Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: performance of this multidimensional penalized spline approach in net survival trend analysis.

Short title: Flexible survival model for trend analyses

Laurent Remontet[1-3], Zoé Uhry [4,1-3], Nadine Bossard[1-3], Jean Iwaz[1-3], Aurélien Belot[5], Coraline Danieli[6], Hadrien Charvat[7], Laurent Roche[1-3] and the CENSUR Working Survival Group.

[1] Hospices Civils de Lyon, Service de Biostatistique-Bioinformatique, Lyon, France.

[2] CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France.

[3] Université Lyon 1, Villeurbanne, France.

[4] Département des Maladies Non-Transmissibles et des Traumatismes, Santé Publique France, Saint-Maurice, France.

[5] Cancer Research UK Cancer Survival Group, Faculty of Epidemiology and Population Health, Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom.

[6] McGill University Health Center, Department of Epidemiology, Biostatistics and Occupational Health, Montreal, QC, Canada.

[7] Division of Prevention, Center for Public Health Sciences, National Cancer Center, Chuo-ku, Tokyo, Japan.

Corresponding author:

Laurent Remontet

Service de Biostatistique-Bioinformatique - Centre Hospitalier Lyon-Sud – Bât. 4D

F-69495 Pierre-Bénite Cedex, France

E-mail: laurent.remontet@chu-lyon.fr

Abstract

Cancer survival trend analyses are essential to describe accurately the way medical practices impact patients' survival according to the year of diagnosis. To this end, survival models should be able to account simultaneously for non-linear and non-proportional effects and for complex interactions between continuous variables. However, in the statistical literature, there is no consensus yet on how to build such models that should be flexible but still provide smooth estimates of survival. In this article, we tackle this challenge by smoothing the complex hypersurface (time since diagnosis, age at diagnosis, year of diagnosis, mortality hazard) using a multidimensional penalized spline built from the tensor product of the marginal bases of time, age, and year. Considering this penalized survival model as a Poisson model, we assess the performance of this approach in estimating the net survival with a comprehensive simulation study that reflects simple and complex realistic survival trends. The bias was generally small and the root mean squared error was good and often similar to that of the true model that generated the data. This parametric approach offers many advantages and interesting prospects (such as forecasting) that make it an attractive and efficient tool for survival trend analyses.

Keywords: penalized spline, survival model, tensor product, varying coefficient model, generalized additive model, cancer net survival trends, multidimensional smoothing, interaction, non-linear effect, non-proportional effect

1. Introduction

1.1. Epidemiological issues and modeling problems

In cancer descriptive epidemiology, one major indicator is the trend in patient survival according to the year of cancer diagnosis (yod) and the trend in the corresponding mortality hazard. Indeed, these trends show the way advances in medical practices (screening campaigns, diagnostic techniques, treatment options, etc.) have changed patient survival over the yod. Within this context, the age at diagnosis is a major variable because these practices depend strongly on age; actually, elderly cancer patients present frequently comorbidities that may prevent the use of aggressive, though efficient, treatment.[1] Moreover, describing trends according to the time elapsed since diagnosis helps a medical interpretation of the analysis results because that course of time corresponds to different steps in the disease and treatment outcomes (post-surgical mortality during early follow-up, outcome of the first-line treatment during the first year after diagnosis, late relapses, etc.)

Hence, a survival model for trend analysis should model the mortality hazard *h* as function of the age at cancer diagnosis, the yod, and the time since cancer diagnosis and answer at least three questions (assuming an improvement over the yod in survival for example): i) Did mortality decrease gradually over the yod or was the decrease observed only over a few yod? Or, in statistical terms, was the effect of the yod on *h* linear or non-linear? ii)

Was the decrease observed whatever the time elapsed since diagnosis or only at specific moments (such as at early follow-up because of better post-surgical management)? Or, was the effect of the yod on *h* proportional or non-proportional? iii) Was the decrease dependent on patient age at diagnosis? Or, was there an interaction between age and yod? These aspects (non-linearity, non-proportionality, and interaction) are very often met in real data; for example, in the French cancer survival population-based data,[2] the effect of age at diagnosis was almost systematically found to be non-linear and non-proportional whatever the cancer site. Furthermore, in a study of survival trends in six European countries and 15 cancers (90 analyses), Uhry et al.[3] found that the effect of the yod was non-proportional in 70% of the analyses and that it depended on age at diagnosis in two thirds of the analyses. Thus, one key issue in modeling survival trends is to build a flexible model able to reflect simultaneously these three fundamental aspects while providing smooth estimates.

1.2. What has been used so far for studying net survival trends?

Up to now, few attempts have been made to build such a flexible model. Indeed, in international trend studies[4, 5], net survival (NS), the main survival indicator used in the context of cancer descriptive epidemiology, is almost exclusively estimated separately for each period without modeling and using non-parametric estimators of NS.[6, 7] Such stratified analyses have well-known limitations: arbitrary choices of period- and age-

strata, loss of information due to categorization of continuous covariates, imprecision due to multiple stratification, possible inconsistencies in NS trends across age strata and considerable difficulties in studying covariate interactions. In addition, such analyses only provide NS estimates and not a description of the excess mortality hazard (h_E) that constitutes an additional and essential clinical piece of information.

Despite the significant progress in flexible parametric modeling,[8-13] few studies analyzed trends opting for a modeling approach and keeping time, age, and yod as continuous covariates[14, 15]; however in these studies, age-yod or age-yod-time interactions were not considered.

To our present knowledge, only one study has proposed a modeling approach that allows for a potentially complex effect of the yod.[3] However, in this study, defining and selecting the appropriate models were quite challenging and achieving a balance between flexibility and parsimony required building nineteen models that differed in modeling the effect of the yod in terms of linearity, proportionality and interaction with age; the final model was chosen according to the Akaike Information Criterion. This study highlights the difficulty of achieving a flexible modeling of $h_E(t,a,y)$ through a classical model-building strategy (guiding principles were proposed without reaching consensus[11]); the number of candidate models may be very large, which requires sound choices for model specification, choices that become more difficult as the study-period or follow-up lengthens. Moreover, variance is under-estimated if the selection process is not accounted

for and there is no simple solution for taking this phenomenon into account in the statistical inference; a correct variance estimation requires heavy bootstrap techniques or to consider multi-model inference.[16]

1.3. A flexible modeling for survival trend analyses: the MPS approach

To tackle the challenge of modeling of $h_E(t,a,y)$ in a flexible and convenient way, we propose to consider the issue as a problem of modeling a complex hypersurface $h_E(t,a,y)$ and to smooth this surface using a multidimensional penalized spline (MPS). The MPS approach is a powerful tool originally developed for Generalized Linear Models.[17] We adapted it to the survival context. One of the major benefits of this solution is that it reduces the model-building issue evoked above. The objective of the present paper is to evaluate the performance of this adapted MPS approach for usual studies of trends in net survival and excess mortality hazards, using realistic simulations. The proposed approach focuses herein on NS but is obviously suitable for overall survival too.

The present article is organized as follows: after a brief review of the NS concept, section 2 presents the proposed approach, highlighting the relationship between the MPS and the varying-coefficient model.[18] A comprehensive simulation study is carried out based on real data to assess the performance of this approach regarding its ability to fit various NS trends; section 3 presents the design, the theoretical parameters, and the indicators of

performance. Section 4 presents the results of these simulations and section 5 a general discussion. In the online supplementary material, we present an analysis of real data (from the French cancer registries) with the R-code necessary to reproduce this analysis (this code is available on the GitHub repository https://github.com/RocheLHCL/SMMR_Remontet2018).

2. Multidimensional penalized splines for a (net) survival model

2.1. Introduction to the concepts of excess mortality hazard model and net survival

In the competing-risk context of cancer survival, individuals may die from cancer or from another cause but, in cancer registries, the causes of death are not always available or reliable. In addition, cancer treatments may have long-term toxicities and ultimately cause death; these extra-deaths are then "due to cancer". These two reasons make "excess mortality" a relevant concept. This excess mortality can be estimated by supposing that, in cancer patients, the mortality due to others causes than the cancer can be obtained from the (all causes) mortality of the general population; the latter is referred to as the "expected mortality" h_P. Then, the mortality observed in cancer patients (h_O) may be written as:

$$h_O(t, \boldsymbol{x}) = h_E(t, \boldsymbol{x}, \boldsymbol{\beta}) + h_P(a + t, \boldsymbol{z}) \quad (1)$$

In this equation, h_E is the *excess mortality hazard* due to cancer, t is the time elapsed since cancer diagnosis, a is the age at cancer diagnosis, h_P is the mortality of the general

population at age $a+t$ given demographic characteristics $\mathbf{z}$ (h_P is considered known and available from national statistics), $\mathbf{x}$ is a vector of variables that may have an effect on h_E, and $\boldsymbol{\beta}$ is the vector of parameters of interest to be estimated.

Fully parametric models[8-10] have been proposed to model h_E e.g., $\log[h_E(t,x,\beta)] = f(t) + g(x) + h(t)x$ where f, g, and h are flexible functions such as cubic splines.

Let us consider an observation t_i, δ_i, x_i, and z_i of subject i, with $\delta_i=1$ when t_i corresponds to the time of death and $\delta_i=0$ when t_i corresponds to a censored observation, the contribution of that observation to the log-likelihood may be written (up to a constant):

$$l_i(\beta) = -\int_0^{t_i} h_E\left(u, x_i, \beta\right) du + \delta_i log[h_E(t_i, x_i, \beta) + h_P(a_i + t_i, z_i)]$$

In a non-penalized framework, the maximum likelihood method may be used to estimate parameters β of the excess hazard model.[8, 10, 19] However, specific numerical techniques are necessary to approximate the integral involved in the likelihood. In 2007, Remontet et al.[10] showed that using the 'point-milieu' approximation for the integral leads to a likelihood similar to the one obtained with a Poisson model on split data (a model that uses a modified link function so as to incorporate the expected mortality rates). Taking advantage of this similarity, a survival model can then be fitted in a numerically practical manner by using a Poisson regression; herein, this approach will be referred to as "Poisson approach".

Finally, once parameters β are estimated, NS, the survival that would be observed if cancer was the only cause of death, can be directly obtained from h_E using the classical relationship between hazard and survival: $NS(t, x_i) = exp\left[-\int_0^t h_E(u, x_{i,}\beta)\, du\right]$

2.2. Introduction to the multidimensional penalized spline approach

The general principle of penalized splines consists in modeling the parameter of interest (here, h_E) as a function of a vector of variables (here, time since diagnosis, age at diagnosis, and yod) using highly flexible functions. This flexibility is typically obtained with splines with a number of knots higher than what is deemed necessary (this leads to a high number of parameters). In the classical unpenalized likelihood framework, such flexibility leads to a high variability of the estimators and to an overfitting. In the penalized spline framework, these drawbacks are overcome by considering a penalized likelihood as the objective function obtained by adding to the classical likelihood a term that penalizes "wiggly" functions. One common choice among the penalization terms is the integral of the squared second derivative of the fitted function: this choice penalizes the functions that are too wiggly, achieves smoothness, and prevents from erratic estimation. The trade-off between model fit and model smoothness is controlled by a smoothing parameter λ.

The very clear and instructional article by Wood[20] presents the essentials for understanding and using penalized splines; specifically, building the penalization term, optimizing the penalized likelihood (with estimation of λ), and making statistical inference with examples that use *mgcv* package[17] in R (see also Marra and Radice[21] and Eilers and Marx[22]). Another useful and instructional reference is a more general overview by Ruppert et al.[23] that presents a mixed-model representation of penalized splines and Bayesian models with longitudinal and spatial effects.

The multidimensional version of penalized splines, based on tensor product of basis functions, has been already proposed in Generalized Linear Models by Wood[24] and by Marx and Eilers.[25, 26] An interesting example of the use of these MPSs in a Poisson model was given by Ugarte et al.[27] who modeled the number of deaths from prostate cancer as a function of the year of death and the geographic coordinates (longitude and latitude) of the residential area (see also Etxeberria et al.[28]). The tensor product of these three variables constitutes a spatio-temporal model and, together with Currie et al.[29], Ugarte proposes to use it as a projection tool.

2.3 Multidimensional penalized splines in (net) survival models

2.3.1. Modelling the mortality hazard with a varying coefficient model

To model log [$h_E(t,a,y)$] as a function of time since diagnosis t, age at diagnosis a, and yod y, we propose using a MPS whose basis is built by the tensor product of three marginal bases chosen for t, a, and y. To motivate this choice, we show its relationship with the varying coefficient model.[18]

First, let $(m_i(t))_{1\leq i\leq I}$, $(q_j(a))_{1\leq j\leq J}$, and $(b_k(y))_{1\leq k\leq K}$ be three low-rank bases for smooth functions f_t, f_a, and f_y, respectively:

$$f_t(t) = \sum_{i=1}^{I} \mu_i m_i(t);\ \ f_a(a) = \sum_{j=1}^{J} \theta_j\, q_j(a);\ f_y(y) = \sum_{k=1}^{K} \beta_k b_k(y)$$

where μ_i, θ_j, and β_k are the parameters to estimate.

For a clear illustration, let us take a very simple (though unrealistic) example, starting from a basic model in which the dynamics of the hazard according to time is log-linear:

$$log[h_E(t)] = f_t(t) = \mu_1 + \mu_2 t$$

We now want to take age into account in this model, knowing that this dynamics may vary with age. One way of achieving this is to allow the intercept and slope of $f_t(t)$ to change with age using another basis for age, say a quadratic polynomial $f_a(a)=\theta_1+\theta_2 a+\theta_3 a^2$. This gives the following model:

$$log[h_E(t,a)] = f_{ta}(t,a) = (\theta_{11} + \theta_{12}a + \theta_{13}a^2) + (\theta_{21} + \theta_{22}a + \theta_{23}a^2)t$$

In this model, for a given age, log [$h_E(t|a)$] is linear in time and for a given time, log [$h_E(a|t)$] is quadratic in age. This six-parameter model can be seen as a *varying coefficient model* where the coefficients of time (intercept $\mu_1=\theta_{11}+\theta_{12}a+\theta_{13}a^2$ and slope $\mu_2=\theta_{21}+\theta_{22}a+\theta_{23}a^2$) are allowed to change smoothly with age, the "effect modifier".[18] This change occurs in a structured fashion, in the sense that each age has its own intercept and its own slope but two adjacent ages have close intercepts and close slopes. Here, we may assume that, symmetrically, time changes the effect of age: the model then obtained will be the same.

Going back to the general and the most realistic case, the multidimensional function f_{ta} will correspond to:

$$log[h_E(t,a)] = f_{ta}(t,a) = \sum_{i=1}^{I}\sum_{j=1}^{J}\theta_{ij}q_j(a)m_i(t)$$

The construction of the multidimensional function may continue according to the same principle but with changes made now to coefficients θ_{ij} according to y. This leads to:

$$log[h_E(t,a,y)] = f_{tay}(t,a,y) =$$
$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\beta_{ijk}b_k(y)q_j(a)m_i(t)$$

(2)

that is, the multidimensional basis consists of the K×J×I terms $b_k(y)\times q_j(a)\times m_i(t)$ obtained by the product of the terms of the marginal basis. This basis construction is rather simple

and may be extended to any number of variables. It is essential to see from (2) that this model allows simultaneously for non-linearity and various interaction patterns. In particular, i) $q_j(a)\times m_i(t)$ and $b_k(y)\times m_i(t)$ terms allow for non-proportional effects of age and yod, respectively; ii) $b_k(y)\times q_j(a)$ allows for complex second-order interactions between age and yod; iii) $b_k(y)\times q_j(a)\times m_i(t)$ allows for complex third-order interactions between yod, age, and time. However, in (2), there are K×J×I terms to estimate; a penalization is thus required to avoid wiggly surfaces.

2.3.2. Measure of function wiggliness and penalized likelihood

The measure of wiggliness of a multidimensional function f_{tay} to use for penalization is based on the second derivatives and is detailed in the publication of Wood.[24]

Adding this penalization term to the likelihood leads to an excess mortality hazard h_E that varies smoothly with *t*, *a*, and *y*; thus, the change in hazard between adjacent times, adjacent ages, or adjacent yod cannot be rough. This smoothing is very appealing and natural in the context of cancer survival trends where it is not expected that treatment improvements would lead to sudden changes in mortality between close years or close ages.

