Clinical & Experimental Metastasis (2021) 38:77–87
https://doi.org/10.1007/s10585-020-10066-8

RESEARCH PAPER

Modeling breast cancer survival and metastasis rates from moderate-sized clinical data

Esha Maiti[1]

Received: 23 July 2020 / Accepted: 20 November 2020 / Published online: 3 January 2021

Abstract
Predicting time-dependent survival probability of a breast cancer patient using information such as primary tumor size, grade, node spread status, and patient age at the time of surgery can be of immense help in managing life expectations and strategizing postoperative treatment. However, for moderate-sized clinical datasets the application of standard Kaplan–Meier theory to determine survival probability as a function of multiple cofactors can become challenging when continuous variables like tumor diameter and survival time are segmented into a large number of narrow intervals, a problem commonly termed the curse of dimensionality. We circumvent this problem by modeling the patient-to-patient distribution of primary tumor diameter with a realistic, right-skewed function, and then matching the diameter-marginalized survival with the mean Kaplan–Meier survival for the data. We apply this procedure on a recent clinical data from 1875 breast cancer patients and develop parameters that can be readily used to estimate post-surgery survival for an arbitrary time length. Finally, we show that the observed fraction of node-positive patients can be quantitatively explained within a simple tumor growth and metastasis framework. Employing two different tumor growth models from the literature (i.e., Gompertz and logistic growth models), we utilize the observed fraction-node-positive data to determine metastasis rates from the surface of a primary tumor and its patient-to-patient distribution.

Keywords Breast cancer · Survival · Kaplan–Meier · Metastasis rate · Tumor growth

Introduction

Breast cancer is a leading cause of death for women worldwide [1]. When a breast cancer patient is faced with the decision of whether to undergo surgery, it is of immense help to have knowledge of the expected survival time (beyond surgery) as a function of the patient's age and features of the primary tumor, such as size, grade, and whether it has spread to one or multiple nodes. Such knowledge can significantly aid in managing the patient's quality of life along with choosing among postoperative treatment options. Kaplan–Meier (KM) theory [2] has long been the standard nonparametric method of computing survival probability curves from right-censored (or uncensored) data and has been employed by many past studies on estimating effects of cofactors such as tumor features (size, grade, node spread status) and patient age at the time of surgery on post-surgery survival chances [3–5]. However, the accuracy and practical application of the method requires a large dataset of patients comprising maybe tens of thousands to hundreds of thousands of subjects studied over a long period of time [6, 7]. A moderately large dataset (hundreds to a few thousand subjects) segmented into narrow ranges of patient age, tumor diameter, specific grade, and node status quickly becomes too sparse for reliable application of KM.

In this work we address the above challenge by categorizing only the age variable while keeping tumor diameter and time of survival as continuous variables. The post-surgery survival probability of a patient is expressed as a parameterized function of the two continuous variables, i.e., (1) tumor diameter at the time of surgery, and (2) the amount of time (number of years) beyond surgery, with the parameter values being dependent on the categorical cofactors (patient

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10585-020-10066-8) contains supplementary material, which is available to authorized users.

✉ Esha Maiti
eshmait@gmail.com

[1] Data Science, Chief Data Office, Kaiser Permanente, Pleasanton, CA 94588, USA

age group, tumor grade, and node status at the time of surgery). Optimized parameters are determined by representing the patient-to-patient distribution of diameters with a smooth right-skewed function [8], and then matching the resulting diameter-marginalized survival function with the Kaplan–Meier mean survival curve computed from clinical data. The procedure yields the survival probability of a patient as a function of time after surgery, and cofactors such as age group, tumor diameter, and node status.

Finally, we show that the observed fraction of patients with positive lymph nodes (i.e., with metastatic spread to at least one of the nodes) can be quantitatively explained by a simple tumor growth and metastasis model. Using two different tumor growth models from the literature and assuming metastasis rates proportional to tumor surface area, we utilize the observed fraction of node-positive patients in our clinical dataset to determine the metastasis rate constant and the distribution of its patient-to-patient variation.

Data

In this work, we analyze the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) breast cancer dataset on around 2000 patients [9]. This dataset has been the subject of several recent studies, ranging from genomic architecture [9] and signaling pathways [10] to cancer prognosis using machine learning [11] and evolutionary modeling of breast cancer [12]. For our own purposes, we focus on six variables: patient age, tumor diameter, survival status (censored or uncensored), survival time, neoplasm histologic grade (henceforth referred to simply as tumor "grade"), and the Nottingham Prognostic Index (NPI). Tumor grade in this dataset is a three-level categorical variable defined as either 1 (well-differentiated, more benign), 2 (moderately differentiated), or 3 (poorly differentiated). Node status ("node"), was also included as a three-level categorical variable defined as node = 1 (no nodes affected by metastatic spread), node = 2 (1–3 nodes affected), or node = 3 (more than 3 nodes affected). The Nottingham Index, NPI, is the sum of node, grade, and a small contribution proportional to the tumor diameter. Thus, it is straightforward to determine the node status N for each patient.

