



Article

No Excess Mortality up to 10 Years in Early Stages of Breast Cancer in Women Adherent to Oral Endocrine Therapy: A Probabilistic Graphical Modeling Approach

Ramon Clèries ^{1,2,3,*} , Maria Buxó ⁴ , Mireia Vilardell ⁵, Alberto Ameijide ⁶ , José Miguel Martínez ^{7,8}, Rebeca Font ^{1,2}, Rafael Marcos-Gragera ^{4,9,10,11,12} , Montse Puigdemont ^{9,10} , Gemma Viñas ¹³, Marià Carulla ⁶, Josep Alfons Espinàs ^{1,2}, Jaume Galceran ⁶, Ángel Izquierdo ^{9,10,13} and Josep Maria Borràs ^{1,2,3}

- ¹ Hospitalet de Llobregat Avenue Gran Vía 199-203, 08908 Barcelona, Spain; rfont@iconcologia.net (R.F.); ja.espinas@iconcologia.net (J.A.E.); jmborras@iconcologia.net (J.M.B.)
 - ² Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Hospitalet de Llobregat, Avenue Gran Via de l'Hospitalet, 199-203-1a Planta, 08908 Barcelona, Spain
 - ³ Department de Ciències clíniques, Universitat de Barcelona, 08907 Barcelona, Spain
 - ⁴ Institut d'Investigació Biomèdica de Girona (IDIBGI), C/Dr. Castany s/n, Edifici M2, Parc Hospitalari Martí i Julià, 17190 Girona, Spain; mbuxo@idibgi.org (M.B.); rmarcos@iconcologia.net (R.M.-G.)
 - ⁵ Independent Researcher, 08700 Barcelona, Spain; mvilardell@uoc.edu
 - ⁶ Registre de Càncer de Tarragona, Servei d'Epidemiologia i Prevenció del Càncer, Hospital Universitari Sant Joan de Reus (IISPV), 43204 Reus, Spain; alberto.ameijide@salutsantjoan.cat (A.A.); maria.carulla@salutsantjoan.cat (M.C.); jaume.galceran@salutsantjoan.cat (J.G.)
 - ⁷ Departament de Estadística i Investigació Operativa, Universitat Politècnica de Catalunya (EDIFICI H), Diagonal 647, 08028 Barcelona, Spain; jose.miguel.martinez-martinez@upc.edu
 - ⁸ Grupo de Investigación en Salud Pública, Universidad de Alicante, 03690 Alicante, Spain
 - ⁹ Registre de Càncer de Girona-Unitat d'Epidemiologia, Institut Català d'Oncologia, 17005 Girona, Spain; mpuigdemont@iconcologia.net (M.P.); aizquierdo@iconcologia.net (Á.I.)
 - ¹⁰ Grup d'Epidemiologia Descriptiva, Genètica i Prevenció del Càncer de Girona-IDIBGI, 17005 Girona, Spain
 - ¹¹ Facultat de Medicina, Universitat de Girona (UdG), 17071 Girona, Spain
 - ¹² Centro de Investigación Biomédica en Red, Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain
 - ¹³ Servei d'Oncologia Mèdica, Institut Català d'Oncologia, Hospital Universitari de Girona "Doctor Josep Trueta", 17005 Girona, Spain; Gvinyes@iconcologia.net
- * Correspondence: r.cleries@iconcologia.net; Tel.: +34-93-260-74-17



Citation: Clèries, R.; Buxó, M.; Vilardell, M.; Ameijide, A.; Martínez, J.M.; Font, R.; Marcos-Gragera, R.; Puigdemont, M.; Viñas, G.; Carulla, M.; et al. No Excess Mortality up to 10 Years in Early Stages of Breast Cancer in Women Adherent to Oral Endocrine Therapy: A Probabilistic Graphical Modeling Approach. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3605. <https://doi.org/10.3390/ijerph19063605>

Academic Editor:
Jon Øyvind Odland

Received: 17 February 2022

Accepted: 16 March 2022

Published: 18 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Breast cancer (BC) is globally the most frequent cancer in women. Adherence to endocrine therapy (ET) in hormone-receptor-positive BC patients is active and voluntary for the first five years after diagnosis. This study examines the impact of adherence to ET on 10-year excess mortality (EM) in patients diagnosed with Stages I to III BC (N = 2297). Since sample size is an issue for estimating age- and stage-specific survival indicators, we developed a method, ComSynSurData, for generating a large synthetic dataset (SynD) through probabilistic graphical modeling of the original cohort. We derived population-based survival indicators using a Bayesian relative survival model fitted to the SynD. Our modeling showed that hormone-receptor-positive BC patients diagnosed beyond 49 years of age at Stage I or beyond 59 years at Stage II do not have 10-year EM if they follow the prescribed ET regimen. This result calls for developing interventions to promote adherence to ET in patients with hormone receptor-positive BC and in turn improving cancer survival. The presented methodology here demonstrates the potential use of probabilistic graphical modeling for generating reliable synthetic datasets for validating population-based survival indicators when sample size is an issue.

Keywords: breast cancer; excess mortality; adherence; endocrine therapy; synthetic dataset; graphical modeling

1. Introduction

Breast cancer (BC) is the most common cancer and the leading cause of cancer death in European women [1]. A decrease in BC mortality is correlated with improvements in survival [2,3], an indicator of the success of cancer control efforts in a population-based setting. Conditional five-year survival is an outcome that measures the efficacy of cancer management, since it responds to the question of “once a patient survives for T years, what is the probability of surviving another five years?” [4]. Most population-based cancer survival indicators are derived from relative survival (RS), defined as the ratio between the overall survival (OS) and expected survival of the cohort with respect to the general population [5]. RS is as an estimate of the patients’ cancer-specific survival compared to the survival of the general population, and one can also assess the conditional RS(CRS) at five additional years after surviving T years [5]. On the basis of the CRS(T), one can determine the five-year excess mortality as $EM(T) = 1 - CRS(T)$, which is used to assess whether patient mortality surpasses the mortality of the general population, that is, when $EM(T) > 0$ [6].

These conditional survival or mortality indicators provide very relevant information on the prognosis of BC over time, as they are a starting point to identify prognostic factors related to long-term survival [6–9]. For instance, the BC cohort’s mortality is not different from the general population’s mortality when EM equals 0 beyond a certain time interval T [6]. Moreover, population-based cancer registries can define the time to cure of cancer as “the number of years after cancer diagnosis when the EM, expressed as a percentage, becomes negligible” [4,8]. That situation occurs when the EM remains clearly below 5% for more than 10 years, and CRS consequently surpasses 95% [8]. A recent study using European cancer registry data showed that an EM of 5% could persist for at least 15 years in BC patients [9]. However, the EM in that study was an overall indicator that could only be adjusted for age because other prognostic factors could not be retrieved from all participating cancer registries.

Stage, molecular subtype, and adherence to endocrine therapy (ET) are key predictors for providing population-based BC survival estimates [10]. Indeed, tamoxifen and aromatase inhibitors are pillars of adjuvant therapy for patients with hormone receptor positive (HR+) BC diagnosed at Stages I–III [11]. Randomized clinical trials showed that five years of adherence to ET positively impact BC survival [11]. In a previous study, we found that nonadherence to ET is significantly and independently associated with recurrence and all-cause mortality at Stages I–III of hormone receptor positive BC after adjusting for age [12]. A question arises regarding the impact of ET adherence on long-term survival and risk of death in patients with BC versus the general population [9].

The sample size of the cohort could be an issue when trying to estimate age-specific survival according to stage and molecular subtype; however, generating a large cohort of simulated survival data on the basis of observed cohort data could help overcome this limitation [13]. This simulation could be achieved in two ways: (1) only simulating survival times [14,15] or (2) generating a set of cohort covariates as a function of survival times [13,16]. For the latter, oversampling techniques such as SMOTE [17], Borderline SMOTE [18], and MWMOTE [19] can also be used to generate balanced subsets of data, where the efficiency of these methods in simulating new datasets must be assessed with the observed survival patterns of real data [19]. However, if we are interested in detecting new patterns of survival, the specific modeling of probabilistic dependencies between the variables of the observed data is needed, which requires estimating a joint probability distribution of the variables [13,20–22]. For that purpose, our research team developed Modelling Graphical Probabilistic Dependencies (ModGraProDep) and suggested that future work should be oriented toward selecting data subsets across several synthetic datasets (SynD) that better mimic the cohort’s survival pattern [13].

In the present study, we developed a method to validate the survival estimates of the original cohort by using a synthetic cohort that combines the “best” subsets of simulated data derived from graphical models. Survival indicators are generated by fitting the cohort data and the simulated SynD to a Bayesian RS model developed for that purpose.