2.3.3. Choice of the marginal bases used for the mortality hazard

To represent functions f_t, f_a, and f_y, we opted for restricted (or "natural") cubic splines as low-rank bases; the dimensions of these bases depend thus on the number of knots chosen. In the penalized framework, the basis dimension should be set to a value slightly higher than that deemed necessary, which brings flexibility. Here, we will focus on net survival trend analyses over 20 yod with 5 years of follow-up (see the details in section 3). According to our experience in cancer survival analysis and the recommendations of Herndon and Harrell,[30, 31] we have chosen to use six knots (including boundary knots) to model f_t, the dynamics of hazard (6 parameters), five knots to model f_a (the effect of age at diagnosis), and four knots to model f_y (the effect of the yod). Given the number of knots, 120 parameters have to be estimated in Formula 2.

According to Gray[32] and Herndon and Harrell,[31] knot location may be based on the empirical percentiles observed in the population of patients who died, which yields in our case, considering the number of knot we have chosen: i) 0^{th}, 20^{th}, 40^{th}, 60^{th}, 80^{th}, and 100^{th} percentiles of survival time for f_t; ii) 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles of age at diagnosis for f_a; and, iii) 0^{th}, 33^{th}, 66^{th}, 100^{th} percentiles of yod for f_y.

2.3.4. Parameter estimation and practical implementation in survival models

Survival models with penalized splines face the same integration problem as models with non-penalized splines (see section 2.1). Previous works have addressed this issue for

unidimensional penalized splines: in 2013, Rodriguez et al.,[33] used a penalized Poisson approach to fit a penalized survival model (an approach that parallels that of Remontet et al.[10] in the unpenalized framework). This approach allows benefiting from the well-known Generalized Additive Model (GAM) framework.[17, 34, 35] In 2005, Kauerman[36] and Becher et al.[37] considered also a penalized survival model as a Poisson GAM model using trapezoid techniques to approximate the integrals. In 2016, Liu et al.[38] proposed a penalized generalized survival model based on the parametric Royston-Parmar approach.[39] The latter authors modeled directly the cumulative hazard, which avoids the difficult integration and replaces it by a simple derivation.

Here, we adopted the Poisson approach described by Rodriguez to implement the MPS in the excess hazard model. To this end, we split the original data into small intervals (as described by Remontet et al.[10]) and, being in the context of the excess hazard model, changed the link function of the Poisson model to allow for the population mortality hazard h_P.[19] In practice, using version 1.8-3 of *mgcv*, model (1) with tensor product (2) has been implemented on split data using function *te()* to build the tensor product basis and using *gam* function to fit the model. Function *gam* makes it possible to adjust the penalized Poisson model and use a personalized link function. The model was fitted using a P-IRLS algorithm and the smoothing parameters λ_t, λ_a, and λ_y were estimated by Restricted Maximum Likelihood (REML).[17]

2.3.5. Derivation of the net survival and the confidence intervals

Once the parameters of the model are estimated as detailed above, the net survival at given time since diagnosis, age at diagnosis, and yod is obtained by integration of the excess mortality hazard using Gauss-Legendre quadrature. Using the delta-method and assuming normality of the logs of the cumulative hazards, the confidence intervals of the net survival (age-specific or age-standardized) are obtained from the Bayesian posterior covariance matrix (V_p in package *mgcv*) of the 120 parameters.[17, 40]

3. The simulation study

To assess the performance of this MPS approach in net survival trend analysis, we simulated survival data under 5 scenarios that represent a variety of trends seen in real data analyses.

3.1. The simulation design

Type of models used for the scenarios

The five scenarios represent gradually complex trends; we chose then five cancers sites according to the results of SUDCAN study for France[3] (esophagus, stomach, breast, cervix uteri, and ovary). For each cancer site, we adjusted a flexible parametric excess

hazard model on the French survival data.[41] The model mimics the effects found in SUDCAN (see details in Supplementary table S1). The scenarios differ in the effect of the yod that could be linear or non-linear, proportional or non-proportional, age-dependent or age-independent, and the most complex scenarios included third-order interactions (table S1). However, the adjusted model used fractional polynomials[11] instead of splines because splines are the basis functions of the tensor; thus, using the same bases for data generation and analyses would overestimate the performance of the MPS approach. Practically, to choose the powers of the fractional polynomials, we adapted the model-building strategy proposed by Sauerbrei et al.[12] Finally, for each scenario, the parameters obtained were regarded as theoretical parameters and used to generate the data.

Sample characteristics

Two sample sizes were considered for each scenario (N=2,000 and N=10,000). In the settings with 2,000 patients, we simulated M=1,000 datasets whereas in the settings with 10,000 patients, we simulated only M=200 datasets to limit the computing time. In each scenario, we simulated a cohort with a similar age-distribution as observed in the French data used to obtain the theoretical parameters. The yod was randomly sampled from a uniform distribution between 1990 and 2010. Patients were censored at 5 years or at end of follow-up in 2013.

Generation of time to death T

For each individual, the time to death T was the minimum of the time to death due to cancer (T_E) and the time to death from other causes (T_P), each being generated separately according to the age and the yod of the individual. T_P was generated using a piecewise exponential distribution as described by Danieli et al.[42] T_E was generated using the inverse transform approach described by Crowther and Lambert[43] for the survival context. The cumulative distribution $F_E(t)$ of T_E, which is derived from the parametric expression of h_E as $F_E(t) = 1 - \exp\left(-\int_0^T h_E(v)dv\right)$, is then numerically inverted to generate T_E from a uniform distribution; i.e., $T_E = F_E^{-1}(u)$ where u is drawn from a uniform distribution.

3.2. Description of the theoretical trends of each scenario

Theoretical excess mortality hazards for 3 ages, 3 years, and each scenario are shown in Figure 1. The theoretical trends of NS by age can be seen in Supplementary Figure S7 for example (solid line).

Scenario 1 used data on esophagus cancer and assumed no effect of the yod to allow assessment of the performance of the MPS approach in a context where the smoothing of the yod effect must be important. Another interesting feature of this scenario is the non-monotony of the hazard function according to the time elapsed since diagnosis: the excess mortality hazard increased up to one year after diagnosis then decreased dramatically

(Figure 1). This means that the smoothing of the effect of time should not be too important to correctly reproduce this curvature.

Scenario 2 used data on stomach cancer and considered that the effect of the yod was non-linear, non-proportional, and had no interaction with age at diagnosis. In this scenario, the strengths of these non-linearity and non-proportionality are rather low.

Scenario 3 used breast cancer data and considered that the effect of the yod was non-linear, proportional, and had an interaction with age at diagnosis. In this scenario, the net survival increased with the yod whatever the age but the magnitude of the increase depended on age (Figure S7). In addition, unlike the two previous scenarios, the theoretical excess mortality hazard was moderate and relatively constant along the time elapsed since diagnosis (Figure 1).

Scenario 4 used cervical cancer data. This cancer showed a linear and non-proportional effect of the yod, with a strong age-yod interaction and a triple interaction between time, age, and yod. In young women, at *fixed* time points since diagnosis between 0 and 24 months, the excess mortality decreased with the yod. In contrast, in elderly women, the excess mortality increased with the yod whatever the position of this time since diagnosis (Figure 1). This leads to an interesting feature in NS trends because NS improves in young patients but worsens in the elderly (Figure S7). All these hazard variations of cervical cancer are shown in Figure S1 that presents the theoretical excess mortality hazards in

3D-plots. This figure illustrates the complexity of the hypersurface we attempt to model with MPS.

Scenario 5, the most complex one, used ovarian cancer data and included a non-linear and non-proportional effect of yod together with a triple interaction. Indeed, at intermediate ages and fixed times since diagnosis less than 3 years, the excess mortality hazard decreased with the yod, whereas it increased with the yod for greater time spans. However, this reverse trend was less marked in young or old ages, which explains the need for a triple interaction.

Figure 2 shows the theoretical 1- and 5-year standardized net survival (sNS) according to the yod in each of the five scenarios (black solid curve). In Scenario 2 (stomach cancer), the curve shows an atypical pattern during the first yod; the sNS in 1990 is slightly higher than the one in 1991 while the survival increases afterwards (this atypical pattern is due to the use of a fractional polynomial for the effect of the yod with powers equal to 2 and -2, see Table S1). In Scenario 3 (breast cancer), the 5-year sNS curve starts flattening in year 2007. In Scenario 5 (ovary cancer), the flat part occurs between 1990 and 1995.

3.3 Simulated data analysis

Simulated data were analyzed using a tri-dimensional MPS in which the bases, the number and positions of knots, the parameter estimation method, and the practical implementation are described in sections 2.3.3 and 2.3.4.

Furthermore, in order to have comparative elements to assess the performance of the MPS, we also analyzed the data with two alternative models:

i) the true model (i.e., the model that generated the data) that can be considered as a "gold-standard", and which also enables us to validate our data generation algorithm,

ii) a basic "PH model" that may be written: $log[h_E(t,a,y)] = f_t(t) + f_a(a) + f_y(y)$, where functions *f* are defined as in the above MPS approach; i.e., restricted cubic splines with same number and positions of knots as in the MPS approach (except that the intercept is dropped for f_a and f_y for identifiability purposes). This PH model had 13 parameters and allowed for non-linear but proportional effects without interactions.

In the true and the PH model, the maximum likelihood parameter estimates (without penalization) were obtained with a homemade procedure based on Cavalieri-Simpson integral approximation and a Newton-Raphson algorithm.[10]

3.4. Assessment of the performance of the MPS approach

We examined the performance of the MPS approach and of the two alternatives in estimating the age-standardized net survival for a given yod (denoted sNS(y)): this parameter of interest was calculated in two steps using a refined annual age standardization as described by Uhry et al.[3] First, NS for each 5-year age-classes was calculated by averaging the NS predicted from the model for each annual age, using within-age-class weights as observed over the whole data. This way, the age structure within age-classes is fixed and does not vary with the year of diagnosis. The age-standardized NS was then derived from these age-class estimates using the ICSS 5-years weights.[44]

For each of the ten settings considered (5 scenarios × 2 sample sizes) and over the M simulated datasets (M=1,000 or 200, *see* section 3.1), we estimated: i) the bias, defined as the difference between the average of the M estimated values and the theoretical value of the parameter of interest; ii) the Root Mean Squared Error (RMSE), defined as the square root of the average of the squared differences between the M estimated values and the theoretical value; iii) the empirical coverage probability (CP), defined as the proportion of 95% confidence intervals that include the theoretical value.

4. The simulation study results

4.1. Bias, RMSE, and coverage probabilities

Figure 2 shows the mean (over 200 simulated datasets with 10,000 patients) of the sNS estimates obtained with the MPS approach and the PH model (see also Figure S2 for sample size 2,000 patients). Figure 3 shows for each scenario the bias made in estimating the 1- and 5-year sNS according to the yod, the method (true model, MPS, or PH model) and the dataset size (2,000 or 10,000 patients). As expected, the bias is null with the true model. With the PH model, the bias is generally greater than with MPS and does not decrease when the sample size increases. With MPS, the bias is generally low (-1 to +1) and lower with 10,000 than with 2,000 patients. Nevertheless, a bias of nearly -2 was seen with stomach cancer data of 1990 whatever the sample size; this shows that the MPS did not reproduce the atypical pattern of 1990. The MPS did not reproduce the flat trend observed in ovary and in breast cancers but led to only a slight bias in sNS with 10,000 patients. In section 4.2, we will focus on two noticeable behaviors of the MPS: i) a small bias at 1 year observed with esophagus data (N=2,000) ii) the absence of bias with cervical data despite highly complex trends.

Figure 4 shows the RMSEs according to the yod (same panel order as in Figure 3). In all five scenarios, the RMSE of the PH model was much higher than that of the true model or the MPS. In the simplest Scenario 1 (esophagus cancer, no effect of the yod), the true model had a lower RMSE than MPS. However, in all other scenarios and settings, the MPS and the true model had very close RMSEs. With stomach (at 5 years after diagnosis

in yod 1990) and breast data (at 5 years in 2010), the RMSE of MPS was very close to that of the true model despite the biases seen in Figure 3. Thus, MPS returned slightly biased but less variable estimates than those of the true model. On the basis of the results shown in Figure 4, Table 1 classifies the methods according to the RMSEs of the estimators of sNS at 5 year after diagnosis. This table shows that the performance indicators of MPS are identical to those of the true model, except in the simplest Scenario 1, and always better than those of the PH model, especially with the large sample size 10,000.

Figure 5 shows the coverage probabilities according to the yod (same panel presentation as in Figures 3 and 4). We should recall first that these probabilities are estimated with 1,000 and 200 simulated datasets when the sample size is, respectively, equal to 2,000 and 10,000 patients. So, to check whether MPS provides coverage probabilities close to the nominal value of 95%, it is better to focus on cases with 2,000 patients, those for which the accuracy of the estimation is the highest. Figure 5 shows then that the coverage probabilities of MPS are generally very satisfactory, though they are unsurprisingly lower than 95% in case of bias (e.g., stomach or ovary data of 1990) and higher than 95% in one case (breast around yod 1997).

In the online supplement, additional results are presented to detail the performance of the MPS approach and of the PH. Figures S3 to S6 show the theoretical $h_E(t)$ for 3 ages and 3 years and the mean of the estimate obtained with the two methods and two samples sizes. In the same spirit, Figures S7 to S10 show the theoretical trends of NS(t=1) and NS(t=5) by age and the mean of the estimates.

4.2. Focus on two behaviors of the MPS approach

The 1st focus point concerns the +0.5 bias seen with MPS at one year after diagnosis with esophagus data and 2,000 patients (Figure 3). Figure S3 shows that, at the beginning of the follow-up, the theoretical high curvature seen in young patients is too smoothed by the MPS: the estimated hazards are too small and the estimated NS too high, which leads to a small positive bias in young patients (Figure S7) and to an overall bias of +0.5. However, with 10,000 patients, this oversmoothing disappears practically and the fit becomes adequate (Figures S4 and S8).

The 2nd focus point concerns the good adjustment made with MPS in the cervical cancer scenario. The strong interaction between age and yod that leads to opposite trends in function of age is perfectly rendered by the MPS approach (Figure S7 and S8). On the

contrary, the PH model leads to a poor fit to the data; this is shown in Figures S5 and S6 in terms of hazard and Figures S9 and S10 in terms of NS. For example, in patients aged 79 years at diagnosis and at five years after diagnosis, the bias with the PH model is -6 in 1990 and +4 in 2010 whatever the sample size (Figures S9 and S10).

Lastly, a practical illustration of the MPS and PH approaches is presented in the last section of the online supplement, using real data from 5977 cervical cancer cases (the dataset used to determine the theoretical parameters in the cervix uteri scenario).

5. Discussion

5.1. Main finding: the good performance of the MPS approach

In this work, we propose a MPS modeling to describe the changes in cancer excess mortality in function of time since diagnosis, age at diagnosis, and yod and explore thus the trends of net survival. The excess mortality hazard $h_E(t,a,y)$ is modeled through a tensor product of the marginal basis of the three variables. This approach allowed a simultaneous modeling of non-linear effects and all types of interaction between variables (including non-proportionality). The work adapted the statistical framework developed by Wood for Generalized Linear Models[17] to the survival and net survival contexts. The extensive simulation study performed here showed that the performance indicators of the MPS approach are close to those of the true model (except in a scenario where there is no

effect of the yod). One major strength of this result is that it was obtained through the analysis of realistic simulated data generated from a model that used fractional polynomials and thus that cannot be considered as a submodel of the MPS under evaluation.

