G, Laurent A, Rondeau V. Dynamic prediction of risk of death using history of cancer recurrences in joint frailty models. Stat Med. 2013;32(30):5366–80. doi:10.1002/sim.5980.
7. Altman D, Royston P. What do we mean by validating a prognostic model?. Stat Med. 2000;19(4):453–73.
8. Konig I, Malley J, Weimar C, Diener H, Ziegler A. Practical experiences on the necessity of external validation. Stat Med. 2007;26(30):5499–511. doi:10.1002/sim.3069.
9. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683–90. doi:10.1136/heartjnl-2011-301246.
10. Rondeau V, Pignon J-P, Michiels S. A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. Stat Methods Med Res. 2011;897:1–19. doi:10.1177/0962280211425578.
11. Gerds T, Schumacher M. Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical J. 2006;48(6):1029–40. doi:10.1002/bimj.200610301.
12. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18(17-18):2529–2545. doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.
13. Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2010 edn. New York: Springer; 2010.
14. Rondeau V, Mazroui Y, Gonzalez JR. frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. J Stat Software. 2012;47(4):1–28.
15. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. J Stat Software. 2012;50(11):1–23.
16. Walton J, Lagord C, Lawrence C, Lawrence G. The development of an algorithm to identify breast cancer recurrences. Brighton. NCIN Cancer Outcomes Conference 2013. Brighton, June 12th-14th, 2013. 2013. [Poster 121, url: http://www.ncin.org.uk/news_and_events/conferences/2013_posters, consulted on December 2013].
17. Proust-Lima C, Sene M, Taylor JM, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. Stat Methods Med Res. 2014;23(1):74–90. doi:10.1177/0962280212445839.
18. Parast L, Cheng S, Cai T. Incorporating short-term outcome information to predict long-term survival with discrete markers. Biometrical J. 2011;53(2):294–307.
19. Faucett CL, Schenker N, Taylor JM. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. Biometrics. 2002;58(1):37–47.
20. Mazroui Y, Mathoulin-Pélissier S, MacGrogan G, Brouste V, Rondeau V. Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data. Biometrical J. 2013;55(6):866–84. doi:10.1002/bimj.201200196.
21. Vergouwe Y, Royston P, Moons K, Altman D. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63(2):205–14. doi:10.1016/j.jclinepi.2009.03.017.
22. Berrino F, De Angelis R, Sant M, Rosso S, Lasota M, Coebergh J, et al. Survival for eight major cancers and all cancers combined for european adults diagnosed in 1995-99: results of the EUROCARE-4 study. Lancet Oncol. 2007;8(9):773–83. doi:10.1016/S1470-2045(07)70245-0.