2. Materials and Methods

2.1. Data: BCStage Dataset

BC data were obtained from the population-based cancer registries of Girona and Tarragona (northeastern Spain) covering an average annual population of 560,120 women from 2005 to 2009 [23]. During this time period, 4053 women under the age of 75 years were diagnosed with invasive BC (code C50 of the 10th edition of the International Classification of Diseases, ICD-10). A total of 352 women (8.7%) were excluded from the analyses due to missing data on estrogen and progesterone status, and another 1215 (30.0%) were excluded due to missing data on stage, Stage IV at diagnosis, or diagnosis of HER2-enriched or triple-negative BC tumors, and we could not retrieve follow-up status (if the patient died or not at the end of follow-up) in $N = 189$. Each woman with BC diagnosed from 2005 to 2009 was followed up to 31 December 2019; we considered a maximal follow-up of 10 years. Of the patients eligible for ET ($N = 2297$), information could only be retrieved for BC patients diagnosed from 2007 to 2009 who met the inclusion criteria: patients presenting positivity for estrogen and/or progesterone receptors diagnosed at Stages I, II, or III, who were eligible for ET ($N = 1243$). Survival times for patients not found to be dead at the end of follow-up were censored. Stage classification was based on the TNM classification system, as described in the 6th edition of the American Joint Committee on Cancer staging manual [24], classifying patients at Stage I, II, or III when TNM was available at the moment of diagnosis.

Adherence to ET for patients with HR+ BCs was tracked during the first five years after BC diagnosis. Any switch to tamoxifen or aromatase inhibitor was considered to be a continuation of treatment. Adherence was estimated as “the proportion of days covered by a filled drug prescription over the treatment period (up to five years from the date of first prescription)”, deeming a cumulative adherence rate of 80% or more as satisfactory [12]. Data on ET prescription refills for BC were collected for the entire study period (2007–2015) from the community pharmacy database, which is mandatory for drug reimbursement in Catalonia.

Collected variables were: age (26, 27, . . . , 73, 74), stage at diagnosis (I, II, or III), adherence to ET (yes: adherence rate > 80% vs. no: adherence rate \leq 80%), follow-up years (1, . . . , 10) and exitus (died vs. survived). Age was also considered to be a categorical variable with three age groups: ≤ 49 , 50–59, and 60–74 years. Patients were additionally classified according to the tumor positivity of the human epidermal growth factor receptor (HER2) expression (HER2+ vs. HER2–).

2.2. Synthetic Data Simulation

2.2.1. Fitting Graphical Models through ModGraProDep

Four synthetic datasets were simulated by modeling the probabilistic dependencies between variables using ModGraProDep [13]. In brief, let Γ be the set of cells in a contingency table, where c_{ash} is a cell of the table with indices $a(\text{age}) - s(\text{stage}) - h(\text{adherence})$. Let $p(c_{ash})$ be the cell probabilities of the contingency table Γ . Using a hierarchical expansion of $\log(p(c_{ash}))$ we considered a saturated log-linear model, a model including the main effects, and all interactions between these, that is

$$\log(p(C_{ash})) = \alpha + \beta_a + \beta_s + \beta_h + \gamma_2^I + \gamma_3^I \quad (1)$$

where parameter α is an intercept, β refers to main effects, and γ_q^I refers to the set of interaction parameters of order q , where $q \in \{2, 3\}$. We can also specify a model with fewer interaction terms by setting higher-order interaction to zero.

Assuming that there is a set of candidate models $M(j) | j \in \{1, \dots, J\}$, ModGraProDep uses a heuristic search based on penalized log-likelihood

$$H(j,k) = -2\log(p(c_{ash})) + k * z(j) \quad (2)$$

where $z(j)$ is the number of model parameters, and k is a penalty factor. Changing the value of k can result in several models using backward stepwise elimination of graph arches. Starting from the saturated model, ModGraProDep fits four models: three by using the k penalty factor, GMk1 for $k = 1$, GMAIC for $k = 2$ (Akaike information criterion [25]), and GMBIC for $k = \log(N)$ (Bayesian information criterion [25]), and another by testing the arch’s conditional independence, GMTEST. Once these four models had been fitted, we first imputed adherence in the BC cases with missing adherence, and then generated the synthetic datasets. We used the junction-tree simulation algorithm implemented in ModGraProDep for simulating four datasets of size $N = 1,000,000$ from each of the four models and according to the probabilistic relationships between variables (see Vilardell et al. (2020) for technical details [13]).

2.2.2. ComSynSurData: Combining Synthetic Survival Datasets

Figure 1 presents the scheme for generating a combined synthetic dataset that selects the best subsets of data that better mimic the survival pattern of the cohort. These are summarized as follows:

Step 0. Use ModGraProDep for generating the four SynDs.

Step 1. Produce a partition of the cohort dataset into L subsets according to A age groups and S levels of a stratification variable, such as stage at diagnosis; then, $L = A \times S$. For instance, if strata were stage at diagnosis with levels {I, II, III}, and three age groups were considered, then $L = 3 \times 3 = 9$ subsets (one for each age group and stage combination). In the same line, the same partition is made for each SynD.

Step 2. For each of the L subsets of the cohort data, find its “best” counterpart among the $4 \times 9 = 36$ subsets of SynDs by comparing survival estimates between the observed cohort and that derived from the SynDs through a scoring method.

Step 3. Once L subsets of SynD are selected in each age stratum, generate a combined synthetic cohort by merging these L subsets, from which Kaplan–Meier survival estimates according to stage and corresponding age groups can be derived.

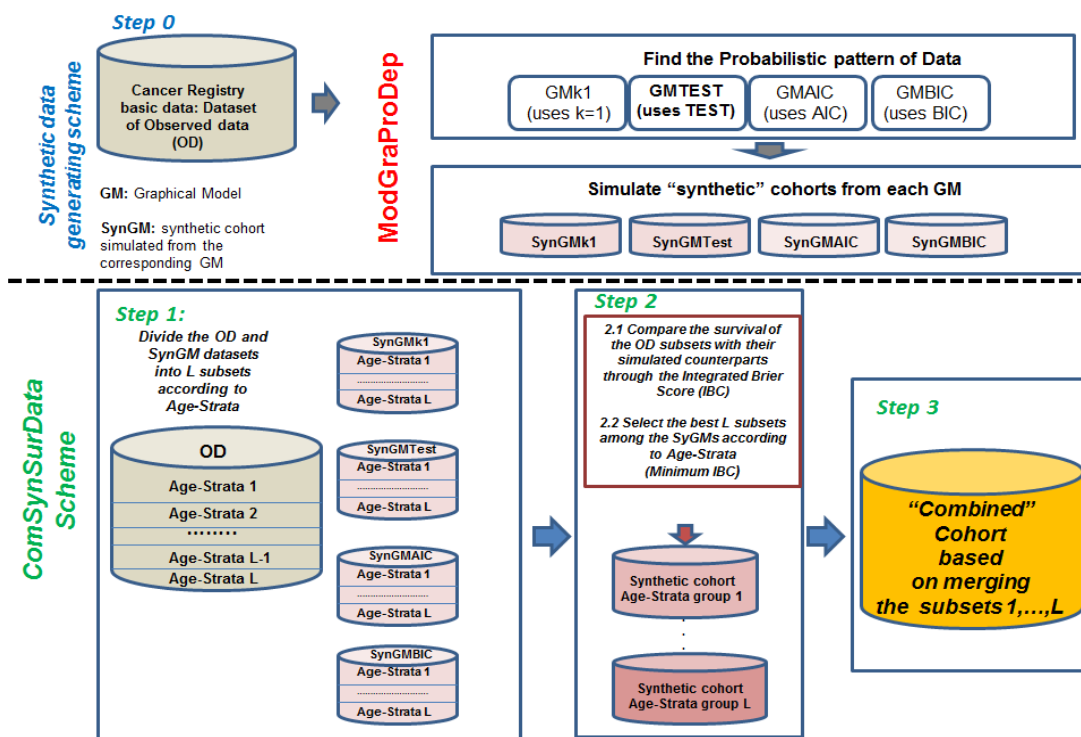


Figure 1. Scheme of procedure for generating combined cohort by using best synthetic cohort for each of the considered L age-stratum groups. Synthetic cohorts generated according to ModGraProDep.

2.2.3. Scoring Method for Comparing Observed versus Predicted Survival in Step 2

ComSynSurData uses the integrated Brier score (IBS), a scoring method to detect inaccuracies in the prognostic classification scheme, that is, disagreement between the survival curves of cohort and simulated data at a certain time T [26]. Let $\hat{S}(T)$ be the predicted survival function, and $\hat{G}(T)$ the censoring distribution, both functions estimated using the Kaplan–Meier method and using the SynD. Here, we used the following definition of the Brier score at time T for censored data [27]:

$$BS^C(t) = \frac{1}{n} \left[\sum_{i=1}^n \left[\frac{(0 - \hat{S}(t))^2}{\hat{G}(t)} I(t_i \leq t) + \frac{(1 - \hat{S}(t))^2}{\hat{G}(t)} I(t_i > t) \right] \right] \quad (3)$$

where t_i is the follow-up of the i -th patient in the cohort, and $I(\cdot)$ are indicator functions, such that $I(t_i \leq t) = 1$ and $I(t_i > t) = 0$ if the i -th patient dies before t , and $I(t_i > t) = 1$ and $I(t_i \leq t) = 0$ if the i -th patient does not die before t .

IBS is an overall measure up to a certain time target t^* , which uses weights defined as $W(t) = t/t^*$ [27]. Here, we used maximal follow-up $t^* = 10$ years. The IBS was calculated as

$$IBS^C(t^*) = \int_0^{t^*} BS^C(u) dW(u) = \frac{1}{10} \int_0^{t^*} BS^C(t) dt \quad (4)$$

For each age and stage stratum, the selected subset of *SynD* would be that with the smallest IBS score, which could lie between 0 and 1, where IBS = 0 shows a perfect match between observed and predicted survival [26,27]. The Supplementary Material file includes the R code for running *ComSynSurData*.