The simulation study explored five realistic scenarios that allow for some complexity in terms of effects or interaction(s). We noted a lack of fit of the MPS in two situations. The first is that of esophagus cancer where the hazard shows an important curvature: with a small sample size --thus a weak signal-- the MPS smoothing was too important and the curvature could not be fully fitted, which generated a slight but systematic bias. However, this bias faded with a larger sample size. This case illustrates the bias induced by penalization: when the information is insufficient, the MPS tends to oversmooth the curves and show simpler effects than the theoretical ones. The second situation happens when changes in sNS occurred at the beginning or at the end of the diagnosis period which led, respectively, to bias in 5-year SNS in stomach cancer in 1990 and in breast cancer in 2010. However, smoothing may provide more stable estimates and, in both situations, despite the bias, the RMSE of the MPS was equivalent to that of the true model.

In summary, the MPS has shown its ability to fit simple as well as complex trends (as in ovary and cervical cancers whose trends depend on age).

5.2. Computing aspects, smoothing parameter estimation, basis function, and type of penalties

In the present work, the practical implementation of the MPS was greatly simplified by the recourse to the Poisson approach and to the powerful package *mgcv*. Actually, after data splitting, the adjustment was made within a Poisson model that had a setting-specific link function. Package *mgcv* (especially function *gam*) is remarkably stable: over 6,000 simulated dataset runs (5,000 runs on 2,000 patients and 1,000 runs on 10,000 patients), only one failed to converge (Scenario 3 with 2,000 patients). Furthermore, function *gam* is relatively fast: with a dataset on 10,000 patients (that is, a split dataset with nearly 200,000 lines), the model fitting took nearly 3 minutes on a single desktop computer with Intel i7-4790 3.60 GHz and 16 GB of RAM. Despite the computing efficiency of package *mgcv*, the algorithms used in GAM were demanding in the survival context. Indeed, the Poisson approach requires data augmentation and the number of parameters to estimate is important due to the tensor product of three dimensions (here, 6×5×4=120 parameters). These two aspects imply dealing with huge matrices that require large RAMs. For example, above 1,500,000 lines (about 70,000 patients), the analysis failed due to lack of memory.

Generalized Cross-Validation (GCV) and REML can be used to estimate the smoothing parameters. When the Bayesian covariance matrix is used, additional simulations showed

that the results obtained with REML and GCV were almost identical in terms of bias, RMSE, and coverage probability (data not shown): in the present context, both methods are good solutions to estimate the smoothing parameters.

In this work, knot location was based on the quantiles of variables of patients who died, which is a common choice in survival analyses.[9, 30, 31] Restricted cubic splines were used as basis functions for time since diagnosis, age, and yod; they are implemented in package *mgcv* via options *bs='cr'* of the *te* function. The P-splines proposed by Marx and Eilers[22, 25, 26] are also implemented in *te* function (option *bs="ps"*). In the P-spline approach, the basis functions are cubic B-splines, the knots should be evenly spaced, and the penalization is directly imposed on the coefficients. Additional results suggested that, in our setting, performances of the P-spline tensor and MPS were roughly comparable. Other choices concerning the bases are theoretically possible with function *te* but the user should then build his/her own bases (see function *smooth.construct* of package *mgcv*).

5.3. Interest and limits of the present approach, applicability in case of small sample size

The simulation results support the fact that the MPS approach is well-adapted to descriptive cancer epidemiology, especially to the analysis of the trends of net survival

as well as, naturally, the trends of "overall" survival (which may be simply obtained by setting the population mortality hazard h_P to zero).

The main objective was to build a flexible model able to provide smooth estimates at the same time. This objective was achieved with the MPS: the tensor product of cubic regression splines provides flexibility, including high-order interaction, while the smoothing parameters provide smoothness in each of the three directions (*t,a,y*).

Another asset of the MPS is the simultaneous specification of non-linearity, non-proportionality, and other interactions, which is directly obtained by modeling hypersurface (*t,a,y,h(t,a,y)*). This simultaneous specification is essential in survival analysis because non-linearity and non-proportionality do interact: omitting or misspecifying the functional form of a continuous variable may lead to spurious non-proportionality and, conversely, omitting or misspecifying non-proportionality may lead to a spurious functional form.[45, 46] When the number of variables is low, the very challenging issue of model-building strategy can then be reduced with MPS.

The parametric aspect of MPS allows explicit derivations of hazard, survival functions, and predictions. This allows a large choice of graphical displays of the results (examples can be found in Supplementary data). In particular, the dynamics of the mortality hazard

is valuable for clinicians and epidemiologists. Moreover, predictions can be made for any variable value, which allows: i) deriving other interesting outcomes such as the crude probability of death[47] or the number of years of life lost due to cancer;[48] ii) performing fine standardization, which avoids residual age effects that may affect the classical standardization[3], iii) using MPS as a forecasting tool.

In the MPS framework, the three variables (time since diagnosis, age at diagnosis, and yod) are kept in their original continuous form, which prevents a loss of information that occurs inevitably upon variable categorization and allows an accurate description of variable effects. One may note that these three variables are dealt with equally; i.e., in specifying the model, variable *t* has no particular mathematical role (contrarily to what is generally seen in survival models).

To provide practical guidelines for practitioners in case of small sample sizes, we checked the behavior of the MPS approach by running additional simulations on 250 to 1,000 cases in the ovarian cancer scenario (data not shown). The MPS worked satisfactorily with no convergence problems and provided better RMSE values than with the true model or the PH model. Thus, given its favorable bias-variance trade-off with small sample sizes, the MPS approach seemed to be robust and efficient. However, as in any other statistical analysis, the amount of information needed to study a phenomenon depends on

the magnitude of this phenomenon; in the ovarian cancer example, N=1000 was the minimum necessary to reach a reasonable precision of NS estimates to allow studying their trends.

The present method reaches its limit when the number of variables increases. For example, dealing with 10 variables having each a marginal basis with four parameters leads to estimate 4^{10} parameters, which requires the use of other approaches. In the presence of a high number of variables, the penalized likelihood approach can still be used but requires further variable selection strategies: Marra and Wood[49] gave an overview of this subject in GAM (not in survival model) and Rodriguez-Girondo et al.[33] evaluated some of these strategies within the context of survival model but without dealing with the presence of interaction between continuous variables. One may cite other approaches in survival but these are based on unpenalized likelihood: Sauerbrei et al.[12] proposed a complex algorithm stemming from fractional polynomials and, in a simulation study, Wynant and Abrahamowicz[6] evaluated four strategies based on a stepwise procedure. However, none of these two works dealt with the issue of interaction and the procedures became very complex when the number of variables increased. Another strategy is the Hazard Regression (HARE) proposed by Kooperberg et al.[50]

5.4. Prospects

Adding one or two other dimensions to the present MPS approach is an interesting prospect. Adding a spatial dimension (as in the work of Ugarte et al.[27]) would lead to a spatio-temporal model whereas adding a deprivation dimension[51] would allow analyzing survival trends according to the socio-economic status. Furthermore, forecasting being a logical consequence of smoothing[29] and survival projections being an important public health topic, it would be interesting to use the MPS as a projection tool.

Acknowledgments

The authors thank the ANR (Agence Nationale de la Recherche) for supporting this study of the CENSUR group (ANR grant number ANR-12-BSV1-0028). This research was carried out in the context of a four-institute research-program partnership involving the Institut National du Cancer (INCa), Santé Publique France (SPF), the French network of cancer registries (FRANCIM), and Hospices Civils de Lyon (HCL). The authors are also grateful to Jacques Estève for valuable advice and to FRANCIM for the access to the data.

Declaration of conflicting interests

The authors declare no conflicts of interest linked with the research, the authorship, and/or publication of this article.

References

1. Colonna M, Bossard N, Remontet L and Grosclaude P. Changes in the risk of death from cancer up to five years after diagnosis in elderly patients: a study of five common cancers. *Int J Cancer*. 2010; 127: 924-31.
2. Bossard N, Velten M, Remontet L, et al. Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *Eur J Cancer*. 2007; 43: 149-60.
3. Uhry Z, Bossard N, Remontet L, Iwaz J and Roche L. New insights into survival trend analyses in cancer population-based studies: the SUDCAN methodology. *Eur J Cancer Prev*. 2016; 26 Trends in cancer net survival in six European Latin Countries: the SUDCAN study: S9-S15.
4. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet*. 2015; 385: 977-1010.
5. De Angelis R, Sant M, Coleman MP, et al. Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE--5-a population-based study. *Lancet Oncol*. 2014; 15: 23-34.
6. Ederer F, Axtell LM and Cutler SJ. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr*. 1961; 6: 101-21.
7. Perme MP, Stare J and Esteve J. On estimation in relative survival. *Biometrics*. 2012; 68: 113-20.
8. Crowther MJ and Lambert PC. A general framework for parametric survival analysis. *Stat Med*. 2014; 33: 5280-97.
9. Giorgi R, Abrahamowicz M, Quantin C, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med*. 2003; 22: 2767-84.
10. Remontet L, Bossard N, Belot A and Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med*. 2007; 26: 2214-28.
11. Royston P and Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for continuous variables*. New-York: Wiley, 2008.
12. Sauerbrei W, Royston P and Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biom J*. 2007; 49: 453-73.
13. Wynant W and Abrahamowicz M. Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Stat Med*. 2014; 33: 3318-37.

14. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med*. 2016; 35: 3066-84.
15. Mounier M, Bossard N, Remontet L, et al. Changes in dynamics of excess mortality rates and net survival after diagnosis of follicular lymphoma or diffuse large B-cell lymphoma: comparison between European population-based data (EUROCARE-5). *Lancet Haematol*. 2: e481-91.
16. Burnham KP and Anderson DR. *Model selection and multimodel inference: a pratical information -Theoretic Approach*. 2nd ed. New York: Springer-Verlag, 2010, p.488.
17. Wood SN. *Generalized additive models: an introduction with R*. 2nd ed. London: Chapman & Hall/CRC, 2017.
18. Hastie T and Tibshirani R. Varying-coefficient models. *J R Stat Soc Series B*. 1993; 55: 757-96.
19. Dickman PW, Sloggett A, Hills M and Hakulinen T. Regression models for relative survival. *Stat Med*. 2004; 23: 51-64.
20. Wood SN. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*. 2002; 157: 157-77.
21. Marra G and Radice R. Penalised regression splines: theory and application to medical research. *Stat Methods Med Res*. 2010; 19: 107-25.
22. Eilers PH and Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science*. 1996; 11: 89-121.
23. Ruppert D, Wand MP and Carroll RJ. Semiparametric regression during 2003-2007. *Electron J Stat*. 2009; 3: 1193-256.
24. Wood SN. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*. 2006; 62: 1025-36.
25. Marx BD and Eilers PH. Multidimensional Penalized Signal Regression. *Technometrics*. 2005; 47: 13-22.
26. Eilers PH and Marx BD. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*. 2003; 66: 159-74.
27. Ugarte MD, Goicoa T, Etxeberria J and Militino AF. Projections of cancer mortality risks using spatio-temporal P-spline models. *Stat Methods Med Res*. 2012; 21: 545-60.
28. Etxeberria J, Ugarte MD, Goicoa T and Militino AF. On predicting cancer mortality using ANOVA-type P-spline models. *Revstat*. 2015; 13: 21-40.
29. Currie ID, Durban M and Eilers PH. Smoothing and forecasting mortality rates. *Statistical Modelling*. 2004; 4: 279-98.

30. Herndon JE and Harrell FE. The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods*. 1990; 19: 639-63.
31. Herndon JE and Harrell FE. The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Stat Med*. 1995; 14: 2119-29.
32. Gray RJ. Flexible methods for analysing survival data using splines, with applications to breast cancer prognosis. *J Am Stat Assoc*. 1992; 87: 942-51.
33. Rodriguez-Girondo M, Kneib T, Cadarso-Suarez C and Abu-Assi E. Model building in nonproportional hazard regression. *Stat Med*. 2013; 32: 5301-14.
34. Hastie T and Tibshirani R. *Generalized additive models*. London: Chapman and Hall, 1990.
35. Ruppert D, Wand MP and Carroll RJ. *Semiparametric regression*. New York: Cambridge University Press, 2003.
36. Kauermann G. Penalized spline smoothing in multivariable survival models with varying coefficients. *Comput Stat Data Anal*. 2005; 49: 169-86.
37. Becher H, Kauermann G, Khomski P and Kouyate B. Using penalized splines to model age- and season-of-birth-dependent effects of childhood mortality risk factors in rural Burkina Faso. *Biom J*. 2009; 51: 110-22.
38. Liu XR, Pawitan Y and Clements M. Parametric and penalized generalized survival models. *Stat Methods Med Res*. 2016.
39. Royston P and Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002; 21: 2175-97.
40. Nychka D. Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Association*. 1988; 83: 1134-43.
41. Cowppli-Bony A, Uhry Z, Remontet L, et al. Survie des personnes atteintes de cancer en France métropolitaine, 1989-2013. Partie 1 – Tumeurs solides. Saint-Maurice: Institut de veille sanitaire, 2016, p. 274.
42. Danieli C, Remontet L, Bossard N, Roche L and Belot A. Estimating net survival: the importance of allowing for informative censoring. *Stat Med*. 2012; 31: 775-86.
43. Crowther MJ and Lambert PC. Simulating biologically plausible complex survival data. *Stat Med*. 2013; 32: 4118-34.
44. Corazziari I, Quinn M and Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer*. 2004; 40: 2307-16.
45. Abrahamowicz M and MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med*. 2007; 26: 392-408.
46. Buchholz A and Sauerbrei W. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biom J*. 2011; 53: 308-31.
47. Charvat H, Bossard N, Daubisse L, Binder F, Belot A and Remontet L. Probabilities of dying from cancer and other causes in French cancer patients based on an

unbiased estimator of net survival: a study of five common cancers. *Cancer Epidemiol*. 2013; 37: 857-63.
48. Andersson TM, Dickman PW, Eloranta S, Lambe M and Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Stat Med*. 2013; 32: 5286-300.
49. Marra G and Wood SN. Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*. 2011; 55: 2372-87.
50. Kooperberg C, Stone CJ and Truong YK. Hazard Regression. *J Am Stat Assoc*. 1995; 90: 78-94.
51. Guillaume E, Pornet C, Dejardin O, et al. Development of a cross-cultural deprivation index in five European countries. *J Epidemiol Community Health*. 2016; 70: 493-9.

Table 1 - Comparison of model performance according to the RMSEs of the estimators of the standardized 5-years net survival.

Scenario	Sample size	Rank
Esophagus	2,000; 10,000	True model > MPS > PH model
Stomach, breast, cervix uteri, ovary	2,000	MPS ~ True model > PH model
Stomach, breast, cervix uteri, ovary	10,000	MPS ~ True model >> PH model

"~" equivalent to, ">" more performant than, ">>" much more performant than

Legends to the figures

Figure 1. Theoretical excess mortality hazard as a function of time since diagnosis in the five scenarios, at 3 ages (10th, 50th, and 90th percentiles of the age distribution of the cases). Black solid curve: year of diagnosis 1990; red dashed curve: year of diagnosis 2000; green double-dashed curve: year of diagnosis 2010

Figure 2. Standardized 1 and 5-years net survival as a function of the yod in the five scenarios with 10,000 patients. Black solid curve: theoretical net survival trends. Red dashed curve: mean of the standardized net survival estimated using MPS. Blue double-dashed curve: mean of the standardized net survival estimated using the PH model.

Figure 3. Bias in estimating the standardized 1 and 5-years net survival as a function of the yod in the five scenarios with 2,000 and 10,000 patients. Black solid curve: bias with the true model. Red dashed curve: bias with the MPS. Blue double-dashed curve: bias with the PH model.

Figure 4. Root Mean Squared Errors in estimating the standardized 1 and 5-years net survival as a function of the yod in the five scenarios with 2,000 and 10,000 patients.

Black solid curve: RMSE with the true model. Red dashed curve: RMSE with the MPS. Blue double-dashed curve: RMSE with the PH model.

Figure 5. Coverage probability (CP) of the 95% confidence intervals of the estimates of the standardized 1 and 5-years net survival as a function of the yod in the five scenarios with 2,000 and 10,000 patients. Black solid curve: CP with the true model. Red dashed curve: CP with the MPS. Blue double-dashed curve: CP with the PH model.