After removing patients with incomplete data, we ended up with 1875 patients in our dataset, which had the following frequency breakdown: node = 1 (968), node = 2 (603), node = 3 (304); grade = 1 (167), grade = 2 (763), and grade = 3 (945). Figure 1 summarizes the distribution of tumor diameter and (uncensored) survival times for the different node and grade categories. From the boxplots of Fig. 1a and b, we can see a positive association between diameter size and node status and a negative association between survival times and node status, i.e., with higher levels of metastatic spread to the nodes the diameter distribution changes to larger sizes and survival times shift to smaller values, respectively. Also, the probability of survival as a function of tumor grade is qualitatively similar to the probability of survival as a function of node status, although there are some quantitative differences, as discussed in the sections below.

Kaplan–Meier survival

From the censored survival data, it is straightforward to compute the survival probability (along with uncertainty estimation) using KM theory. The mean KM survival probability for the whole dataset of 1875 patients, as well as for subsets segmented according to node status, age group, and tumor grade, is summarized in Fig. 2. The uncertainty in the estimate of overall survival is represented by the 95% confidence bounds in Fig. 2a. Such uncertainty is known to increase proportionally to the time of survival and decrease inversely proportionally to the square root of the number of patients at risk [2]. Thus, for categories with a low number of samples, the uncertainty in the KM estimate can be significantly higher than the tight bounds indicated in Fig. 2a.

Prior to discussing survival results, we would like to note that age was categorized into three levels (or age groups), i.e., < 55 (age group 1), 55–65 (age group 2), and > 65 (age group 3), with 611, 496, and 768 patients, respectively. Such choices of groups were based on an exploration of different segmentations and performing logrank tests [13] to decide if statistically significant differences in survival existed among different age groups. For instance, if we segment the data into four age groups, < 50, 50–60, 60–70, and > 70, we find no significant survival difference between the first two age groups, but significant differences among all other pairs. The chosen three-level age categorization in this work was not based on any rigorous statistical procedure, but was rather the result of seeking a segmentation that satisfies: (1) a small number of categories; (2) round values of age boundaries; (3) significant survival difference between each age group pair (by logrank test); and (4) significant patient population in each age group.

From the KM curves of Fig. 2b–d several interesting trends were noted. The overall decreasing survival probability with increasing node status, age group, and tumor grade is intuitive, and thus, not surprising. However, some results were unexpected. For instance, Fig. 2c shows that for times less than 8 years, the survival probability is higher for age group 55–65 than for age group < 55. Logrank tests [13] indicate significant survival differences among the three age groups (< 55, 55–65, and > 65), although the difference between the first two age groups is relatively smaller than their difference from age group 3. Such results are consistent with the survival curves