2.3. Statistical Modeling of Excess Mortality

We used an RS model to derive the survival indicators. Let $\lambda_O(T)$ be the overall hazard of death in the cohort at a specific time T , and $\lambda_P(T)$ is the expected hazard in the cohort using the general population mortality [28]. Applying additive modeling, the excess hazard of death in the cohort due to BC is $\lambda_X(T) = \lambda_O(T) - \lambda_P(T)$ [29], where $OS(T) = \int_0^T \exp(-\lambda_O(t)) dt$ is the observed survival in the cohort at time T , and $ES(T)$ its expected survival in the cohort, $ES(T) = \int_0^T \exp(-\lambda_P(t)) dt$. Relative survival (RS) at time t is calculated as [28]:

$$RS(T) = \frac{OS(T)}{ES(T)} \quad (5)$$

$RS(T)$ could reach (or even surpass) 1 when $OS(T)$ is equal to the survival of the general population [28]. From $RS(T)$, one can derive the five-year conditional relative survival at T years of follow-up as [5]

$$CRS(T) = \frac{RS(T+5)}{RS(T)} \quad (6)$$

From this, the five-year conditional excess mortality (EM) at T years of follow-up [5,6].

$$EM(T) = 1 - CRS(T) \quad (7)$$

Using (7), one can assess temporal changes in the *EM* by monitoring this quantity during follow-up [5]. Moreover, it is of interest for both the patient and clinician to estimate the probability of death due to cancer in the presence of other causes at time T , $PCa(T)$ and the crude probability of death due to other causes in the presence of cancer mortality at time T , $POC(T)$ [6]. These quantities can be derived from the $RS(T)$ by using competing risks modeling as

$$PCa(T) = \int_0^T OS(u) \lambda_X(u) du \quad (8)$$

$$POC(T) = \int_0^T OS(u)\lambda_P(u)du \quad (9)$$

where the sum of these two probabilities gives the probability of death from any cause at time T [6]. Since all these indicators are related to $\lambda_X(t)$ and $\lambda_O(t)$, these last two risks can be estimated by $\hat{\lambda}_O(T) = O(T)/Y(T)$ and $\hat{\lambda}_P(T) = E(T)/Y(T)$, where $O(T)$ is the observed number of deaths at T and $E(T)$ is the expected number of deaths at T , which is calculated from applying the age-specific mortality rates of the general population to each one of the individuals at risk within the T interval, and finally, $Y(T)$ is the number of individuals at risk in T .

Since $O(T)$ is usually considered to be a Poisson-distributed random variable with mean μ_T , we used a Bayesian autoregressive modeling of order 1 to estimate $\hat{\lambda}_O$, assuming a prior precision (inverse of variance) of 0.001 [30], defined as

$$O(T) \sim \text{Poisson}(\mu_T) \log(\mu_1) = \log(Y(1)) + \delta_1 \quad \delta_1 \sim N(0, 0.001) \quad (10)$$

$$\log(\mu_T) = \log(Y(T)) + \delta_T | T > 1 \quad \delta_T \sim N(\delta_{T-1}, 0.001) | T > 1$$

Posterior distributions and the corresponding 95% credible intervals of aforementioned survival indicators (5)–(9) were calculated through posterior estimates of μ_T , and fixed quantities $E(T)$ and $Y(T)$. The model was implemented using WinBUGS [31] (see the program code in supplementary material file), which was run within R (<http://www.R-project.org>, accessed on 5 December 2021) through the R2WinBUGS library [32].

2.4. Analysis Scheme

First, the GM was fitted to the original dataset, and adherence was imputed in cases with missing information. Second, four SynDs were generated using ModGraProDep, and from these SynDs, ComSynSurData selected the best L age-stage subsets of synthetic data that were used to generate the combined synthetic dataset. Survival indicators were derived from fitting the Bayesian relative survival model to this combined cohort, and these were also validated with those obtained using the original cohort. Lastly, age-specific survival indicators for epidemiologic or clinical use were calculated.

3. Results

Table 1 presents the clinical and pathological characteristics of the observed cohort in Girona and Tarragona in 2005–2009, stratified according to HER2+/HER2– expression. Main differences were detected in the distribution of BC stage: stages II and III were more frequent in patients with HER2+ compared to HER2– tumors. Mean age at diagnosis was 55.3 years: 32.7% of the patients were diagnosed with BC before 50 years of age, 29.6% were diagnosed at age 50 to 59 years, and 37.7% were 60 years or older. Most patients were diagnosed at early stages, whereas only 17.5% were diagnosed at Stage III. Mean follow-up was 8.2 years, and 11.7% of patients died during that period. Of these, information about adherence could be retrieved in those diagnosed from 2007 to 2009 ($N = 1243$), 75% of whom showed a cumulative adherence rate of 80% or higher during the first five years after the BC diagnosis. In cases with missing adherence data, a value for adherence was imputed making use of ModGraProDep.

Table 1 also shows the distribution of the number of BC cases according to adherence and HER2 status after the imputation of these four models. We did not find any difference in the distribution of the percentages according to adherence status when comparing the observed frequencies in the cohort (the $N = 1243$ BC patients) with those obtained after using each of the four models implemented in ModGraProDep (see Table 1, *Distribution of BC cases in the cohort after the imputation of adherence to ET when missing*). However, the distribution of adherence status in the cohort was identical when GMAIC and GMBIC models were used, indicating that the probabilistic graphical pattern of the dependencies between variables in the observed data ($N = 1243$) was likely to be identical when fitting these two graphical models to the cohort data.

Table 1. Characteristics of patients diagnosed with breast cancer from 2005 to 2009 in Girona and Tarragona. Of the 2297 BC patients, complete data for endocrine treatment (ET) were available for 1243 in 2007–2009. Imputation of adherence through ModGraProDep was performed for the remaining 1054 BC patients.

		HER2– (N = 1736; 75.6%)	HER2+ (561; 24.4%)	Total (N = 2297; 100%)
Registry	Girona	876 (50.5%)	301 (53.6%)	1176 (51.2%)
	Tarragona	860 (49.5%)	260 (46.4%)	1121 (48.8%)
Age	Mean (SD)	55.6 (10.6)	54.3 (10.7)	55.3 (10.6%)
	≤49 years	556 (32.0%)	196 (35.0%)	751 (32.7%)
	50–59 years	502 (28.9%)	178 (31.7%)	680 (29.6%)
	60–74 years	678 (39.1%)	187 (33.3%)	866 (37.7%)
Stage at diagnosis	I	769 (44.3%)	195 (34.7%)	997 (43.4%)
	II	641 (36.9%)	257 (45.9%)	898 (39.1%)
	III	326 (18.8%)	109 (19.5%)	402 (17.5%)
Deceased (%)		11.9	10.9	11.7
Follow-up in years, mean (SD)		9.2 (1.7)	9.3 (1.5)	9.2 (1.6)
Adherence to ET	No: ≤80%	234 (13.4%; 24.9% ^b)	75 (13.5%; 24.8% ^b)	309 (13.5%; 24.9% ^b)
	Yes: >80%	706 (40.7%; 75.1% ^b)	228 (40.6%; 75.2% ^b)	934 (40.6%; 75.1% ^b)
	Total ^a	940 (54.1%; 100.0% ^b)	303 (54.0%; 100.0% ^b)	1243 (54.1%; 100.0% ^b)
	Missing ^c	796 (45.9%; -)	258 (45.9%; -)	1054 (45.9%; -)
Distribution of BC Cases in Cohort after Imputation of Adherence to ET when Missing				
	Adherence to ET	HER2– (N = 1736; 75.6%)	HER2+ (N = 561; 24.4%)	Total (N = 2297; 100%)
GMK1^d	No: ≤80%	426 (24.5%)	126 (22.4%)	552 (24.0%)
	Yes: >80%	1310 (74.5%)	435 (77.6%)	1745 (76.0%)
	Total	1736 (100.0%)	561(100.0%)	2297 (100.0%)
GMAIC^e	No: ≤80%	420 (24.2%)	120 (21.4%)	540 (23.5%)
	Yes: >80%	1316 (74.8%)	441 (78.6%)	1745 (76.5%)
	Total	1736 (100.0%)	561(100.0%)	2297 (100.0%)
GMBIC^f	No: ≤80%	420 (24.2%)	120 (21.4%)	540 (23.5%)
	Yes: >80%	1316 (74.8%)	441 (78.6%)	1745 (76.5%)
	Total	1736 (100.0%)	561(100.0%)	2297 (100.0%)
GMTEST^g	No: ≤80%	424 (24.4%)	121 (21.5%)	545 (23.7%)
	Yes: >80%	1312 (74.6%)	440 (78.5%)	1752 (76.3%)
	Total	1736 (100.0%)	561(100.0%)	2297 (100.0%)

^a Cases with available information on endocrine therapy in 2007–2009, N = 1243; ^b percentage with respect to ^a; ^c cases with no available information on endocrine therapy; ^{d–g} distribution of cases according to adherence, imputing adherence status in BC cases with missing information by applying ModGraProDep models.