Statistical Methods in Medical Research

Online supplementary material for:

Flexible and structured survival model for a simultaneous estimation of non-linear, non-proportional effects and complex interactions between continuous variables: performance of this multidimensional penalized splines approach in net survival trend analysis.

Laurent Remontet, Zoé Uhry, Nadine Bossard, Jean Iwaz, Aurélien Belot, Coraline Danieli, Hadrien Charvat, Laurent Roche

Corresponding author:
Laurent Remontet
Email: Laurent.remontet@chu-lyon.fr

Contents

Supplementary Table S1. Model used for $\log[h_{\mathrm{E}}(t,a,y)]$ to generate the data in the simulation study (FP: Fractional Polynomial).

Scenario	Cancer site	Type of model used for $\log[h_{\mathrm{E}}(t,a,y)]$ [(1)]	Detail of the model used for $\log[h_{\mathrm{E}}(t,a,y)]$ [(2)]	Characteristics
1	Esophagus	$\sim f(t) + h(t) \times a + g(a)$	$\sim FP_0(t) + FP_1(t) \times a + \beta_7 a + \beta_8 \left(1/(\sqrt{a})\right) + \beta_9 \log(a)$	No effect of y
2	Stomach	$\sim f(t) + h(t) \times a + g(a) +$ $k(y) + y{:}\,\gamma(t)$	$\sim FP_0(t) + FP_1(t) \times a + \beta_7 a + \beta_8 (1/a^2) + \beta_9 a^3 +$ $\beta_{10}\left(\frac{1}{y^2}\right) + \beta_{11} y^2 + y{:}\left[\beta_{12}\left(\frac{1}{(t+1)}\right) + \beta_{13}\log(t+1) + \beta_{14} t\right]$	NLIN-NPH effect of y and no a-y interaction
3	Breast	$\sim f(t) + h(t) \times a + g(a) +$ $k(y) + n(a){:}\,k(y)$	$\sim FP_0(t) + FP_1(t) \times a + \beta_7 a + \beta_8 a^3 + \beta_9 [a^3 \log(a)] +$ $\beta_{10} y^2 + \beta_{11} y^3 + \beta_{12} a^2 y^2 + \beta_{13} a^2 y^3 + \beta_{14} a^3 y^2 + \beta_{15} a^3 y^3$	NLIN-**PH** effect of y with a-y interaction
4	Cervix uteri	$\sim f(t) + h(t) \times a + g(a) +$ $y + y{:}\,t +$ $n(a){:}\,(y + y{:}\,t)$	$\sim FP_0(t) + FP_1(t) \times a + \beta_7 a + \beta_8 a^2 + \beta_9 \log(a) +$ $\beta_{10} y + \beta_{11} \mathrm{yt} +$ $\beta_{12} a^2 y + \beta_{13} a^2 \mathrm{yt} + \beta_{14} a^3 \log(a)\, y + \beta_{15} a^3 \log(a)\, \mathrm{yt}$	**LIN**-NPH effect of y with **high** a-y interaction and **triple** t-a-y interaction
5	Ovary	$\sim f(t) + h(t) \times a + g(a) +$ $k(y) + y{:}\,t +$ $n(a){:}\,[k(y) + y{:}\,t]$	$\sim FP_0(t) + FP_1(t) \times a + \beta_7 a + \beta_8 \sqrt{a} + \beta_9 a^2 +$ $\beta_{10} y^2 + \beta_{11} y^2 \log(y) + \beta_{12} y{:}\, t +$ $\beta_{13} y^2 \sqrt{a} + \beta_{14} y^2 \log(y) \sqrt{a} + \beta_{15} y t \sqrt{a} + \beta_{16} y^2 log(a) \sqrt{a} +$ $\beta_{17} log(a) \sqrt{a} y^2 \log(y) + \beta_{18} log(a) \sqrt{a} yt$	**NLIN-NPH** effect of y and **complex triple** t-a-y interaction

(1) Choice of scenarios (cancer site combined with type of model) were guided by the French results of SUDCAN study. Note that for esophagus (scenario 1 with no year-effet), a slight linear effect of y was actually observed in SUDCAN and that for cervix uteri, the type of model is based on an updated analysis of French data (in SUDCAN, where period of diagnosis ended in 2004, no year effect was actually found)

(2) $FP_0(t) = \beta_0 + \beta_1 \left(\frac{1}{t+1}\right) + \beta_2\, log(t+1) + \beta_3 t$ and $FP_1(t) = \beta_4 \left(\frac{1}{t+1}\right) + \beta_5\, log(t+1) + \beta_6 t$

Figure S1. Theoretical excess mortality hazard used for the generation of cervix uteri data in scenario 4. First row: 3d-plot of the excess hazard as a function of time and age, at years 1990, 2000, and 2010; second row: 3d-plot of the excess hazard as a function of year and age, at 0.5, 1, and 5 years; third row: 3d-plot of the excess hazard as a function of year and time, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases).

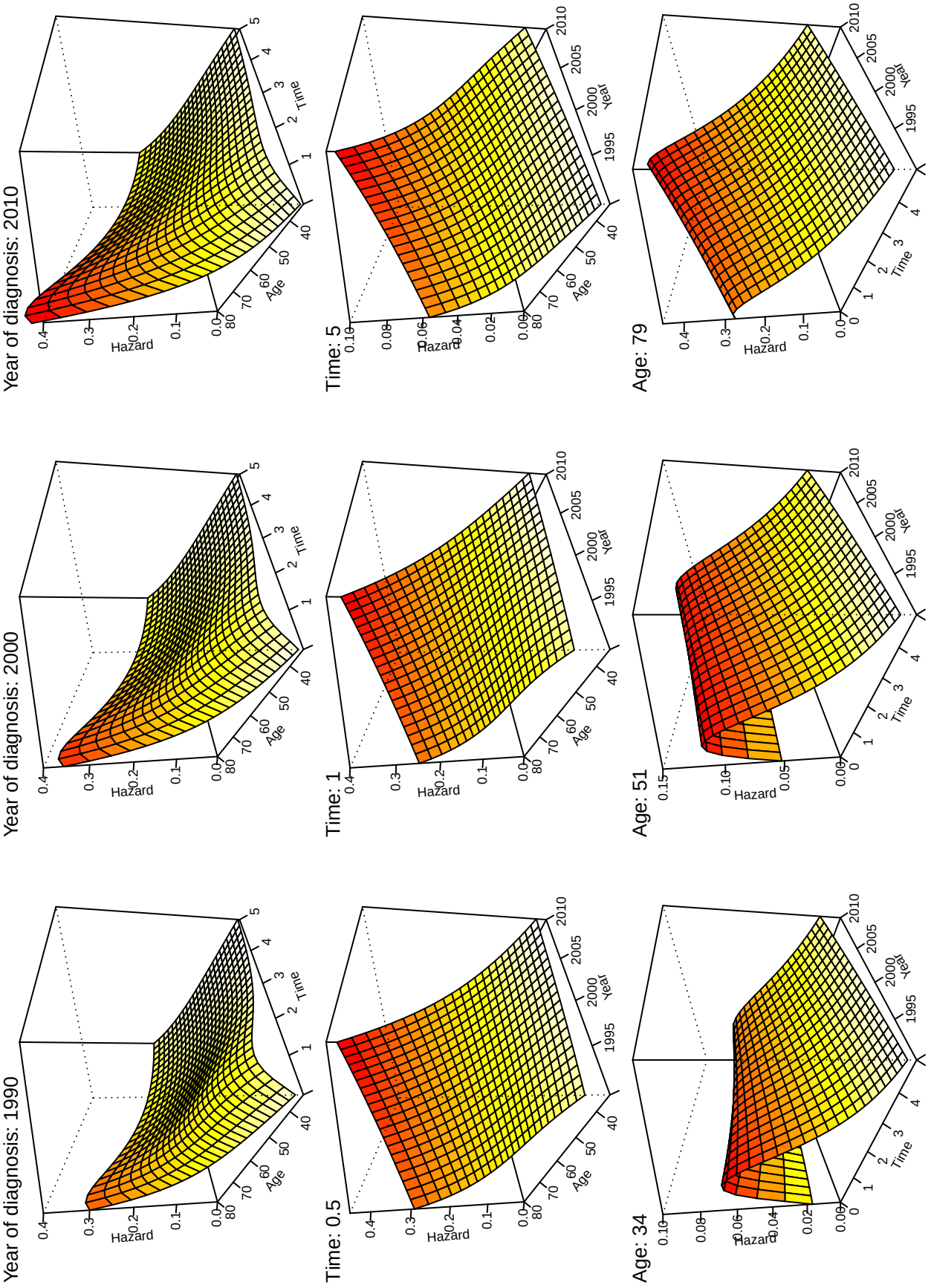


Figure S2. Standardized net survival at 1 and 5 years as a function of the year of diagnosis in the five scenarios, with 2000 cases. Black solid curve: Theoretical standardized net survival; blue double-dashed curve: Mean of the standardized net survival estimated using the Proportional Hazard model (PH); red dashed curve: Mean of the standardized net survival estimated using the multidimensional penalized splines approach (MPS).

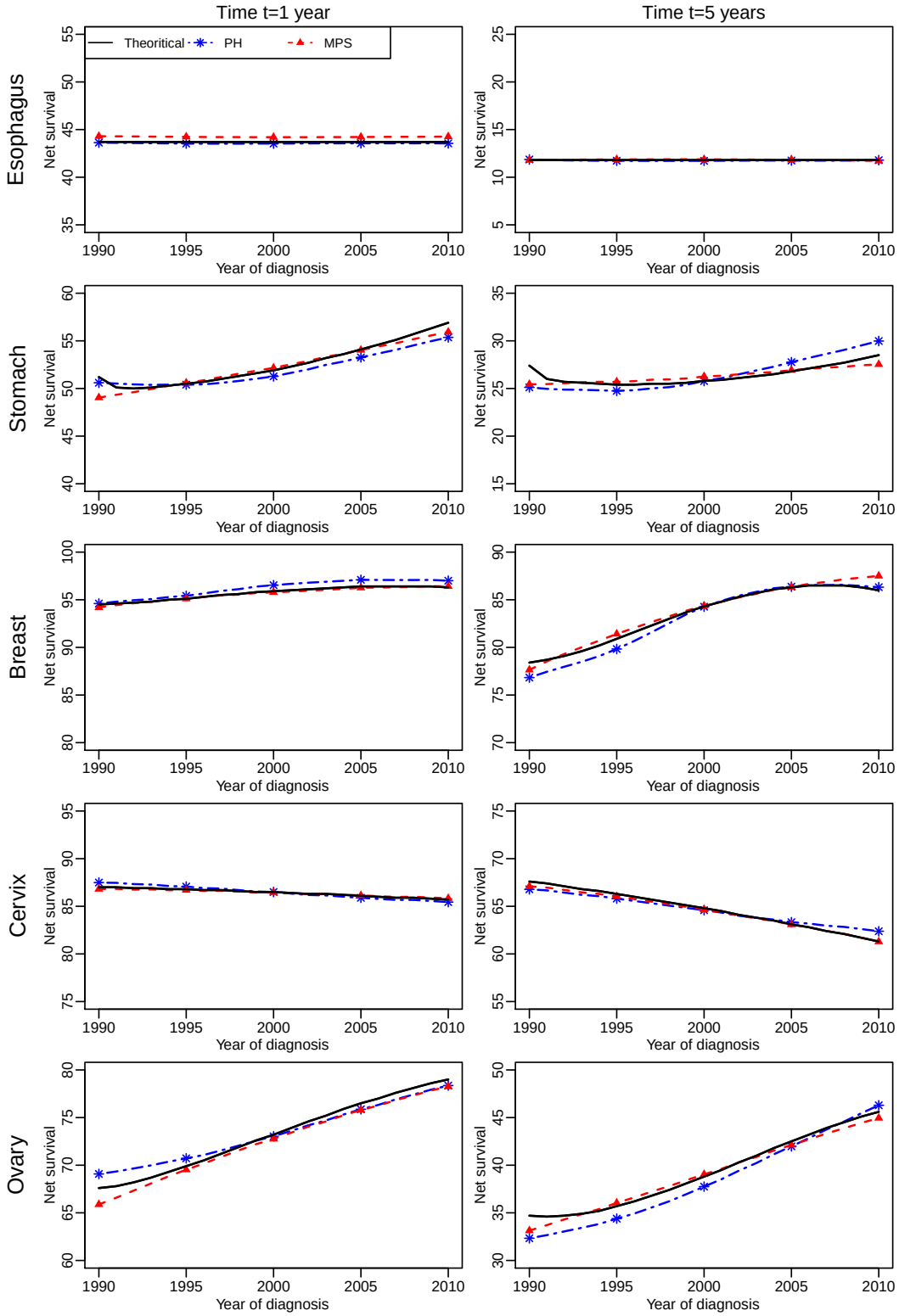


Figure S3. Excess mortality hazard as a function of time since diagnosis in the five scenarios **with 2000 cases**, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases). Solid curve: Theoretical excess mortality hazard; dashed curve: Mean of the excess mortality hazard using the **multidimensional penalized splines approach**.

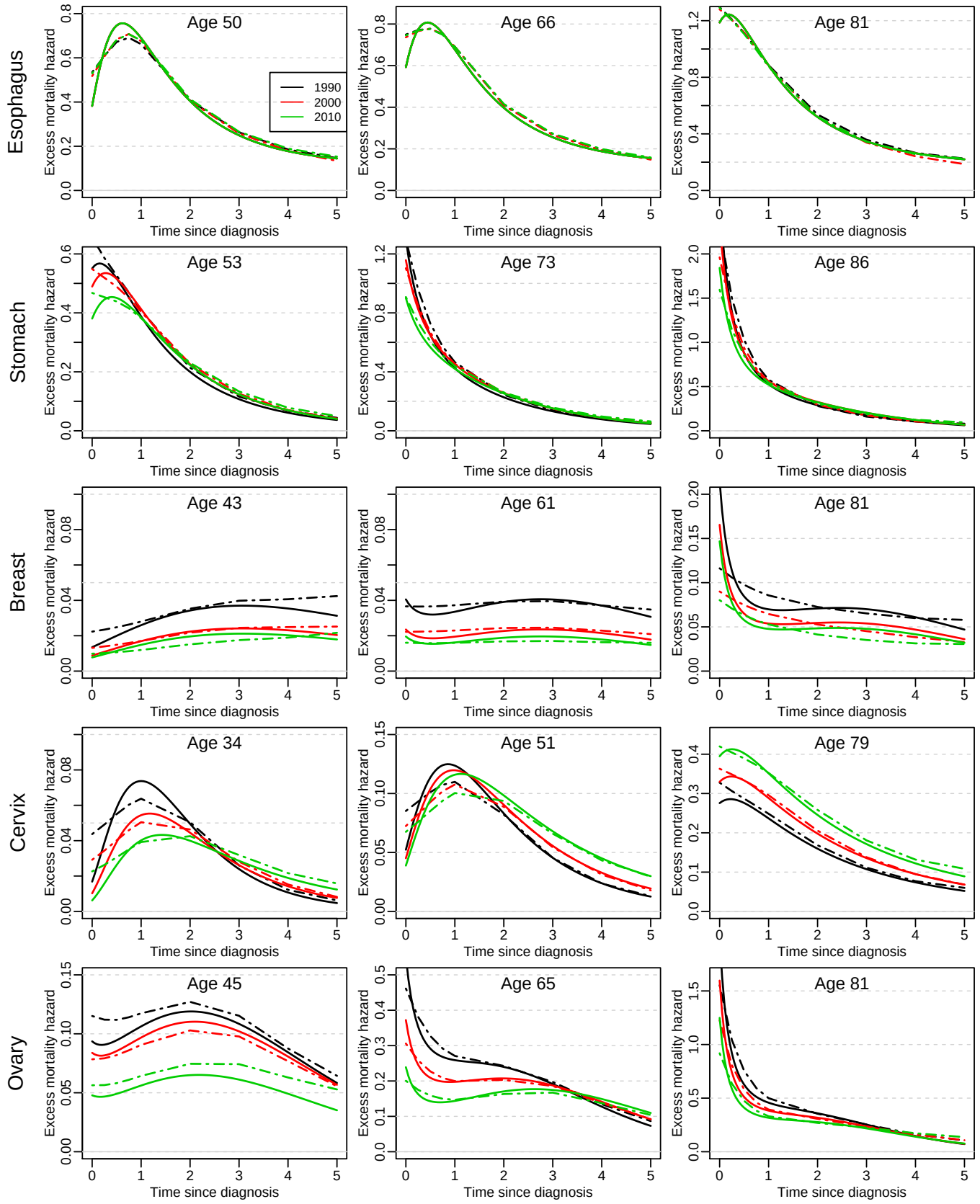


Figure S4. Excess mortality hazard as a function of time since diagnosis in the five scenarios **with 10000 cases**, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases). Solid curve: Theoretical excess mortality hazard; dashed curve: Mean of the excess mortality hazard using the **multidimensional penalized splines approach**.