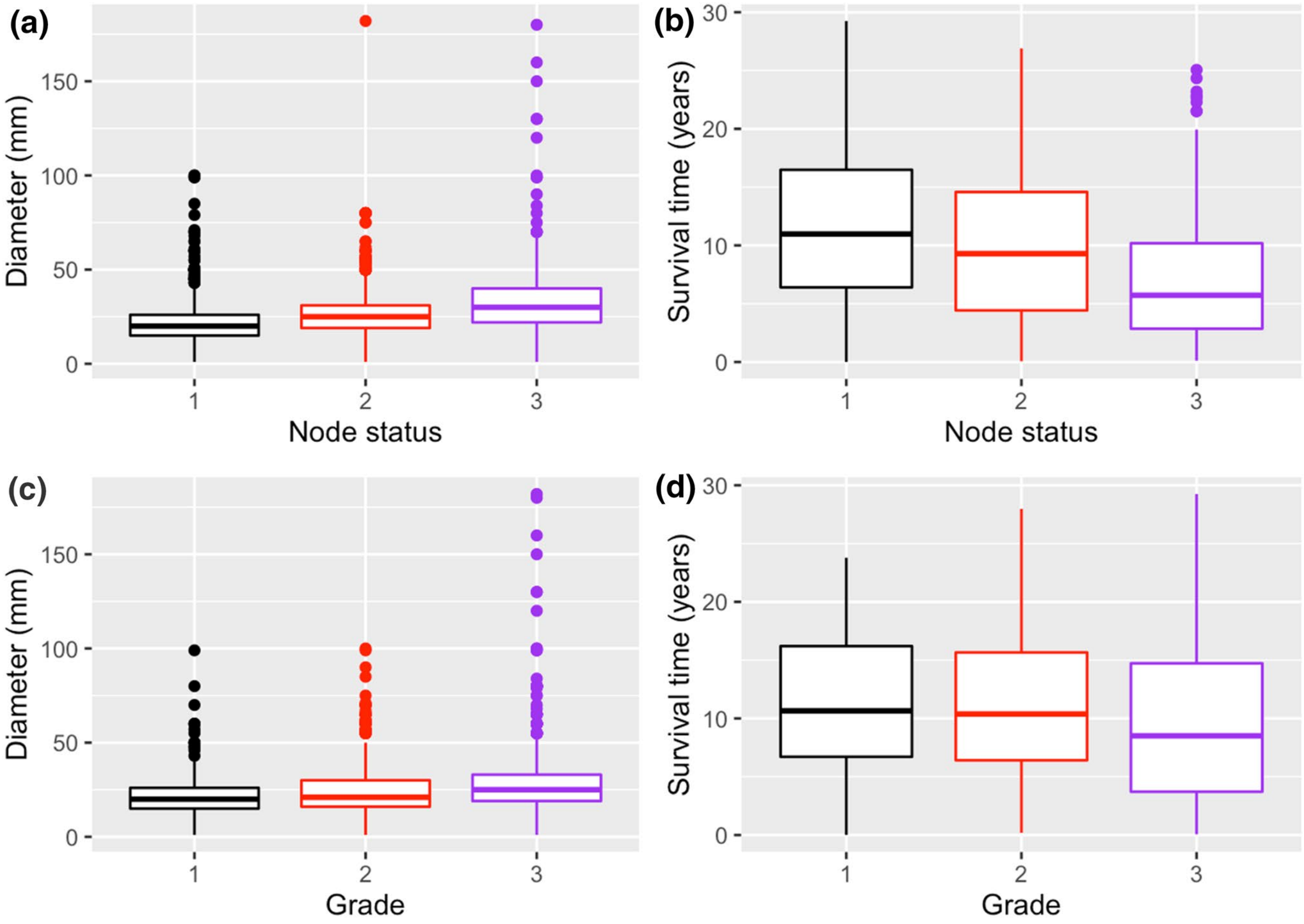

Fig. 1 Exploratory data analysis of the METABRIC dataset. Boxplot representation of the distribution of tumor diameter and survival times (uncensored) for the three levels of node status and tumor grades

of Fig. 2c. Also, as Fig. 2d indicates, the survival differences among the three tumor grades are smaller than those between the node categories. Thus, in the following analysis, dependence on node status has been explored more extensively.

As shown in Fig. 2b–d, the Cox proportional hazard model [14, 15] does not hold as a function of cofactors such as node status, age group, or tumor grade. Figure 3, which plots the hazard function for the three different node status and the three different age groups, supports this finding. Statistical tests on hazard proportionality [16] on node, grade, and age group cofactors resulted in significantly small p-values in all cases, thereby confirming the non-applicability of the Cox model for this dataset.

Distribution of tumor diameters and the survival function

The main strategy in this work, i.e., fitting the distribution of diameters with a smooth probability distribution, arose from noting that in the METABRIC dataset the diameter distribution for any category segment appears to follow the same qualitatively similar right-skewed distribution. This is exemplified in Fig. 4 for four different categories, i.e., (a) the entire dataset of 1875 patients; (b) patients with tumor grade 3; (c) patients of age group 1 (< 55) with node status 1; and (d) patients with node status 3.