Figure 2 shows the graphical modeling of the data, which encodes a factorization of the joint probability distribution of the dataset. Three probabilistic schemes can be distinguished: one obtained using GMK1 (Figure 2a), another using GMTEST (Figure 2b), and another, as noted above, obtained through GMAIC and GMBIC (Figure 2c). Figure 2a shows that the model GMK1 considered that all variables were related (connected). The GMTEST model considers age as related to exitus, but this is conditional on adherence or stage at BC diagnosis, and HER2 as directly related to the other variables through stage. Lastly, GMAIC and GMBIC models consider that age could be independent from the data structure, and all remaining variables are conditionally independent once exitus is known. Stage was related to the remaining variables, conditional on others, regardless of the model used.

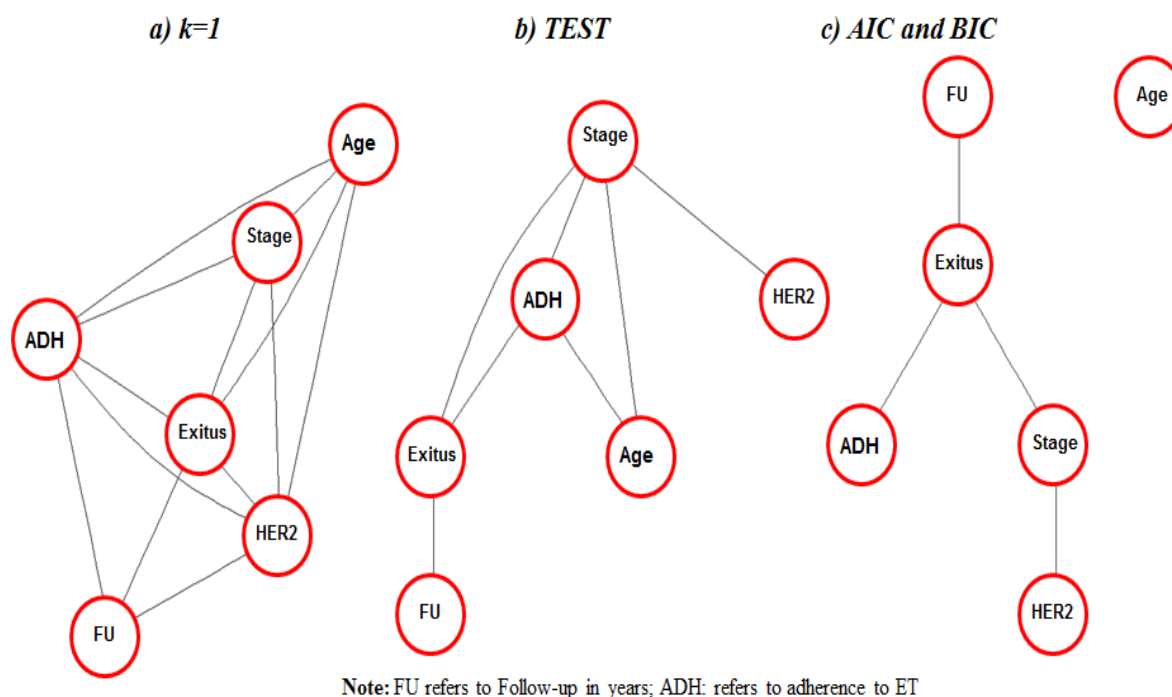


Figure 2. Undirected acyclic graphs generated from fitting the best graphical models to observed data ($N = 1243$) using different criteria: (a) GMK1 model: k -penalty factor of penalized log-likelihood set to 1; (b) GMTEST: testing for statistical significance of arches; (c) GMAIC: Akaike information criterion (AIC) and GMBIC: Bayesian information criterion (BIC).

3.1. Data Simulation

After the imputation of the missing data, ModGraProDep was used for simulating the four SynDs, and from these, ComSynSurData was applied to generate the combined dataset. Four datasets were considered, and on their basis, four SynDs were simulated. Once these models were fitted, the four SynDs were introduced into ComSynSurData, and the combined dataset was generated. Table 2 shows the matrix of internally generated IBS scores by ComSynSurData from which to select the $L = 9$ subsets. From these, seven data subsets were selected from the SynD dataset derived from GMK1, two from the SynDs derived from GMAIC and GMBIC, and none from GMTEST.

3.2. Comparing Observed Survival in the Cohort with Survival in the Combined Cohort

To assess the reliability of these simulated datasets, OS in the cohort with real data ($N = 1243$) was compared with the estimated survival using the combined cohort (Figure 3). Using the posterior distribution of the survival derived from the combined cohort, its median survival overlapped with the 95% credible intervals of observed survival in the original cohort in almost all age groups. In some, however, the median of the survival's combined cohort was slightly lower than the observed survival, but close to the lower bound of the 95% credible interval of the survival in the original cohort: age group ≤ 49 years at Stages I and II, and for the age group of 59–74 years at Stage III.

Table 2. Integrated Brier score at up to 10 years of follow-up by age and stage, comparing the cohort’s absolute survival with the absolute survival estimated using each one of the synthetic datasets derived from the Graphical Models (in bold: minimal integrated Brier score for each age group according to stage of breast cancer at diagnosis).

	Synthetic Dataset			
	<i>Derived from GMk1</i>	<i>Derived from GMTesT</i>	<i>Derived from GMAIC</i>	<i>Derived from GMBIC</i>
Stage I				
≤49 years	0.0149	0.0146	0.0142	0.0143
50–59 years	0.0432	0.0433	0.0433	0.0433
60–74 years	0.0471	0.0485	0.0482	0.0482
Stage II				
≤49 years	0.0484	0.0486	0.0485	0.0485
50–59 years	0.0703	0.0707	0.0706	0.0706
60–74 years	0.0943	0.0986	0.0982	0.0984
Stage III				
≤49 years	0.1183	0.1188	0.1185	0.1185
50–59 years	0.1437	0.1431	0.1427	0.1426
60–74 years	0.1722	0.2020	0.1960	0.1965

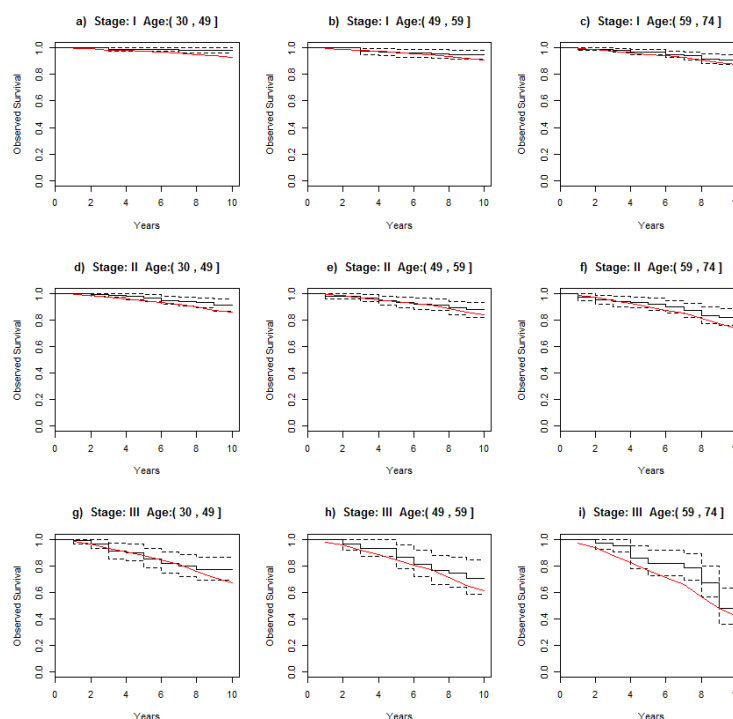


Figure 3. Comparison of 95% credible interval of observed survival derived from original cohort (black) and median survival (red) of combined cohort across stages at diagnosis and stratified by age group.

3.3. Survival Indicators Derived from Combined Dataset

Figure 4 compares the EM observed in the original cohort with that estimated using the combined dataset. Median EM between these datasets did not differ, since the 95% credible intervals derived from the observed cohort overlapped with the estimates derived from the combined cohort. In this line, the patients diagnosed in stages I and II who were adherent to endocrine therapy did not show EM with respect to the general population. However, we found that patients diagnosed in these early stages who were not adherent to

ET had an EM with a median ranging from 5% to 10%, which usually suggests a significant EM. For patients diagnosed at Stage III, the effect of nonadherence to ET might double the EM with respect to adherence.