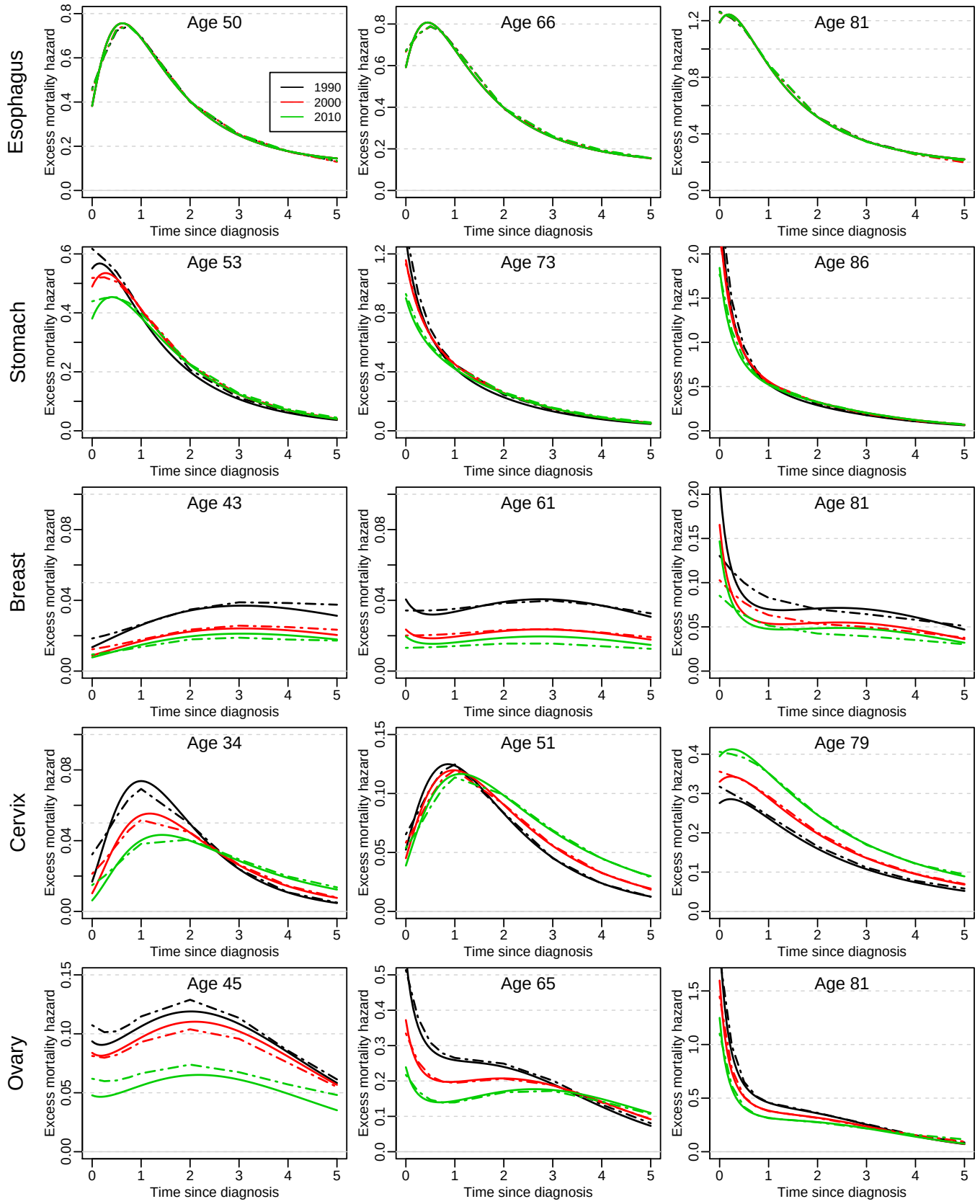


Figure S5. Excess mortality hazard as a function of time since diagnosis in the five scenarios **with 2000 cases**, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases). Solid curve: Theoretical excess mortality hazard; dashed curve: Mean of the excess mortality hazard using the **Proportional Hazard model**.

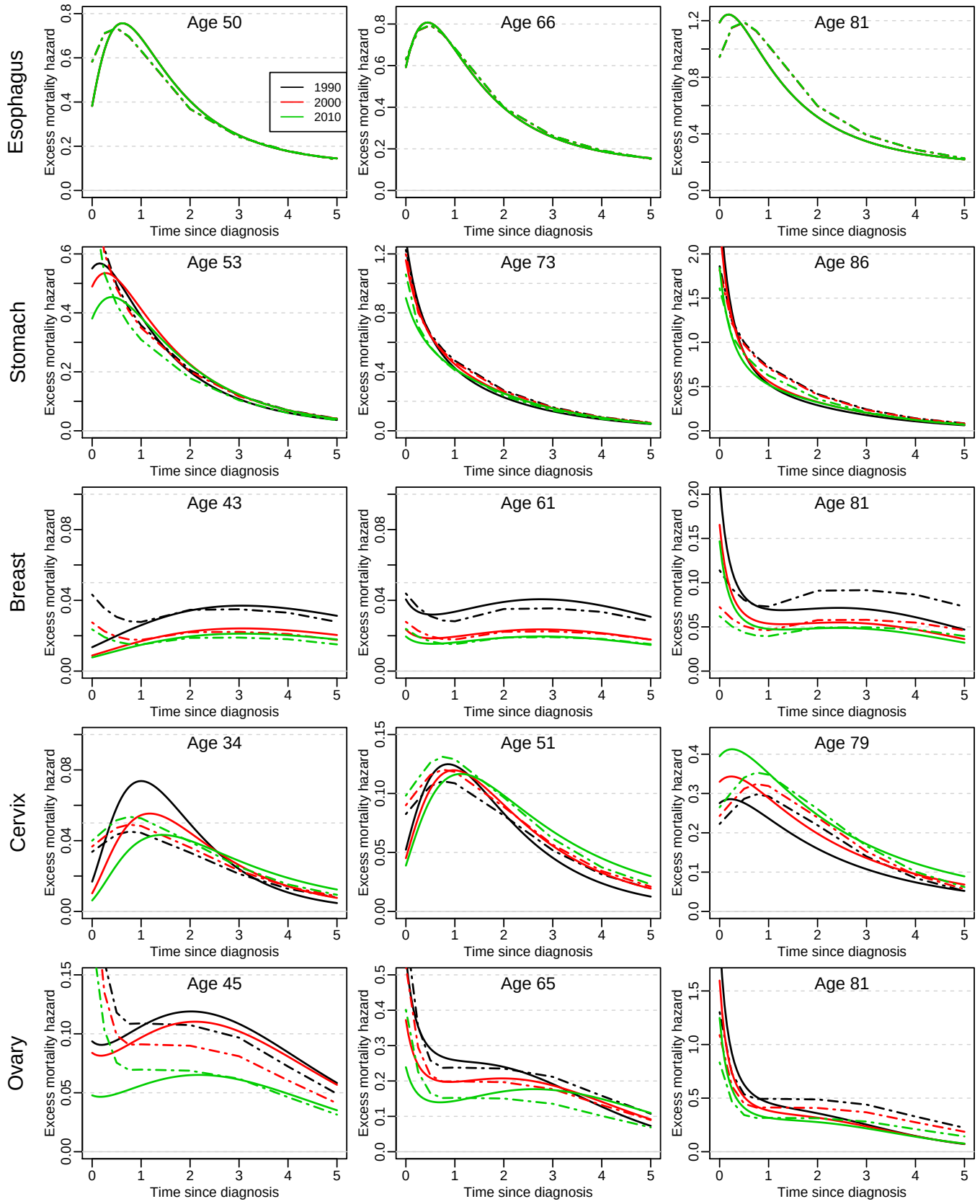


Figure S6. Excess mortality hazard as a function of time since diagnosis in the five scenarios **with 10000 cases**, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases). Solid curve: Theoretical excess mortality hazard; dashed curve: Mean of the excess mortality hazard using the **Proportional Hazard model**.

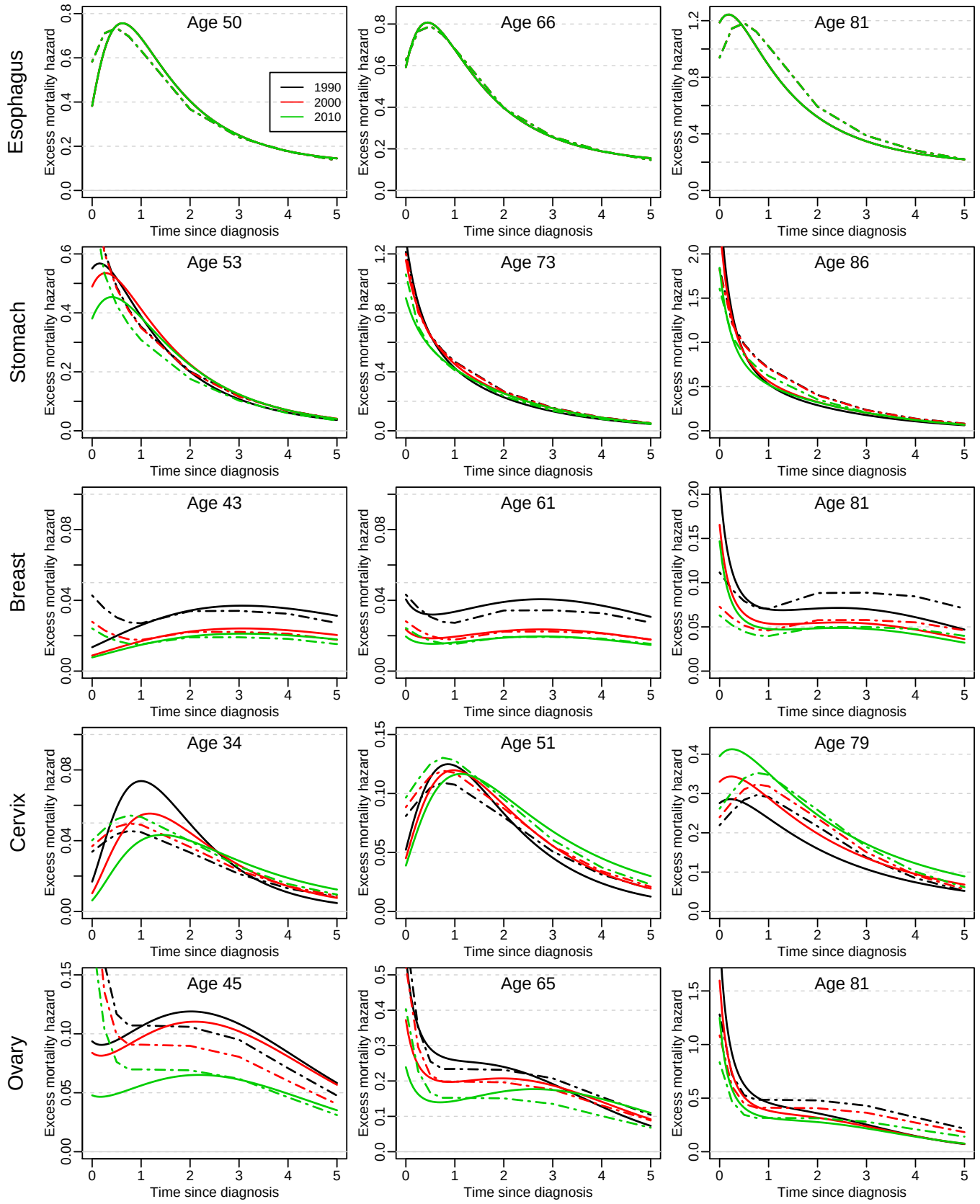


Figure S7. Net survival at 1 and 5 years as a function of the year of diagnosis in the five scenarios **with 2000 cases**, at 3 ages (10th, 50th, and 90th percentiles of the age distribution of the cases). Solid curve: Theoretical age-specific net survival; dashed curve: Mean of the age-specific net survival estimated using the **multidimensional penalized splines approach**.

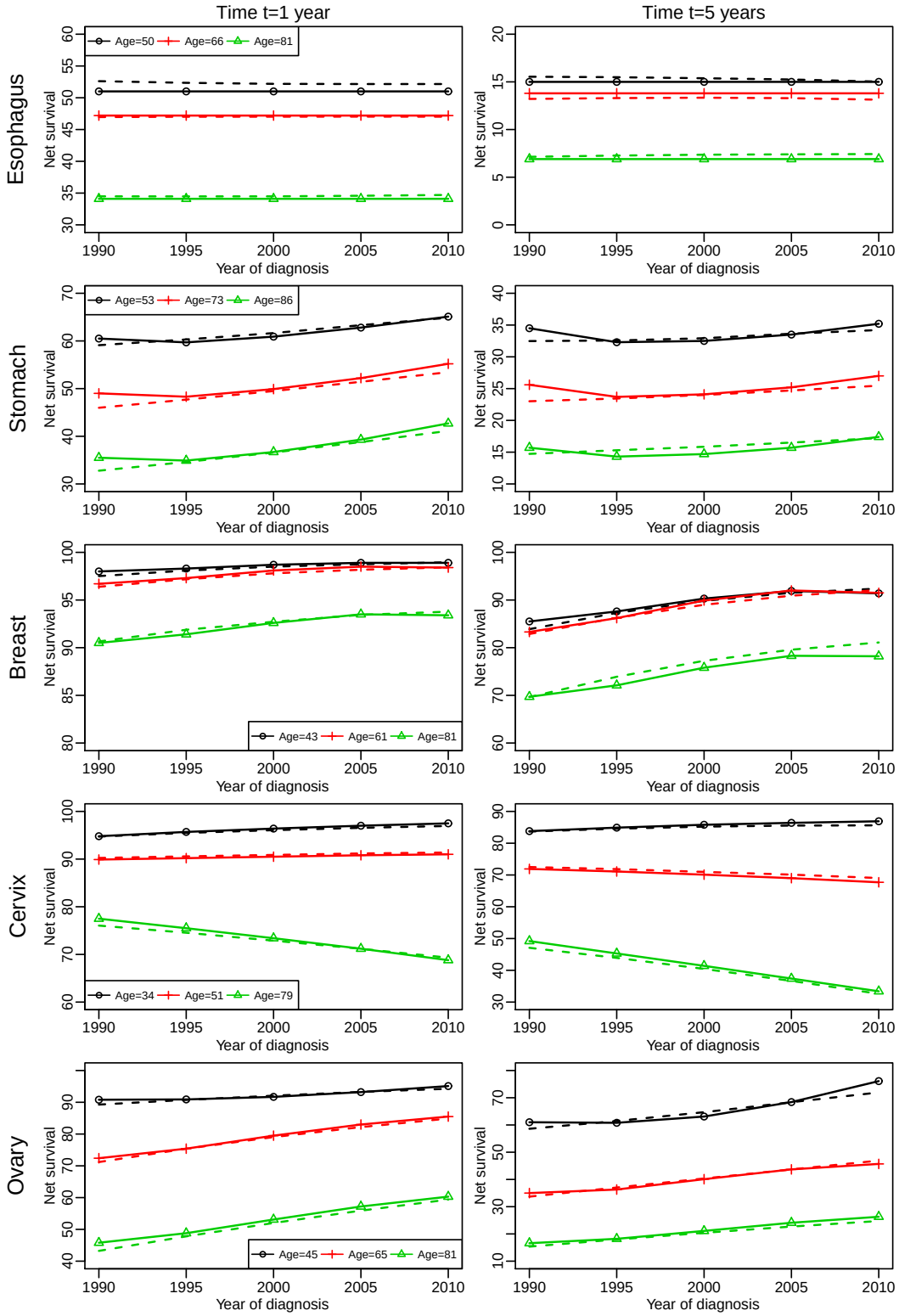


Figure S8. Net survival at 1 and 5 years as a function of the year of diagnosis in the five scenarios **with 10000 cases**, at 3 ages (10th, 50th, and 90th percentiles of the age distribution of the cases). Solid curve: Theoretical age-specific net survival; dashed curve: Mean of the age-specific net survival estimated using the **multidimensional penalized splines approach**.