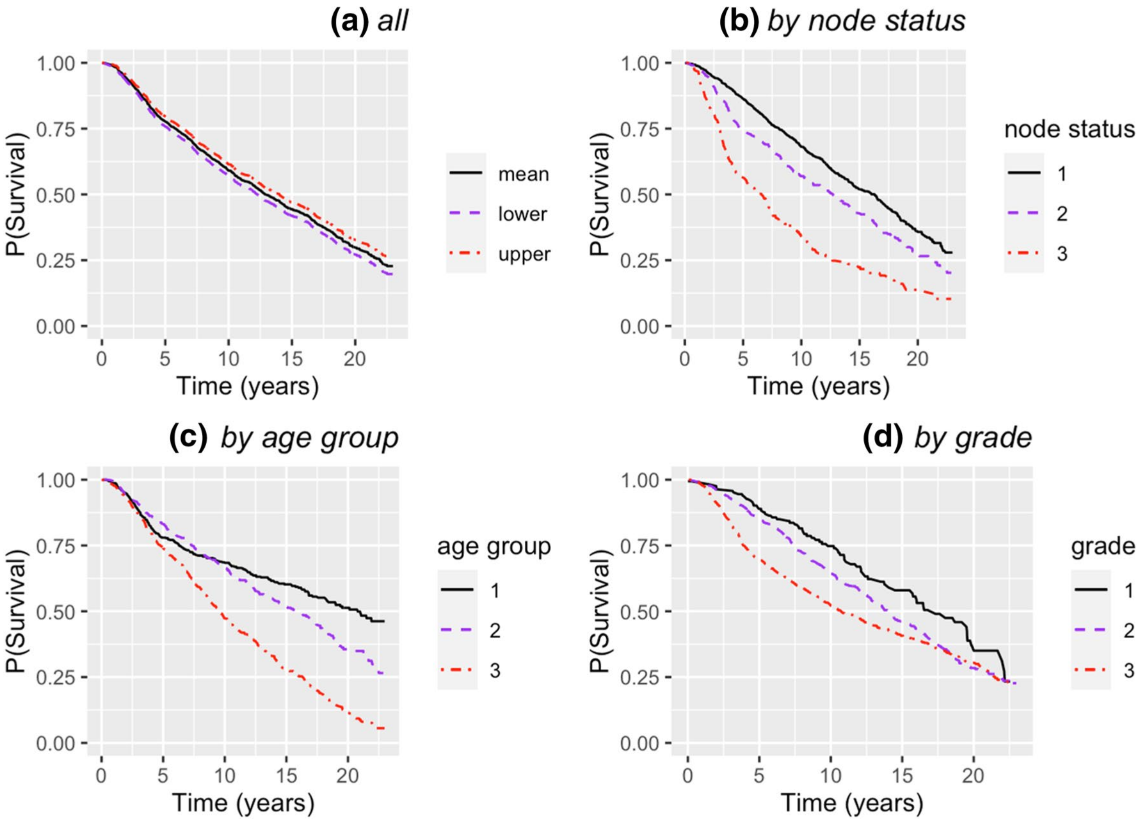

Fig. 2 Mean Kaplan–Meier survival probability for the METABRIC dataset: **a** whole dataset, **b** for each level of node status, **c** for each of the three age groups, **d** for each level of tumor grade. In **a** the 95% confidence margins are also included

To fit the right-skewed histograms of Fig. 4 we chose the following three-parameter probability distribution function, originally used by Dagum [8]:

$$\varphi_{pdf}(D) = \frac{(ap/b)(D/b)^{ap-1}}{\left\{(D/b)^a + 1\right\}^{p+1}} \tag{1}$$

Thus, for each category subset, we have a set of Dagum parameters a, b, p that best fits the corresponding histogram of diameter distribution, which results in the best fit $\varphi_{pdf}(D)$ for that category.

As mentioned in the introduction, the reason for pursuing a smooth diameter distribution in this work is twofold: (1) lack of enough data within an over-specified category with a narrow range of diameters; and (2) resulting increase in the number of model parameters. The idea behind our approach is that each KM survival curve of Fig. 2 is a marginal distribution of a survival function (S) of two continuous variables, survival time (t) and tumor diameter (D), that has been marginalized (or integrated) over variable D. Next, we selected a functional form for S as a function of two continuous variables, t and D. To this end, we segmented the dataset according to many different (node, age group, grade, diameter range) classes and explored the patterns of KM survival probability. From such analyses we found that survival probability as a function of t and D can be modeled by the function:

$$S(D, t; z) = e^{-k(z)D^n t^{\alpha(z)}}, \tag{2}$$

where z represents the set of cofactors such as node status, age group, and tumor grade. In the above model, the exponent n is kept independent of the cofactors z, while the parameters k and α are both z-dependent. The hazard