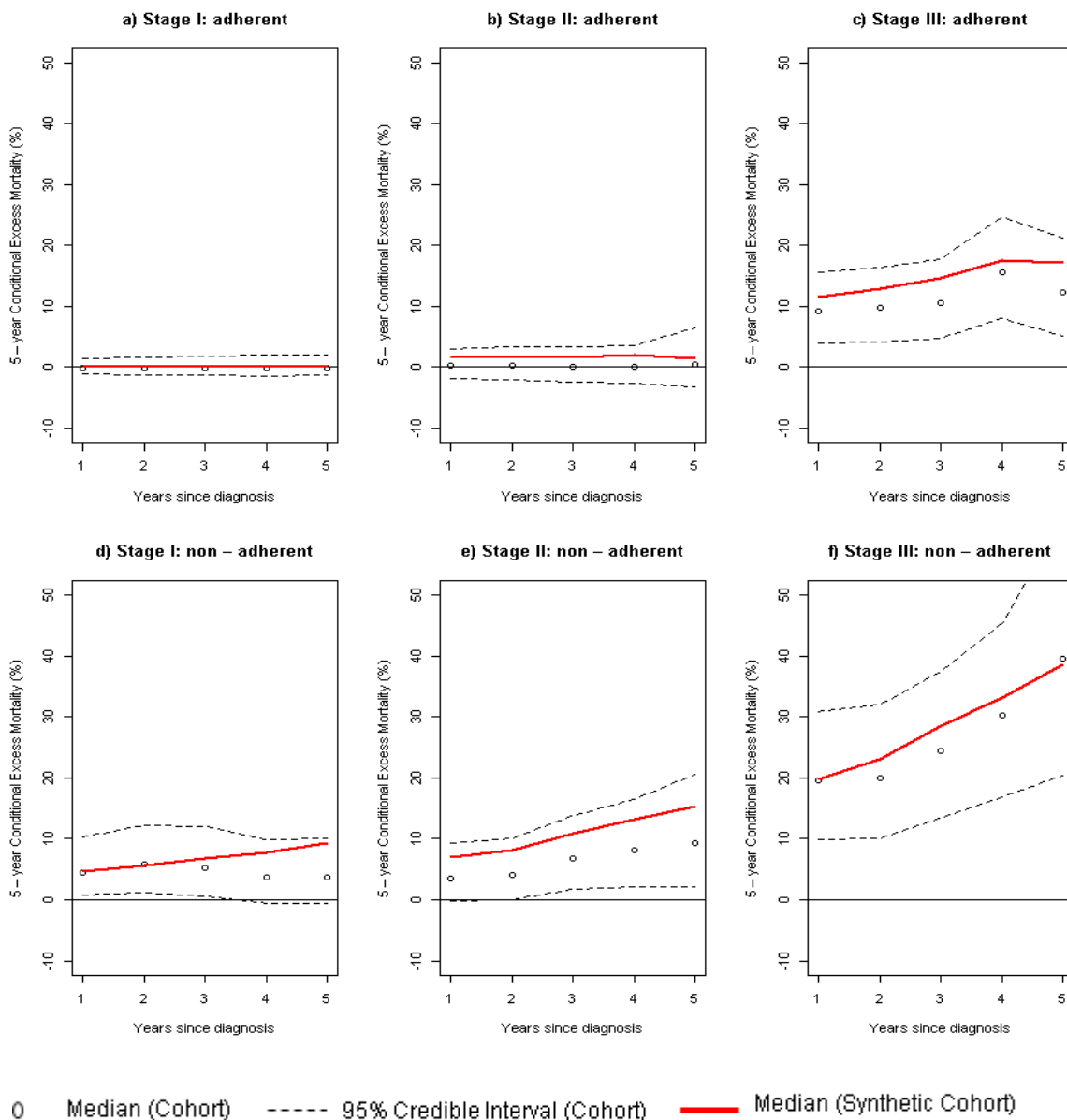


Figure 4. Comparison of 5-year conditional excess mortality (in percentage) between original cohort (black) and combined cohort (red) across stage at diagnosis and stratifying by adherence to endocrine therapy: (a,d) Stage I; (b,e) Stage II; (c,f) Stage III.

Table 3 presents the age-specific epidemiological survival indicators derived from the combined cohort across age groups and stage at diagnosis. The adherence group showed higher OS (+ 6% at 5 years and +15.2% at 10 years) and lower 10-year PCa (−18.7%) and 5-year EM (−14.5%) compared to the nonadherent group.

Table 3 also shows that, at Stage I, adherent patients diagnosed before 50 years of age may present a small but non-negligible 1.1% EM when compared to the general population. In contrast, no EM was detected in patients diagnosed beyond that age. Nonadherent patients present 4.6% to 9% higher EM, depending on the age group. I Stage II, adherent patients diagnosed beyond 59 years did not show EM during the follow-up. The largest differences in survival indicators between adherent and nonadherent patients

were observed in the Stage III group, with better prospects of survival in adherent compared to nonadherent patients, independently of age at BC diagnosis.

Table 3. Survival indicators derived from synthetic cohort comparing breast cancer patients adherent vs. nonadherent to endocrine therapy across age groups and stage at diagnosis.

		OS(5) (%)	OS(10) (%)	PCa(10) (%)	POC(10) (%)	EM(5) (%)
Adherent	N *	Me(95% CI)	Me(95% CI)	Me(95% CI)	Me(95% CI)	Me(95% CI)
Stage I						
≤49 years	72,817	98.3 (98.2; 98.4)	95.7 (95.5; 95.9)	1.8 (1.6; 1.9)	2.5 (2.5; 2.5)	1.1 (1.0; 1.3)
50–59 years	92,526	98.4 (98.3; 98.5)	96.0 (95.9; 96.2)	0.2 (0.1; 0.3)	3.8 (3.6; 3.9)	0.0 (−0.1; 0.1)
60–74 years	167,001	97.3 (97.2; 97.3)	92.9 (92.7; 93.1)	0.2 (0.1; 0.3)	6.9 (6.7; 7.1)	0.0 (−0.1; 0.1)
Stage II						
≤49 years	98,722	95.9 (95.7; 96.1)	89.0 (88.8; 89.3)	8.7 (8.4; 8.9)	2.3 (2.0; 2.6)	5.8 (5.5; 6.2)
50–59 years	92,612	96.5 (96.4; 96.6)	90.6 (90.4; 90.9)	3.4 (3.1; 3.6)	6.0 (5.9; 6.1)	2.3 (2.1; 2.6)
60–74 years	92,919	91.9 (91.7; 92.1)	78.6 (78.3; 79.0)	0.6 (0.4; 0.9)	20.8 (20.3; 21.1)	0.0 (−0.4; 0.4)
Stage III						
≤49 years	40,968	87.8 (87.5; 88.1)	69.6 (69.1; 70.1)	27.9 (27.3; 28.4)	2.5 (2.5; 2.6)	19.7 (19.1; 20.2)
50–59 years	36,659	88.0 (87.7; 88.3)	69.5 (68.9; 70.1)	24.8 (24.2; 25.4)	5.7 (5.7; 5.8)	17.9 (17.3; 18.6)
60–74 years	41,335	78.8 (78.4; 79.2)	47.3 (46.7; 47.9)	28.1 (27.5; 28.8)	24.6 (24.5; 24.7)	25.4 (24.5; 26.3)
Overall	735,559	94.5 (94.4; 94.6)	85.7 (85.6; 85.8)	0.9 (0.8; 1.0)	13.5 (13.3; 13.6)	0.5 (0.4; 0.7)
Nonadherent						
Stage I						
≤49 years	34,356	96.5 (96.3; 96.7)	90.8 (90.4; 91.2)	7.1 (6.7; 7.5)	2.2 (2.2; 2.2)	4.6 (4.3; 5.0)
50–59 years	29,888	92.8 (92.5; 93.1)	81.5 (80.9; 82.0)	13.1 (12.6; 13.7)	5.4 (5.4; 5.4)	9.0 (8.5; 9.6)
60–74 years	33,313	87.7 (87.3; 88.0)	68.8 (68.2; 69.5)	7.4 (6.8; 8.0)	23.9 (23.2; 24.1)	5.4 (4.6; 6.1)
Stage II						
≤49 years	49,897	94.7 (94.5; 94.9)	85.9 (85.5; 86.3)	11.9 (11.5; 12.3)	2.1 (2.1; 2.1)	8.0 (7.6; 8.4)
50–59 years	28,269	87.2 (86.8; 87.6)	66.6 (65.9; 67.4)	28.0 (27.3; 28.7)	5.3 (5.3; 5.4)	20.8 (20.1; 21.6)
60–74 years	26,084	85.6 (85.2; 86.0)	62.7 (61.9; 63.4)	13.9 (13.0; 14.7)	23.5 (23.4; 23.6)	11.9 (10.9; 12.9)
Stage III						
≤49 years	16,468	77.6 (77.0; 78.2)	46.3 (45.3; 47.3)	49.8 (48.9; 50.7)	3.9 (3.8; 4.0)	39.5 (38.3; 40.7)
50–59 years	14,953	78.0 (77.3; 78.6)	46.5 (45.5; 47.5)	47.0 (46.1; 47.9)	6.5 (6.4; 6.6)	38.0 (36.8; 39.3)
60–74 years	13,020	72.2 (71.4; 72.9)	31.3 (30.2; 32.3)	41.1 (40.1; 42.0)	27.7 (27.4; 28)	43.9 (42.1; 45.7)
Overall	246,248	88.5 (88.4; 88.6)	70.5 (70.3; 70.8)	19.6 (19.4; 19.9)	9.8 (9.6; 9.9)	14.5 (14.3; 14.8)
Overall difference						
Adherent vs. nonadherent **		6.0	15.2	−18.7	3.7	−14.0

Note: survival indicators expressed in percentage; **N** *: number of patients in the combined synthetic cohort; **Me**: Median; **95 CI**: 95% Credible Interval; **OS(T)**: observed survival at T = 5 and T = 10 years after cancer diagnosis; **PCa(10)**: crude probability of death due to cancer at T = 10 years; **POC(10)**: crude probability of death due to other causes at T = 10 years; **EM(T)**: 5-year conditional excess mortality at T years after cancer diagnosis; **: difference in the median estimate Adherent minus nonadherent.