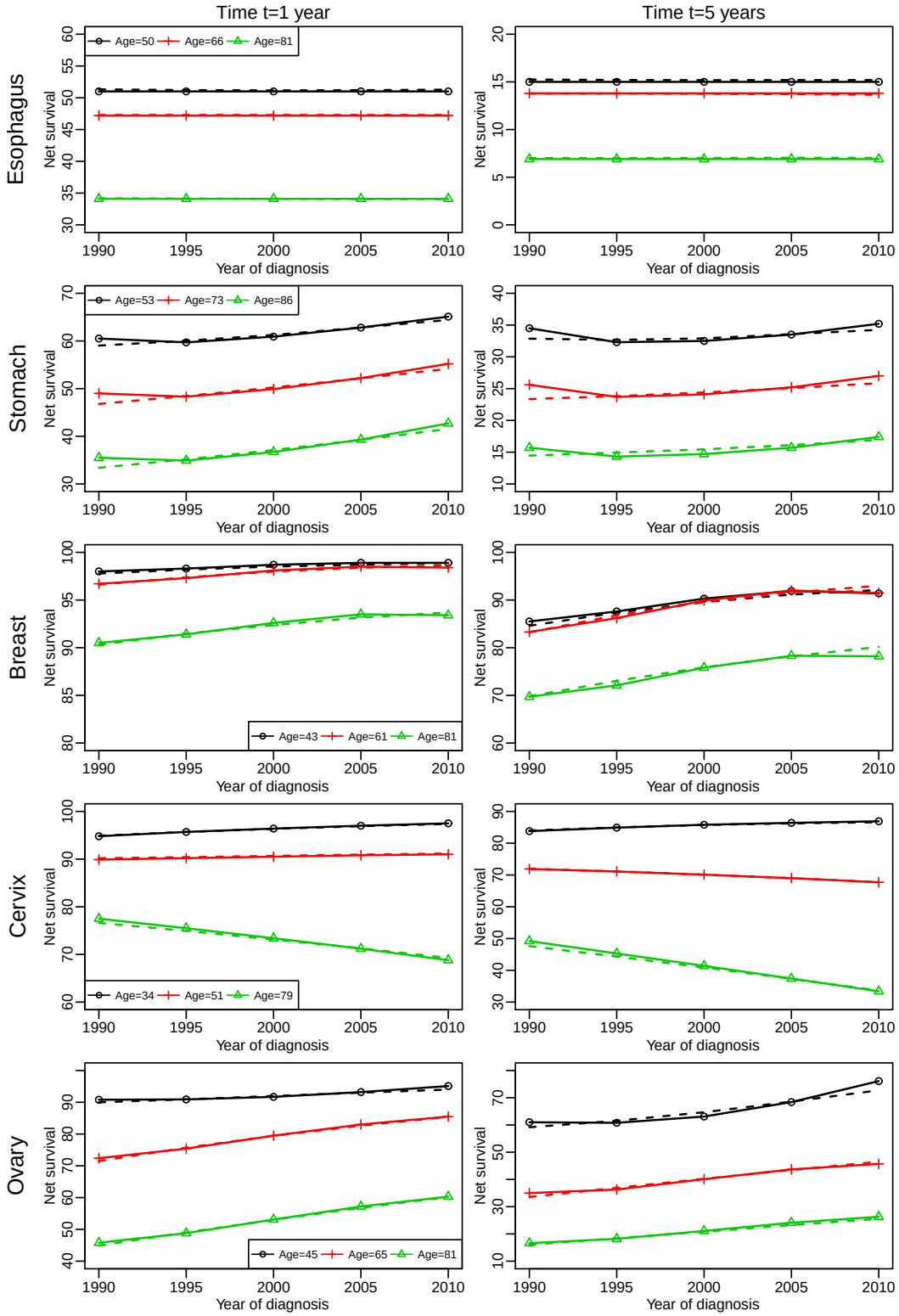


Figure S9. Net survival at 1 and 5 years as a function of the year of diagnosis in the five scenarios **with 2000 cases**, at 3 ages (10th, 50th, and 90th percentiles of the age distribution of the cases). Solid curve: Theoretical age-specific net survival; dashed curve: Mean of the age-specific net survival estimated using the **Proportional Hazard model**.

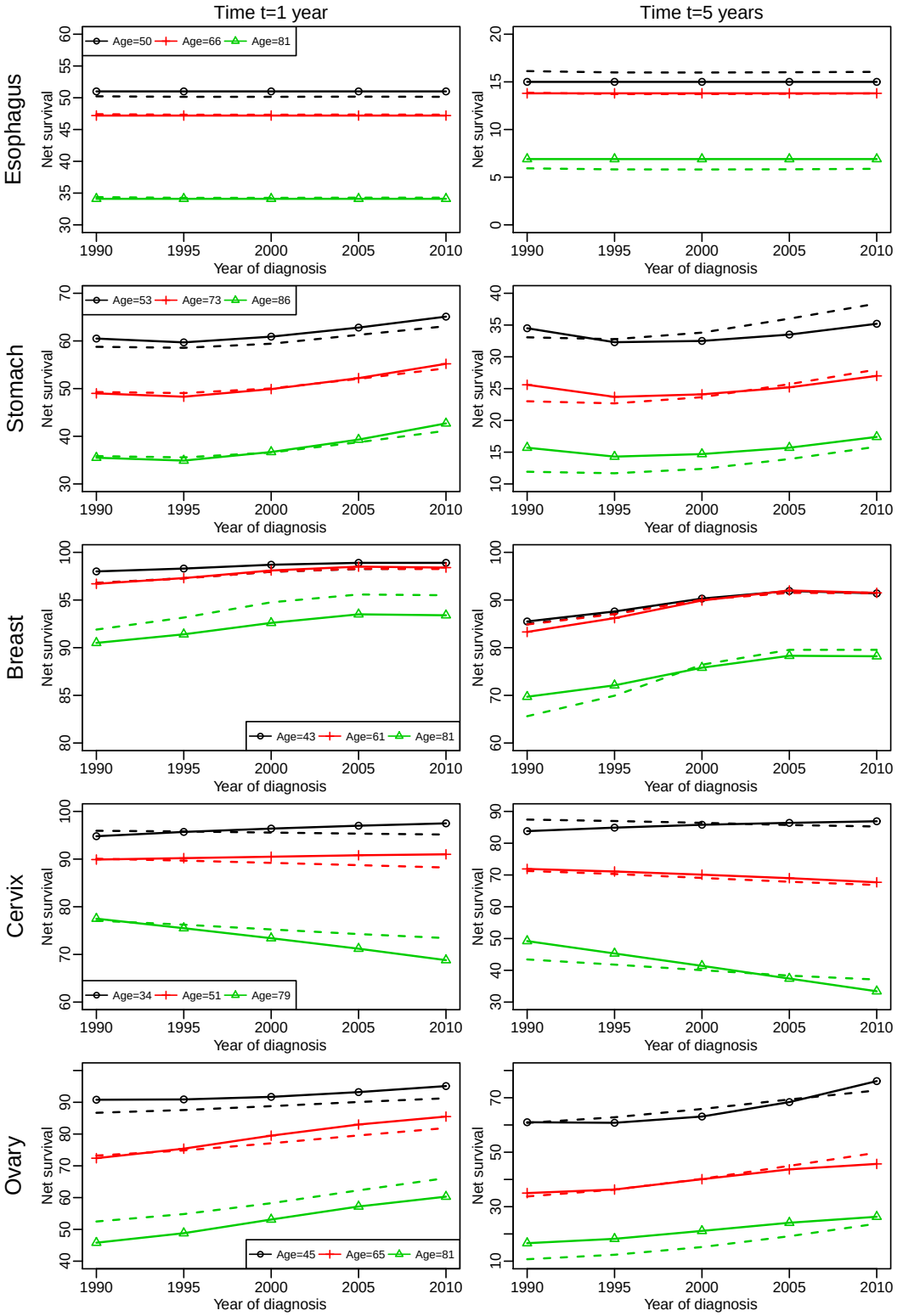
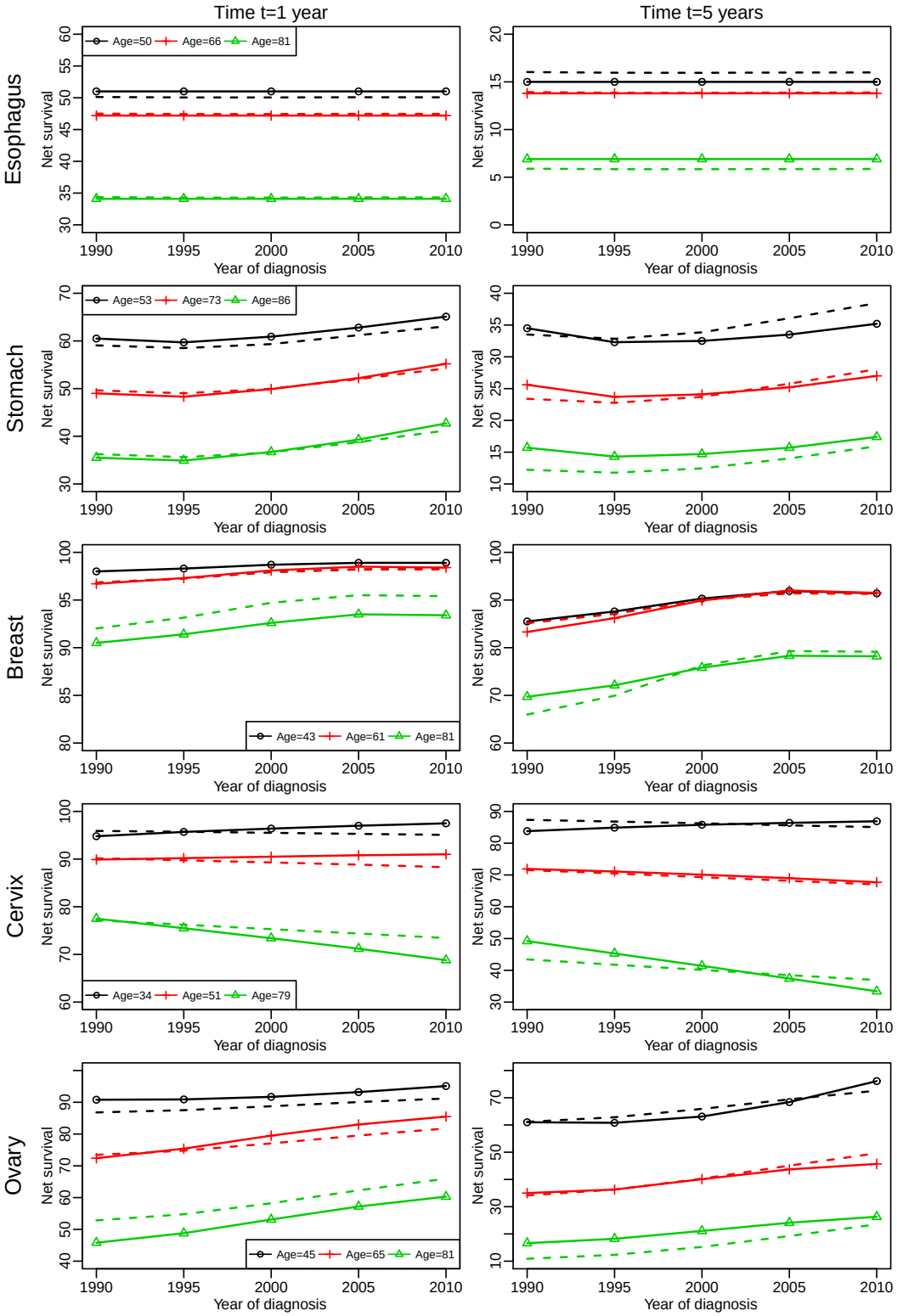


Figure S10. Net survival at 1 and 5 years as a function of the year of diagnosis in the five scenarios **with 10000 cases**, at 3 ages (10^{th}, 50^{th}, and 90^{th} percentiles of the age distribution of the cases). Solid curve: Theoretical age-specific net survival; dashed curve: Mean of the age-specific net survival estimated using the **Proportional Hazard model**.



Case study: trends in net survival and in the dynamics of excess hazard from cervical cancer, in France.

This section is an illustration of a survival trends population-based study, as performed by the Multidimensional Penalized Splines approach (MPS) and the Proportional Hazard model (PH).

Here, we studied trends in net survival (NS) and in the excess hazard for cervical cancer in France; this study included all incident cases of primary invasive cervical cancer (ICD-03 code C53) diagnosed between January 1, 1989 and December 31, 2010 in the area covered by 7 registries of the French Network of Cancer registries (FRANCIM). The end of follow-up was June 30, 2013. This dataset was the one used to determine the theoretical parameters in the cervix uteri scenario (see section 3 of the paper). It included 5977 cervical cancer cases and 2139 (35.8%) deaths were observed within 5 years from diagnosis. Age at diagnosis ranged from 18 to 100 years (median: 49). More information about this dataset can be found in the works of Cowppli-bony and al.[1, 2]

The MPS and PH approaches were identical to those described in the simulation study (see sections 2.3.3, 2.3.4 of the paper). The age-standardized NS for a given year of diagnosis was also calculated as in the paper. We just recall that, for the MPS approach, the log-excess hazard was modelled as a function of time t, age a, and year of diagnosis y using a tensor product smooth which basis was built using restricted cubic splines of dimension 6, 5, and 4, respectively. The knot location of these splines was based on the empirical percentiles observed in the population of patients who died. The smoothing parameters were estimated using the REML criterion. For the PH approach, the excess hazard was modelled as

$$\log(h_E(t,a,y)) = f_t(t) + f_a(a) + f_y(y), \text{where}$$

f_t, f_a, and f_y were restricted cubic splines with the same features as the marginal bases of the MPS approach (same number and location of the knots). The 13 parameters of this PH model were obtained using maximum likelihood method (without any penalization).

We also replicated the analysis performed in Cowppli-bony and al,[1, 2] which is very typical of what has been done up-to-now in survival trends studies. In this study, NS was estimated using the non-parametric estimator of Pohar-Perme[3] (PP) and analysis was stratified by age-class (5 strata), and period of diagnosis (4 strata).

The resulting trends in age-standardized NS at 1 and 5 years are depicted in Figure S11. The MPS estimates are reasonably concordant with the PP estimates, whereas an unobserved increase in standardized NS at 5 years after year 2005 was obtained with the PH approach.

Figure S11. Standardized net survival at 1 and 5 years as a function of the year of diagnosis in Cervical cancer. Red solid curve: Multidimensional Penalized Splines approach (MPS); blue dashed curve: Proportional Hazard model (PH); gray segment: non-parametric estimation using the Pohar-Perme method with 95% CI (vertical bar).

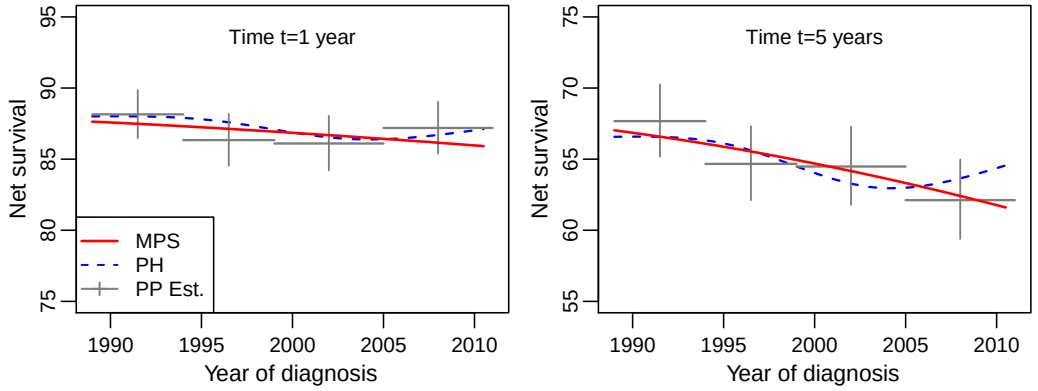


Figure S12. Net survival (NS) at 1 and 5 years as a function of the year of diagnosis in Cervical cancer, by age. The gray segments correspond to the estimates by period and age-class obtained with the Pohar-Perme method with 95% CI (vertical bar). Using the Multidimensional Penalized Splines approach (MPS; red solid curve) and the Proportional Hazard approach (PH; blue dashed curve), NS was estimated at 5 ages, each age corresponding to the median of age within each of the 5 age-classes.

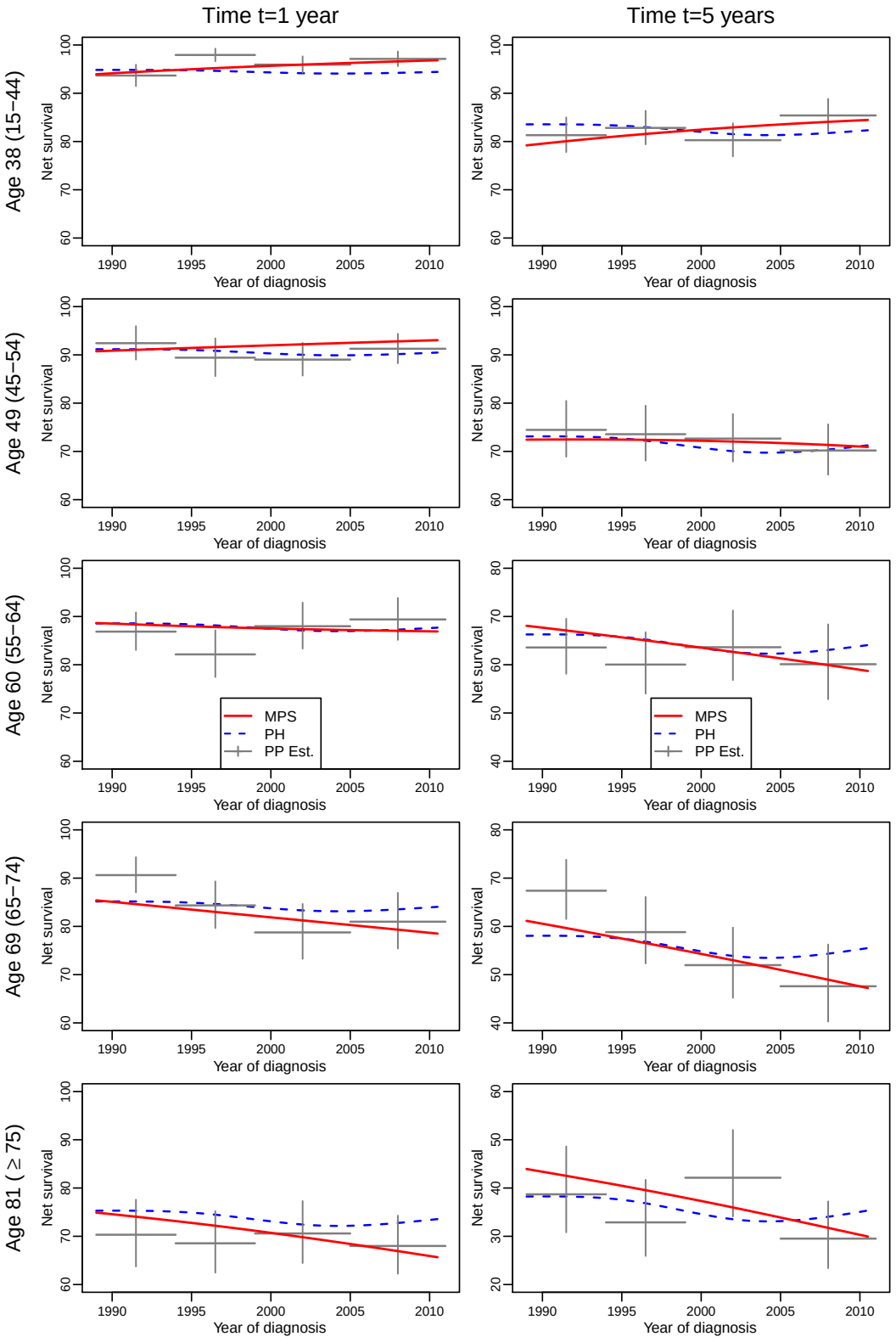
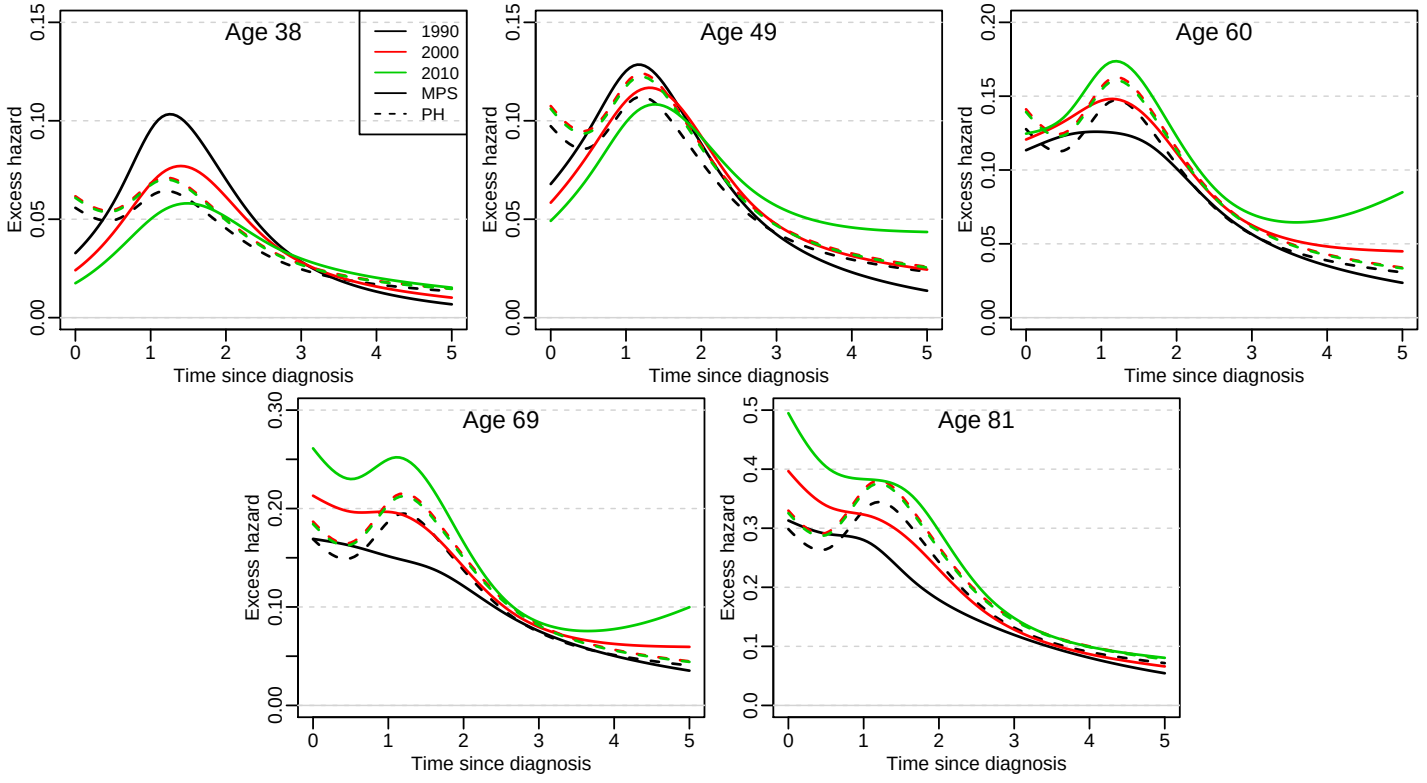


Figure S12 shows the corresponding trends by age-class (PP) or at the median age determined within each class (MPS and PH approaches). The MPS approach (red solid line) showed well distinct trends in survival at 1 and 5 years across ages, with an improvement observed in younger women and deterioration in older women. This pattern was overall confirmed by the PP estimates, although variability of these estimates led to somewhat erratic behaviors. As for the PH approach (blue dashed line), the pattern of trends in survival was inevitably similar whatever the time and the age because of the constraints induced by this model: survival decreased between years 1989 and circa 2004, then increased afterwards.

Figure S13 shows the dynamics of the excess hazard by age and year of diagnosis. The PH assumption and the absence of interaction (dashed curves) can clearly be seen in this graph; for example, the resulting excess hazard for y=2000 was higher than for y=1990 whatever the time and age. Conversely, the MPS approach provided a more complex picture of the dynamics of the excess hazard, exhibiting strong time-age-year interactions. So the dynamics were different according to age; excess hazard decreased regularly with time at older ages whereas it peaked around 1.5 years from diagnosis at younger ages. Furthermore, excess hazard increased with year of diagnosis for women aged 60 and over throughout the follow-up, while, in younger ages, it mainly decreased with years of diagnosis (this led to the different NS trends according to age seen in figure S12).

Figure S13 thus provides fundamentals medical results and this kind of figure is indispensable for clinicians and epidemiologists to help them understand the way medical practises have changed patient mortality over the year of diagnosis.

Figure S13. Excess mortality hazard as a function of time since diagnosis in Cervical cancer, at 5 ages. Solid curve: excess mortality hazard using the Multidimensional Penalized Splines approach; dashed curve: excess mortality hazard using the Proportional Hazard model.

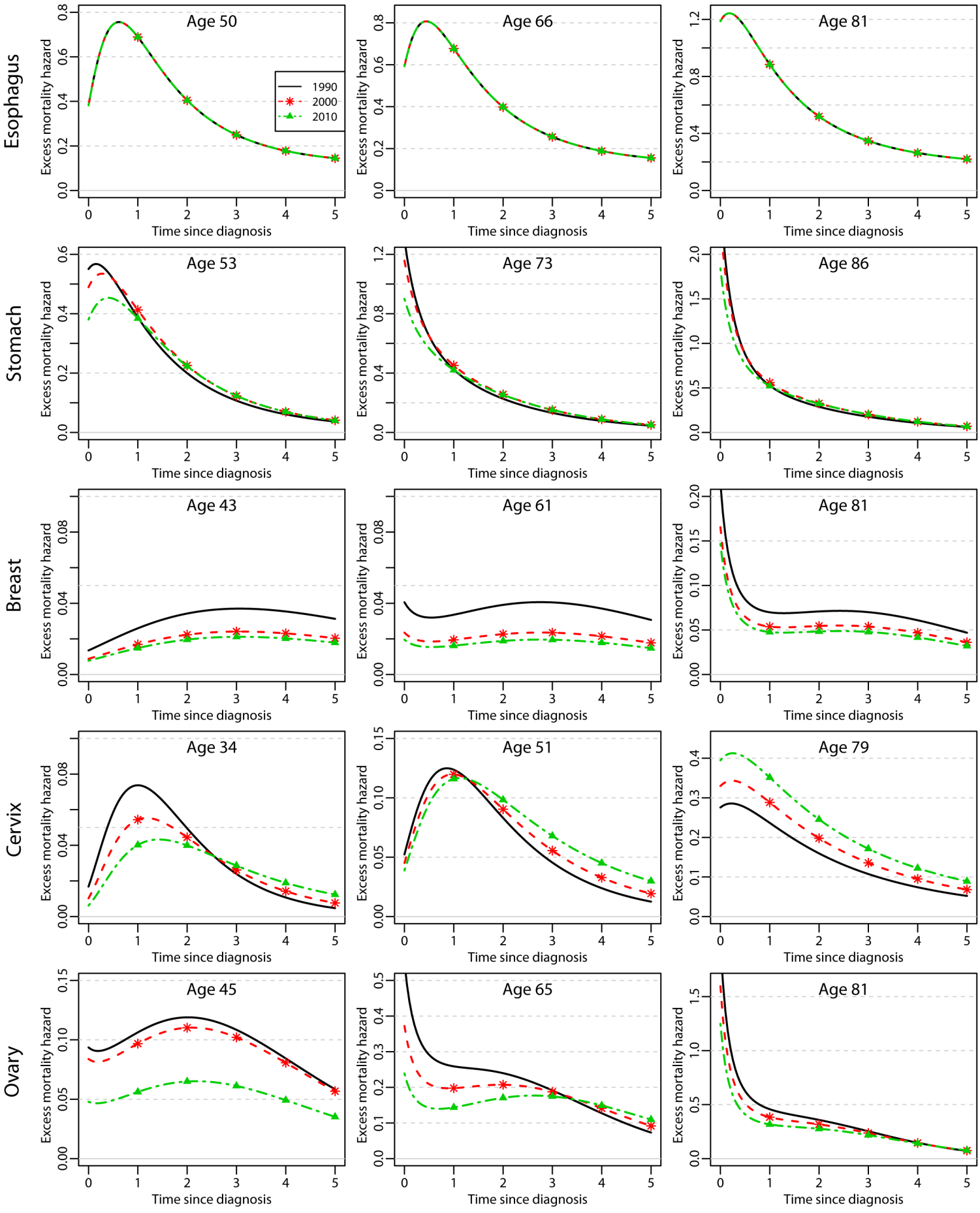


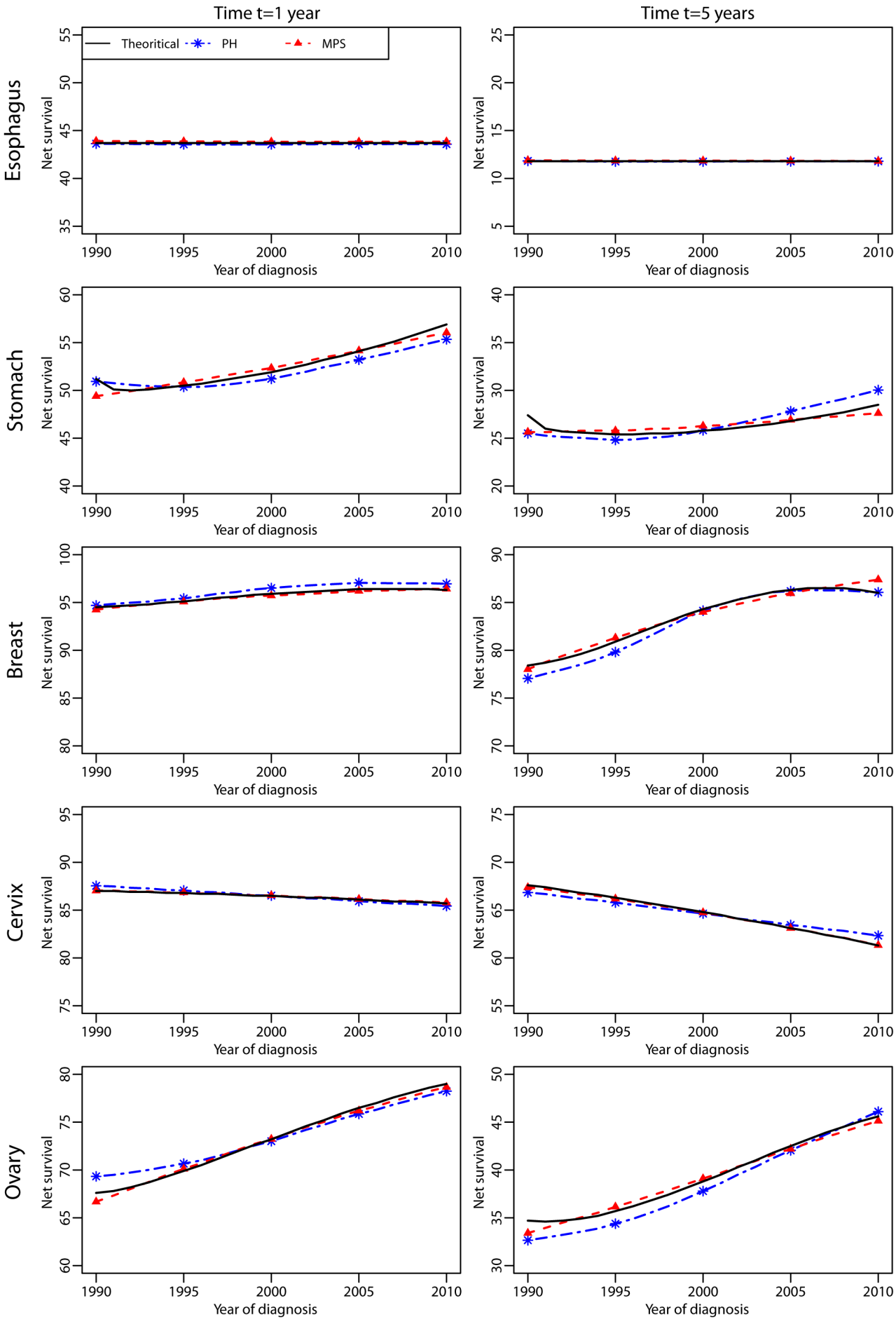
In our view, this example in cervical cancer illustrates the advantages of an efficient modelling approach, such as the MPS one, to study trends in survival and hazard. On one hand, both the degree of details and interpretation of the results are limited with stratified analyses based on PP estimator. On the other hand, the PH approach cannot describe properly the trends in survival or hazard whenever interactions are present. The MPS approach is an appealing alternative to us, as it is able to catch complex trends, but still provides smooth estimates.