Figure 5 shows the comparison of the 3 main population-based survival indicators across age groups and stratified by adherent and nonadherent patients: EM(5), PCa(10) and OS(10). In Stages I and II of BC, differences in EM(5) and PCa(10) between adherent and nonadherent patients were clearly marked and showed their maximum among BC patients diagnosed beyond 50 years. At Stage III, the age trend of these two indicators was similar, showing a marked rise beyond 59 years of age at BC diagnosis. Lastly, OS(10) showed two patterns: (i) for adherent patients, survival was similar up to 59 years of age and decrease thereafter, independently of stage at diagnosis; (ii) for nonadherent patients, OS(10) exponentially decreased with age except in Stage III.

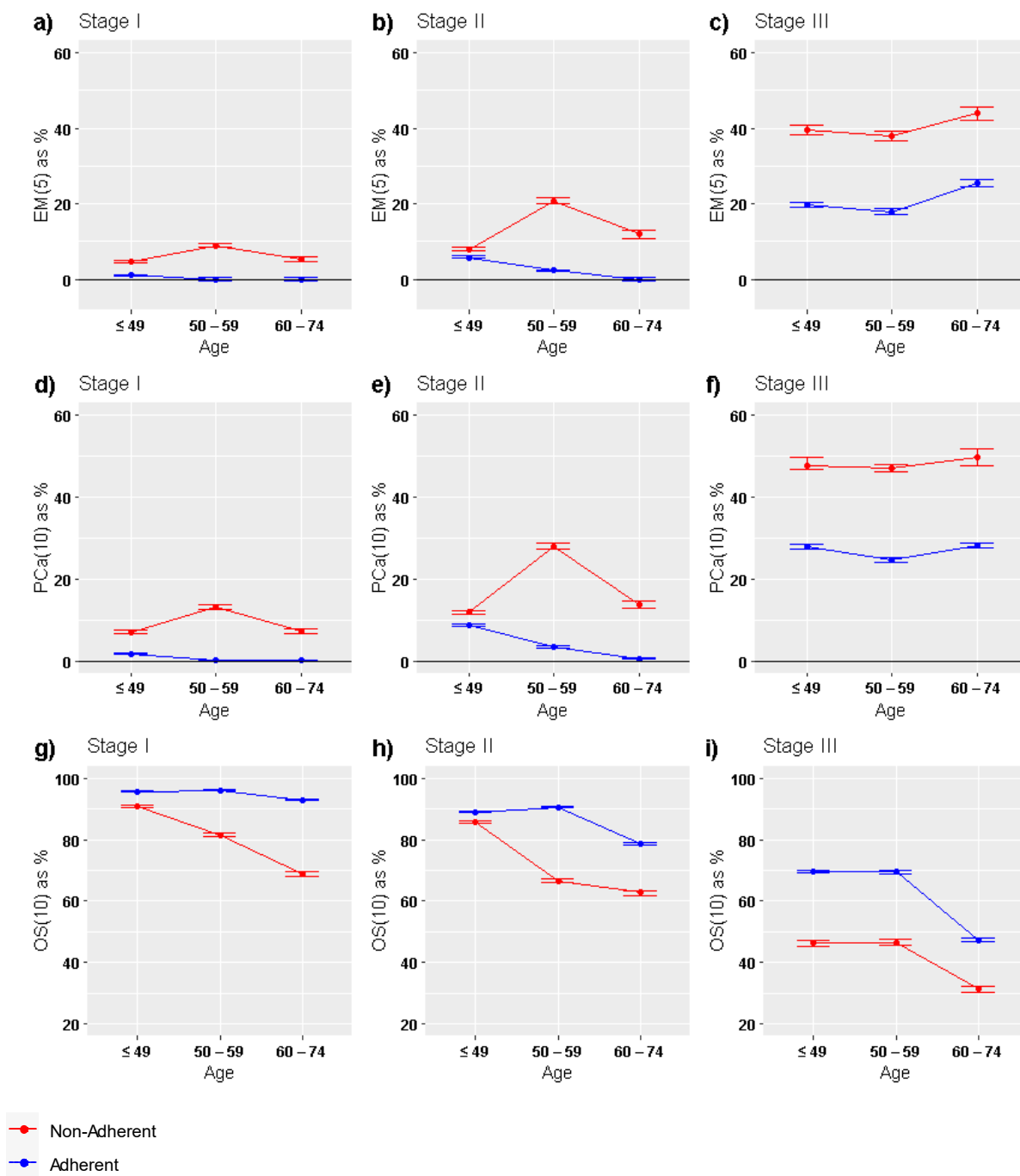


Figure 5. Graphical comparison of main population-based survival indicators between adherent and nonadherent BC patients across age groups: (a–c) EM(5); (d–f) PCa(10); (g–i) OS(10). EM(5): 5-year conditional excess mortality (EM) at $T = 5$ years after cancer diagnosis; PCa(10): crude probability of death due to cancer at $T = 10$ years; OS(10): observed survival at 10 years after cancer diagnosis.

4. Discussion

This study provides estimates of the most common population-based statistical indicators in order to assess the impact of stage, age, and adherence to ET for survival in patients with positive estrogen- and/or progesterone-receptor BC. We compared the estimates from the original cohort with those derived from synthetic datasets generated through graphical models fitted to the cancer registry cohort. Using the advantages of probabilistic graphical modeling, we first identified the probabilistic data structure, used it to impute

the adherence status in patients with missing data for this variable, and simulated data for a large cohort to estimate age-specific survival indicators. We implemented the Comb-SynSurData method in order to select the best subsets of four synthetic datasets derived from ModGraProDep. To the best of our knowledge, this is the first study to show that adherence to ET greatly impacts BC survival among HR+ patients with early-stage breast cancer: no excess risk of death up to 10 years after BC in women diagnosed beyond 49 years of age. This result sheds light into curing BC for this group of patients.

The assessment of treatment response is crucial for evaluating anticancer therapies, treatment planning, and outcomes, where patients' OS is the baseline measure [33]. However, that evaluation requires a large sample and long-term follow-up, which are usually not available in the same study. We used a method for generating a large sample of synthetic data on the basis of the original cohort in order to estimate the observed survival indicators using the original cohort data, which had the minimal required long-term follow-up of 10 years for assessing EM due to BC [8]. Using these indicators, healthcare policy planning should be informed by the estimated prevalence of cancer deaths at a population level, which can be calculated through RS [34]. These indicators are strongly related to the concept of a statistical assessment of the "cure" of BC [35], which entails: (I) long survival time beyond 10 years and equal life expectancy [9], and (II) no cancer relapses up to almost 10 years after BC diagnosis [35].

Our study has a strong limitation in assessing the statistical cure of BC: our follow-up cannot go beyond 10 years. Another limitation is that the simulated cohorts were based on the observed data provided by the original cohort. Therefore, survival indicators derived from these simulated cohorts can only internally validate the indicators estimated from the original data. The availability of external data provided by other cancer registries with similar information would be useful for an additional validation of the results and reproducibility. Information on long-term prognosis by stage, receptor status and adherence to ET is information not usually reported by population-based cancer registries [9]. However, recent studies suggest the need for using these variables for population-based studies in order to assess whether the influence of stage or BC subtype on survival lessens in the long term, which might lead to a consideration of cancer cure in early stages [36–38]. The impact of ET adherence on BC patient survival is significant [39], and our results, which show differences in EM when comparing the cohort's mortality with that of the general population, are relevant to this. Moreover, differences between adherent and nonadherent patients are significant across all age groups, but show different impact depending on stage at diagnosis. This point must be accounted and further investigated, since age, stage, and treatment play a crucial role in the clinical follow-up of BC patient. Studies regarding this are needed.

A small but significant level of EM was detected in the adherent group of younger BC patients (<50 years) diagnosed at Stage I. However, survival estimates for these women using the combined cohort could be slightly lower than the observed survival in the original cohort, and this could limit the use of this subset of data. On the other hand, a previous study carried out on a cohort with ductal carcinoma in situ and diagnosed in Girona also detected statistically significant EM in patients diagnosed before 50 years of age [40]. Evidence suggests that differences in biological characteristics of breast tumors could impact patient survival [41]. Moreover, 5- and 10-year local recurrences at early stages [42] arise depending on age and molecular subtype. Although a high proportion of BCs are HR+ and HER2–, those diagnosed in young women are likely to be more aggressive [43,44], even in luminal-like early BC [45,46]. A study carried out using SEER data noted worse BC-specific survival for women in the oldest age groups for every BC subtype analyzed, with the exception of Stage IV triple-negative disease [10]. In that study and others, worse survival was observed in patients diagnosed before 35 years of age at Stages I–III [10,46]. Other studies showed that young age is also a predictor of decreased adherence to adjuvant ET, which in turn is associated with increased mortality [47]. Although ET is unquestionably a therapeutic tool for HR + BC, these strategies are associated with potential side effects

and toxicity, which may have a differential effect depending on age [48]. On the other hand, randomized trials showed that, in premenopausal women with BC, the addition of ovarian suppression to tamoxifen may increase 8-year rates of both disease-free and overall survival [49]. However, diagnoses in the cohort under study predate results of these randomized trials, so women under 49 years of age in our study could not have had access to these improved treatments. Studies on BC survival and late adverse events due to ET must be considered beyond 10 years of follow-up, since evidence suggests that distant recurrences may arise from 5 to 20 years after diagnosis [49].