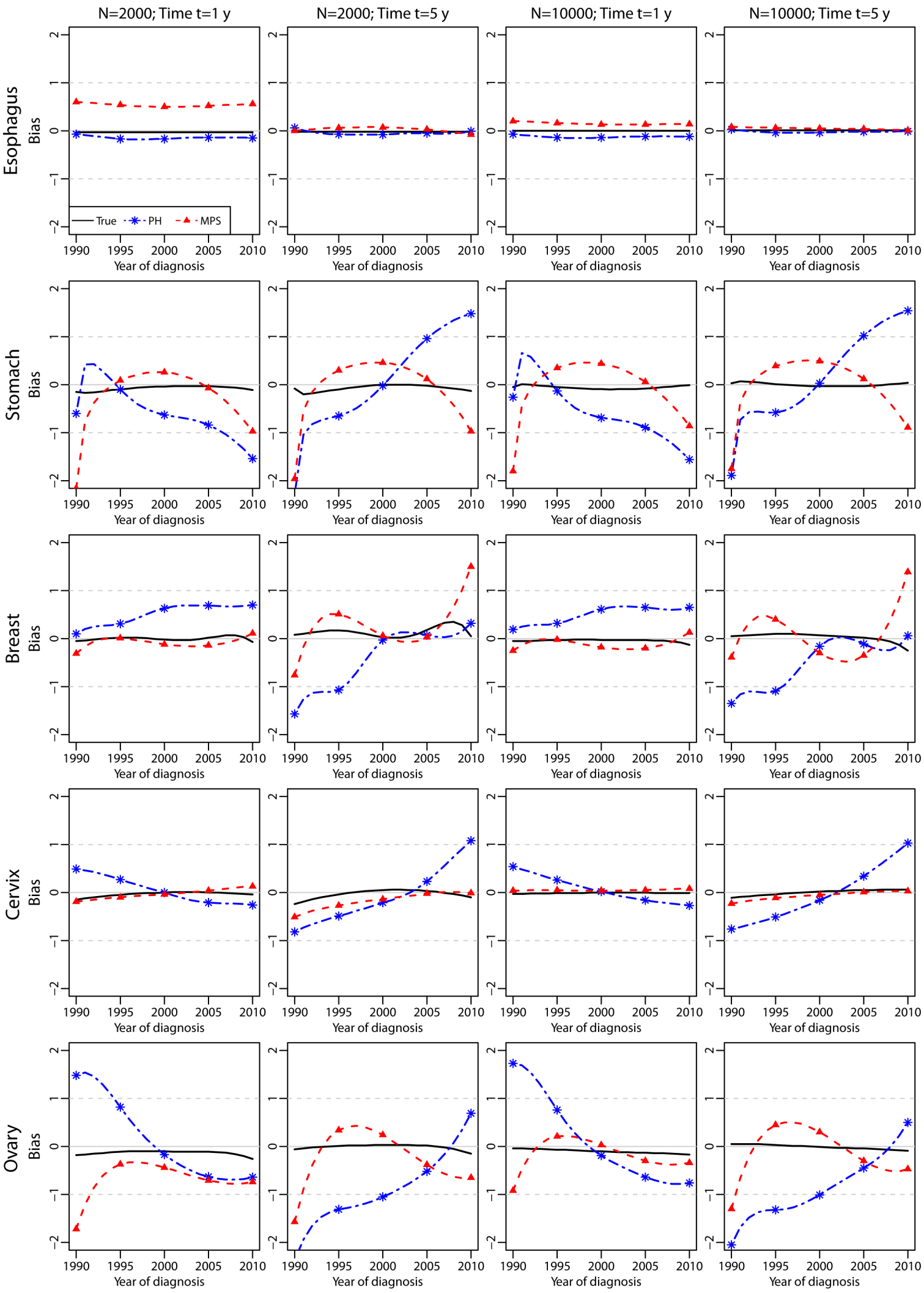
The R-code to reproduce this analysis is available on the GitHub repository **https://github.com/RocheLHCL/SMMR_Remontet2018** (Cf. the readme.pdf for explanations of the contents). However, due to copyright issues, we cannot provide the original real dataset. So, we provided one of the simulated dataset used in the simulation study on cervix uteri cancer data on 10,000 patients. The results may thus differ, to some extent, from those presented in the article.

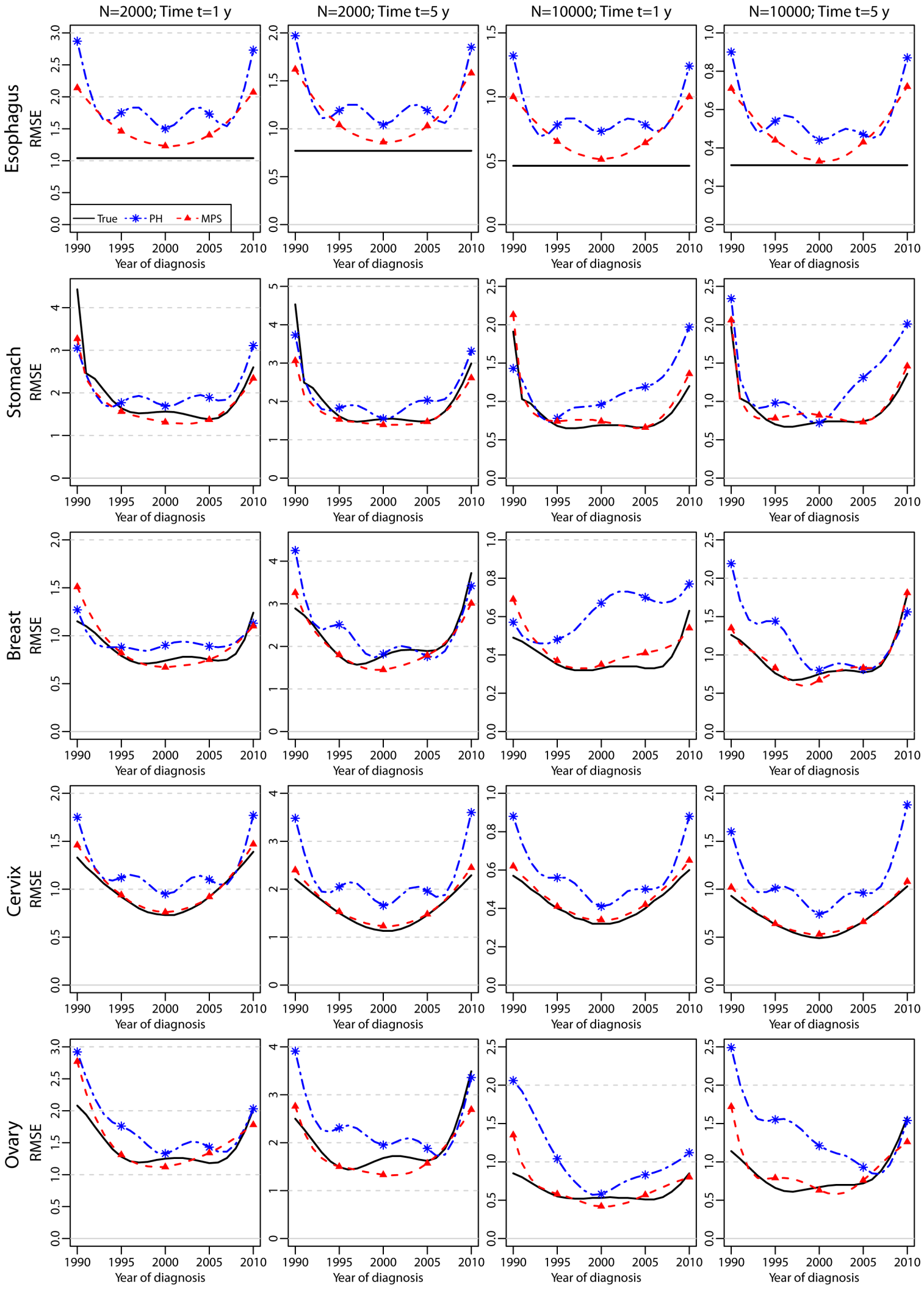
References

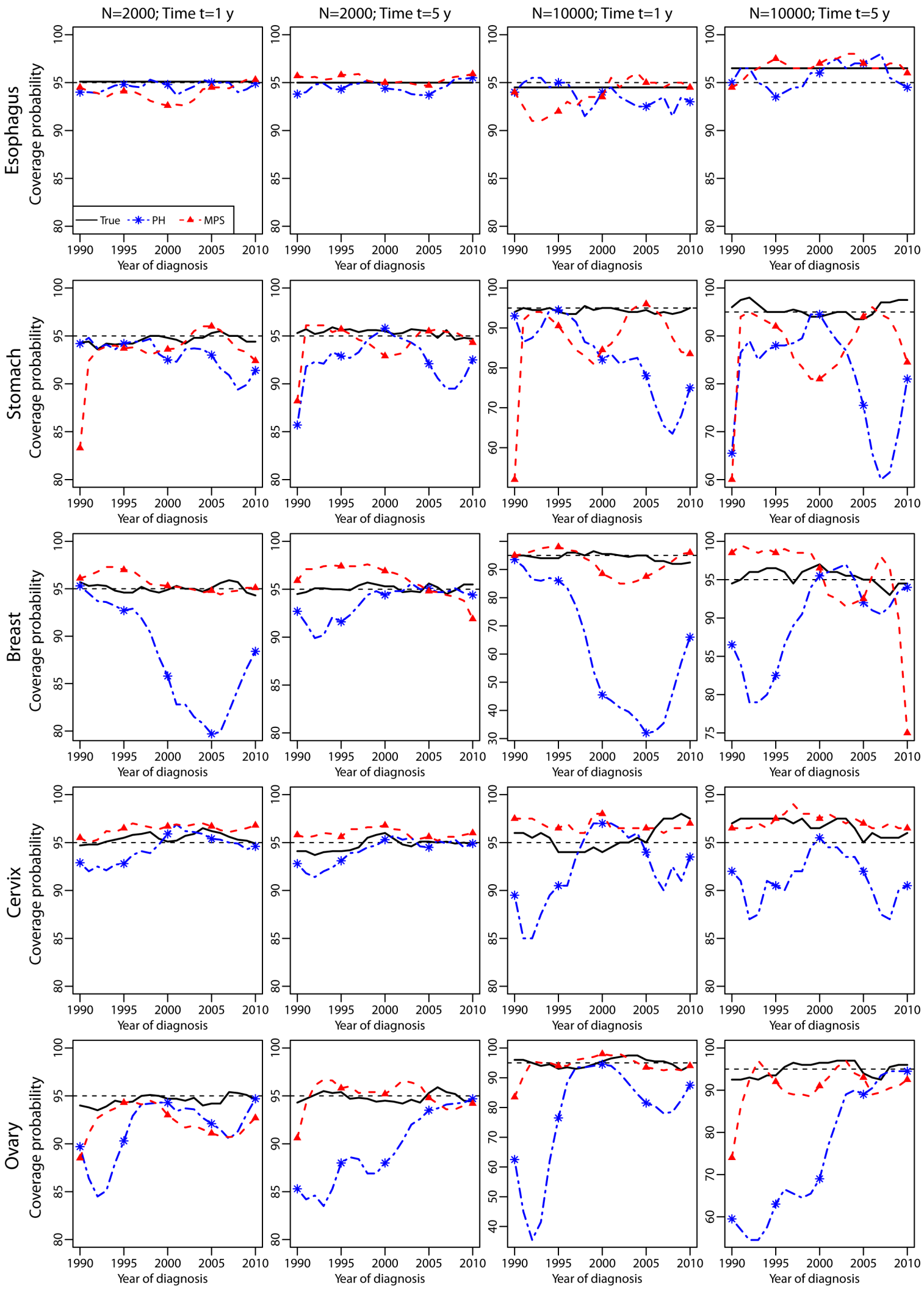
1. Cowppli-Bony A, Uhry Z, Remontet L, et al. Survival of solid cancer patients in France, 1989-2013: a population-based study. *Eur J Cancer Prev*. 2017; 26: 461-8.
2. Cowppli-Bony A, Uhry Z, Remontet L, et al. Survie des personnes atteintes de cancer en France métropolitaine, 1989-2013. Partie 1 - Tumeurs solides. Saint-Maurice: Institut de veille sanitaire, 2016. *http://invs.santepubliquefrance.fr/fr../layout/set/print/Publications-et-outils/Rapports-et-syntheses/Maladies-chroniques-et-traumatismes/2016/Survie-des-personnes-atteintes-de-cancer-en-France-metropolitaine-1989-2013-Partie-1-tumeurs-solides*
3. Perme MP, Stare J and Esteve J. On estimation in relative survival. *Biometrics*. 2012; 68: 113-20.


Esophagus
Stomach
Breast
Cervix
Ovary
Age 50
Age 66
Age 81
Age 53
Age 73
Age 86
Age 43
Age 61
Age 81
Age 34
Age 51
Age 79
Age 45
Age 65
Age 81
1990
2000
2010
Excess mortality hazard
Time since diagnosis


Time t=1 year
Time t=5 years
Theoritical
PH
MPS
Esophagus
Stomach
Breast
Cervix
Ovary
Net survival
Year of diagnosis
1990
1995
2000
2005
2010


N=2000; Time t=1 y
N=2000; Time t=5 y
N=10000; Time t=1 y
N=10000; Time t=5 y
Esophagus
Stomach
Breast
Cervix
Ovary
Bias
True
PH
MPS
Year of diagnosis
1990
1995
2000
2005
2010
−2
−1
0
1
2


N=2000; Time t=1 y
N=2000; Time t=5 y
N=10000; Time t=1 y
N=10000; Time t=5 y
Esophagus
Stomach
Breast
Cervix
Ovary
RMSE
True
PH
MPS
Year of diagnosis


N=2000; Time t=1 y
N=2000; Time t=5 y
N=10000; Time t=1 y
N=10000; Time t=5 y
Esophagus
Stomach
Breast
Cervix
Ovary
Coverage probability
Year of diagnosis
True
PH
MPS