Studies of EM derived from small cohorts of cancer patients must be further evaluated using larger cohorts [40]. Here, we present a procedure for simulating a large sample dataset by fitting graphical models to cohort data and coupling a log-linear model and a Bayesian network. Since our interest was in simulating the most reliable data, one aim was to assess the probabilistic dependencies between variables. ModGraProDep identifies a set of graphical models by using a heuristic search based on changing k , a penalty factor in the partial likelihood (see Equation (2) above) [16–18,21]. Although specific values of k such as $k = 2$ and $\log(N)$ equation lead to two known measures for model choice, AIC and BIC, ModGraProDep identifies two alternative models, one using $k = 1$ and another testing the arch's statistical significance at $\alpha = 0.05$ [13,18]. Vilardell et al. showed that estimating survival from one of these four models could provide reliable survival indicators [13]. Here, we introduced a method for deriving a synthetic dataset that provides better survival indicators by combining the best subsets of data of several synthetic datasets. An interesting feature in ComSynSurData is that it could be adapted to use any set of simulated data, and these could come from oversampling techniques, such as SMOTE [17], Borderline SMOTE [18] and MWMOTE [19]. However, synthetic datasets derived from ModGraProDep provide additional information about the data structure and data relationship between variables. The latter can also be useful for clinicians and epidemiologists in understanding the probabilistic patterns of the disease under study.

5. Conclusions

To sum up, coupling relative survival modeling with synthetic data simulation validated our main clinical result: patients with HR+ breast cancers diagnosed beyond 49 years of age at Stage I and diagnosed beyond 59 years of age in Stage II do not have 10-year EM compared to the general population if they follow the prescribed regimen of ET. These results call for developing interventions that promote adjuvant ET adherence in eligible BC patients given its potential benefits in improving cancer survival. The methodology presented here demonstrates the potential use of probabilistic graphical modeling in generating reliable synthetic datasets to be used for validating population-based survival indicators when sample size is an issue.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/ijerph19063605/s1>, supplementary file: R code implementation of ComSynSurData for replicating analysis.

Author Contributions: Conceptualization, R.C., R.F., R.M.-G., A.A., M.C., J.M.M., R.M.-G., J.G., Á.I. and J.M.B.; data curation, R.F., A.A., M.C., M.P. and R.C.; formal analysis, R.F., M.B., A.A., J.M.M., and R.C.; funding acquisition, R.C. and J.M.B.; investigation, R.C., R.F., M.B., M.C., R.M.-G., J.A.E., G.V., Á.I., J.G. and J.M.B.; methodology, M.B., A.A., J.M.M., M.V. and R.C.; project administration, R.C., J.G., Á.I.; resources, R.C., R.M.-G., Á.I., R.M.-G. and J.M.B.; software, A.A., J.M.M., M.V. and R.C.; supervision, R.C., R.F., M.B., M.C., R.M.-G., Á.I., J.G. and J.M.B.; validation, M.B., M.V., A.A., M.P. and R.C.; visualization, R.C.; writing—original draft, R.F., M.B. and R.C.; writing—review and editing, R.F., M.B. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Instituto de Salud Carlos III PI18/01836 funded by FEDER funds/European Regional Development Fund (ERDF)-a way to Build Europe-//FONDOS FEDER “una manera de hacer Europa”.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (The Clinical Research Ethics Committee of Bellvitge University Hospital (“Comité de Ética del Hospital Universitario de Bellvitge”) Approval Code: PR160/18 Approval Date: 10-11-2018).

Informed Consent Statement: The public health administration of each autonomous community/province in Spain authorized the collection and use of this data for its analysis without requirement of informed consent, covered by the Spanish general and public health laws 14/1986 and 33/2011.

Data Availability Statement: Data supporting reported results can be requested to the corresponding author.

Acknowledgments: We also acknowledge Agència d’Avaluació d’Universitats i Recerca (2017SGR00735) from Generalitat de Catalunya and PGC2018-095931-B-100 (MCIU/AEI/FEDER, UE). We also thank CERCA Programme/Generalitat de Catalunya for institutional support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Chirlaque, M.D.; Salmerón, D.; Galceran, J.; Ameijide, A.; Mateos, A.; Torrella, A.; Jiménez, R.; Larrañaga, N.; Marcos-Gragera, R.; Ardanaz, E.; et al. Cancer survival in adult patients in Spain. Results from nine population-based cancer registries. *Clin. Transl. Oncol.* **2018**, *20*, 201–211. [CrossRef] [PubMed]
3. Clèries, R.; Ameijide, A.; Buxó, M.; Martínez, J.M.; Marcos-Gragera, R.; Vilardell, M.-L.; Carulla, M.; Yasui, Y.; Vilardell, M.; Espinàs, J.A.; et al. Long-term crude probabilities of death among breast cancer patients by age and stage: A population-based survival study in Northeastern Spain (Girona–Tarragona 1985–2004). *Clin. Transl. Oncol.* **2018**, *20*, 1252–1260. [CrossRef] [PubMed]
4. Hieke, S.; Kleber, M.; König, C.; Engelhardt, M.; Schumacher, M. Conditional Survival: A Useful Concept to Provide Information on How Prognosis Evolves over Time. *Clin. Cancer Res.* **2015**, *21*, 1530–1536. [CrossRef]
5. Shack, L.; Bryant, H.; Lockwood, G.; Ellison, L.F. Conditional relative survival: A different perspective to measuring cancer outcomes. *Cancer Epidemiol.* **2013**, *37*, 446–448. [CrossRef]
6. Cronin, K.A.; Feuer, E.J. Cumulative cause-specific mortality for cancer patients in the presence of other causes: A crude analogue of relative survival. *Stat. Med.* **2000**, *19*, 1729–1740. [CrossRef]
7. He, V.Y.F.; Condon, J.R.; Baade, P.D.; Zhang, X.; Zhao, Y. Different survival analysis methods for measuring long-term outcomes of Indigenous and non-Indigenous Australian cancer patients in the presence and absence of competing risks. *Popul. Health Metr.* **2017**, *15*, 1. [CrossRef]
8. Maso, L.D.; Guzzinati, S.; Buzzoni, C.; Capocaccia, R.; Serraino, D.; Caldarella, A.; Tos, A.P.D.; Falcini, F.; Autelitano, M.; Masanotti, G.; et al. Long-term survival, prevalence, and cure of cancer: A population-based estimation for 818,902 Italian patients and 26 cancer types. *Ann. Oncol.* **2014**, *25*, 2251–2260. [CrossRef]
9. Maso, L.D.; Panato, C.; Tavilla, A.; Guzzinati, S.; Serraino, D.; Mallone, S.; Botta, L.; Boussari, O.; Capocaccia, R.; Colonna, M.; et al. Cancer cure for 32 cancer types: Results from the EUROCARE-5 study. *Int. J. Epidemiol.* **2020**, *49*, 1517–1525. [CrossRef]
10. Freedman, R.A.; Keating, N.L.; Lin, N.U.; Winer, E.P.; Vaz-Luis, I.; Lii, J.; Exman, P.; Barry, W.T. Breast cancer-specific survival by age: Worse outcomes for the oldest patients. *Cancer* **2018**, *124*, 2184–2191. [CrossRef]
11. Munzone, E.; Colleoni, M. Optimal management of luminal breast cancer: How much endocrine therapy is long enough? *Ther. Adv. Med. Oncol.* **2018**, *10*, 1758835918777437. [CrossRef] [PubMed]
12. Font, R.; Alfons, J.; Agustí, E.; Angel, B.; Jaume, I.; Saladie, F. Influence of adherence to adjuvant endocrine therapy on disease-free and overall survival: A population-based study in Catalonia, Spain. *Breast Cancer Res. Treat.* **2019**, *175*, 733–740. [CrossRef] [PubMed]
13. Vilardell, M.; Buxó, M.; Clèries, R.; Martínez, J.M.; Garcia, G.; Ameijide, A.; Font, R.; Civit, S.; Marcos-Gragera, R.; Vilardell, M.L.; et al. Missing data imputation and synthetic data simulation through modeling graphical probabilistic dependencies between variables (ModGraProDep): An application to breast cancer survival. *Artif. Intell. Med.* **2020**, *107*, 101875. [CrossRef] [PubMed]
14. Austin, P.C. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Stat. Med.* **2012**, *31*, 3946–3958. [CrossRef] [PubMed]
15. Metcalfe, C.; Thompson, S.G. The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Stat. Med.* **2006**, *25*, 165–179. [CrossRef]
16. Moriña, D.; Navarro, A. The R Package survsim for the Simulation of Simple and Complex Survival Data. *J. Stat. Soft.* **2014**, *59*, 1–20. Available online: <https://www.jstatsoft.org/index.php/jss/article/view/v059i02> (accessed on 18 January 2022). [CrossRef]

17. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. {SMOTE}: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
18. Han, H.; Wang, W.; Mao, B. Borderline-SMOTE: A New Over-Advances in Intelligent Computing. In Proceedings of the International Conference on Intelligent Computing, ICIC 2005, Hefei, China, 23–26 August 2005; Lecture Notes in Computer Science, Sampling Method in Imbalanced Data Sets Learning; Huang, D.S., Zhang, X.P., Huang, G.B., Eds.; Springer: Berlin, Germany, 2005; Volume 3644. [[CrossRef](#)]
19. Barua, S.; Islam, M.M.; Yao, X.; Murase, K. MWMOTE—Majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 405–425. [[CrossRef](#)]
20. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan-Kaufman Publishers: Los Altos, CA, USA, 1988. [[CrossRef](#)]
21. Lauritzen, S.L.; Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Ann. Math. Artif. Intell.* **1988**, *50*, 157–224. [[CrossRef](#)]
22. Højsgaard, S.; Edwards, D.; Lauritzen, S. *Graphical Models with R.*; Springer: New York, NY, USA; Dordrecht, The Netherlands; Heidelberg, Germany; Springer: London, UK, 2012.
23. Ameijide, A.; Clèries, R.; Carulla, M.; Buxó, M.; Marcos-Gragera, R.; Martínez, J.M.; Vilardell, M.L.; Espinàs, J.A.; Borrás, J.M.; Izquierdo, Á.; et al. Cause-specific mortality after a breast cancer diagnosis: A cohort study of 10,195 women in Girona and Tarragona. *Clin. Transl. Oncol.* **2019**, *21*, 1014–1025. [[CrossRef](#)]
24. Singletary, S.E.; Connolly, J.L. Breast cancer staging: Working with the sixth edition of the AJCC Cancer Staging Manual. *CA Cancer J. Clin.* **2006**, *56*, 37–47, Quiz 50, 51. [[CrossRef](#)] [[PubMed](#)]
25. Højsgaard, S. Graphical Independence Networks with the gRain Package for R. *J. Stat. Soft.* **2012**, *46*, 1–26. Available online: <https://www.jstatsoft.org/index.php/jss/article/view/v046i10> (accessed on 18 January 2022).
26. Graf, E.; Schmoor, C.; Sauerbrei, W.; Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **1999**, *18*, 2529–2545. [[CrossRef](#)]
27. Haider, H.; Hoehn, B.; Davis, S.; Greiner, R. Effective ways to build and evaluate individual survival distributions. *J. Mach. Learn. Res.* **2020**, *21*, 1–63.
28. Pohar Perme, M.; Estève, J.; Rachet, B. Analysing population-based cancer survival—Settling the controversies. *BMC Cancer* **2016**, *16*, 933. [[CrossRef](#)] [[PubMed](#)]
29. Seppä, K.; Hakulinen, T.; Läärä, E.; Pitkaniemi, J. Comparing net survival estimators of cancer patients. *Stat. Med.* **2016**, *35*, 1866–1879. [[CrossRef](#)]
30. Clèries, R.; Buxó, M.; Yasui, Y.; Marcos-Gragera, R.; Martínez, J.M.; Ameijide, A.; Galceran, J.; Borrás, J.M.; Izquierdo, Á. Estimating long-term crude probability of death among young breast cancer patients: A Bayesian approach. *Tumori* **2016**, *102*, 555–561. [[CrossRef](#)]
31. Lunn, D.J.; Thomas, A.; Best, N.; Spiegelhalter, D. WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **2000**, *10*, 325–337. [[CrossRef](#)]
32. Sturtz, S.; Ligges, U.; Gelman, A. R2WinBUGS: A Package for Running WinBUGS from R. *J. Stat. Soft.* **2005**, *12*, 1–16. Available online: <https://www.jstatsoft.org/index.php/jss/article/view/v012i03> (accessed on 18 January 2022). [[CrossRef](#)]
33. Wang, S.; Liu, Y.; Feng, Y.; Zhang, J.; Swinnen, J.; Li, Y.; Ni, Y. A review on curability of cancers: More efforts for novel therapeutic options are needed. *Cancers* **2019**, *11*, 1782. [[CrossRef](#)]
34. Mariotto, A.B.; Noone, A.-M.; Howlader, N.; Cho, H.; Keel, G.E.; Garshell, J.; Woloshin, S.; Schwartz, L.M. Cancer survival: An overview of measures, uses, and interpretation. *J. Natl. Cancer Inst. Monogr.* **2014**, *2014*, 145–186. [[CrossRef](#)] [[PubMed](#)]
35. Miller, K.; Abraham, J.H.; Rhodes, L.; Roberts, R. Use of the word “cure” in oncology. *J. Oncol. Pract.* **2013**, *9*, e136–e140. [[CrossRef](#)] [[PubMed](#)]
36. Mariotto, A.B.; Zou, Z.; Zhang, F.; Howlader, N.; Kurian, A.W.; Etzioni, R. Can we use survival data from cancer registries to learn about disease recurrence? The case of breast cancer. *Cancer Epidemiol. Biomark. Prev.* **2018**, *27*, 1332–1341. [[CrossRef](#)] [[PubMed](#)]
37. Van Maaren, M.C.; Strobbe, L.J.A.; Smidt, M.L.; Moosdorff, M.; Poortmans, P.M.P.; Siesling, S. Ten-year conditional recurrence risks and overall and relative survival for breast cancer patients in the Netherlands: Taking account of event-free years. *Eur. J. Cancer* **2018**, *102*, 82–94. [[CrossRef](#)]
38. Van Maaren, M.C.; De Munck, L.; Strobbe, L.J.; Sonke, G.; Westenend, P.; Smidt, M.L.; Poortmans, P.M.; Siesling, S. Ten-year recurrence rates for breast cancer subtypes in the Netherlands: A large population-based study. *Int. J. Cancer* **2018**, *144*, 263–272. [[CrossRef](#)]
39. Männle, H.; Siebers, J.W.; Momm, F.; Münstedt, K. Impact of patients’ refusal to undergo adjuvant treatment measures on survival. *Breast Cancer Res. Treat.* **2020**, *185*, 239–246. [[CrossRef](#)]
40. Roca-Barceló, A.; Viñas, G.; Pla, H.; Carbó, A.; Comas, R.; Izquierdo, Á.; Pinheiro, P.S.; Vilardell, L.; Solans, M.; Marcos-Gragera, R. Mortality of women with ductal carcinoma in situ of the breast: A population-based study from the Girona province, Spain (1994–2013). *Clin. Transl. Oncol.* **2018**, *21*, 891–899. [[CrossRef](#)]
41. Azim, H.H.A.; Michiels, S.; Bedard, P.; Singhal, S.K.; Criscitiello, C.; Ignatiadis, M.; Haibe-Kains, B.; Piccart-Gebhart, M.; Sotiriou, C.; Loi, S. Elucidating prognosis and biology of breast cancer arising in young women using gene expression profiling. *Clin. Cancer Res.* **2012**, *18*, 1341–1351. [[CrossRef](#)]

42. He, X.M.; Zou, D.H. The association of young age with local recurrence in women with early-stage breast cancer after breast-conserving therapy: A meta-analysis. *Sci. Rep.* **2017**, *7*, 11058. [[CrossRef](#)]
43. Johansson, A.L.V.; Trewin, C.B.; Hjerkind, K.V.; Ellingjord-Dale, M.; Johannesen, T.B.; Ursin, G. Breast cancer-specific survival by clinical subtype after 7 years follow-up of young and elderly women in a nationwide cohort. *Int. J. Cancer* **2019**, *144*, 1251–1261. [[CrossRef](#)]
44. Johansson, A.L.V.; Trewin, C.B.; Fredriksson, I.; Reinertsen, K.V.; Russnes, H. In modern times, how important are breast cancer stage, grade and receptor subtype for survival: A population-based cohort study. *Breast Cancer Res.* **2021**, *23*, 17. [[CrossRef](#)] [[PubMed](#)]
45. Liu, Z.; Sahli, Z.; Wang, Y.; Wolff, A.C.; Cope, L.M.; Umbricht, C.B. Young age at diagnosis is associated with worse prognosis in the Luminal A breast cancer subtype: A retrospective institutional cohort study. *Breast Cancer Res. Treat.* **2018**, *172*, 689–702. [[CrossRef](#)]
46. Partridge, A.H.; Hughes, M.E.; Warner, E.T.; Ottesen, R.A.; Wong, Y.-N.; Edge, S.B.; Theriault, R.L.; Blayney, D.W.; Niland, J.C.; Winer, E.P.; et al. Subtype-dependent relationship between young age at diagnosis and breast cancer survival. *J. Clin. Oncol.* **2016**, *34*, 3308–3314. [[CrossRef](#)] [[PubMed](#)]
47. Huiart, L.; Ferdynus, C.; Giorgi, R. A meta-regression analysis of the available data on adherence to adjuvant hormonal therapy in breast cancer: Summarizing the data for clinicians. *Breast Cancer Res. Treat.* **2013**, *138*, 325–328. [[CrossRef](#)] [[PubMed](#)]
48. Condorelli, R.; Vaz-Luis, I. Managing side effects in adjuvant endocrine therapy for breast cancer. *Expert Rev. Anticancer Ther.* **2018**, *18*, 1101–1112. [[CrossRef](#)]
49. Francis, P.A.; Pagani, O.; Fleming, G.F.; Walley, B.A.; Colleoni, M.; Láng, I.; Gómez, H.L.; Tondini, C.A.; Ciruelos, E.; Burstein, H.J.; et al. Tailoring Adjuvant Endocrine Therapy for Premenopausal Breast Cancer. *N. Engl. J. Med.* **2018**, *379*, 122–137. [[CrossRef](#)]