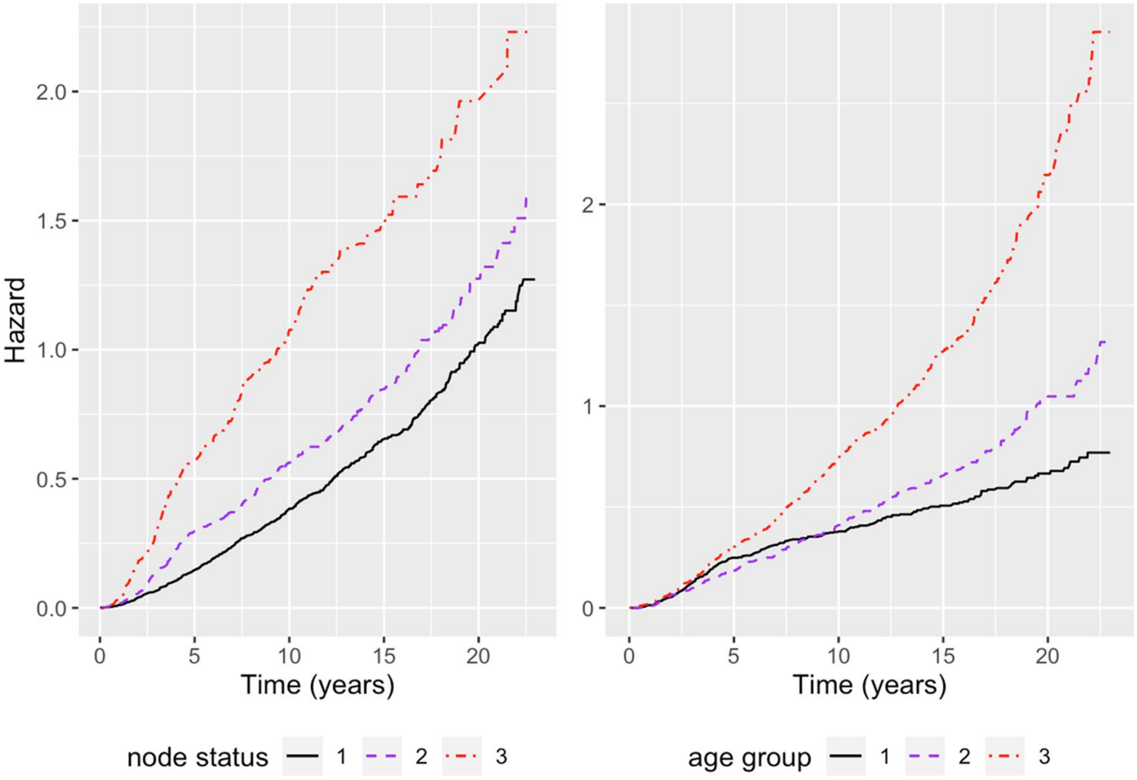

Fig. 3 The hazard function corresponding to the KM survival curves of Fig. 2b and c. The hazard ratio between different curves vary with time, thus showing the inapplicability of the Cox proportional hazard model for cofactors such as node status and age group

function [17, 18] corresponding to the above model is readily obtained as:

$$h(D,t;z) = \alpha(z)k(z)D^n t^{\alpha(z)-1}. \tag{3}$$

The z-dependence of the parameter α (along with marginalization over diameters D as discussed in the following section) leads to the breakdown of the Cox proportional hazard model, consistent with previous discussions.

Optimizing parameters and survival as a function of diameter

The optimized survival parameters $n, k(z)$, and $\alpha(z)$ are obtained by matching the marginalized survival, defined by:

$$S_{\text{model}}(t;z) = \int \varphi_{pdf}(D) e^{-k(z)D^n t^{\alpha(z)}} dD, \tag{4}$$

with the corresponding KM survival curves, e.g., as shown in Fig. 2. From numerical experiments, we see that the root-mean-squared error (RMSE) between the KM survival curves computed from the data and the corresponding marginalized survival S_{model} is low for a range of values of the triplet (n, $k(z)$, $\alpha(z)$). Thus, to reduce the number of parameters, we constrain the parameter n to be of fixed value (independent of cofactors z) and optimize only $k(z)$ and $\alpha(z)$ for each cofactor combination of interest. In the analysis below, we have chosen this value to be 0.8, although a slightly different value of n (e.g., 0.7 or 0.9) would have also yielded comparable results. Table 1 lists the various cofactor sets we have explored in this study and the corresponding optimized survival parameters. Figure 5 displays the results for S_{model} corresponding to the cofactor sets in Fig. 2. These results show good agreement with the KM curves of Fig. 2.

With the optimized parameters shown in Table 1, Eq. (4) can readily be used to estimate survival probability as a function of D. Figure 6 illustrates such a prediction for example cases. When comparing Fig. 6a and b, we see that the 5-year survival probability for age group 2 is higher than that of age group 1, while the trend reverses for the 15-year survival probability, which is consistent with the survival curves of Fig. 2c and 5c. The relative ordering of the curves for other cofactors is also as expected.

As a more direct rationalization of the choice of the optimized parameters in Table 1, we have segmented the

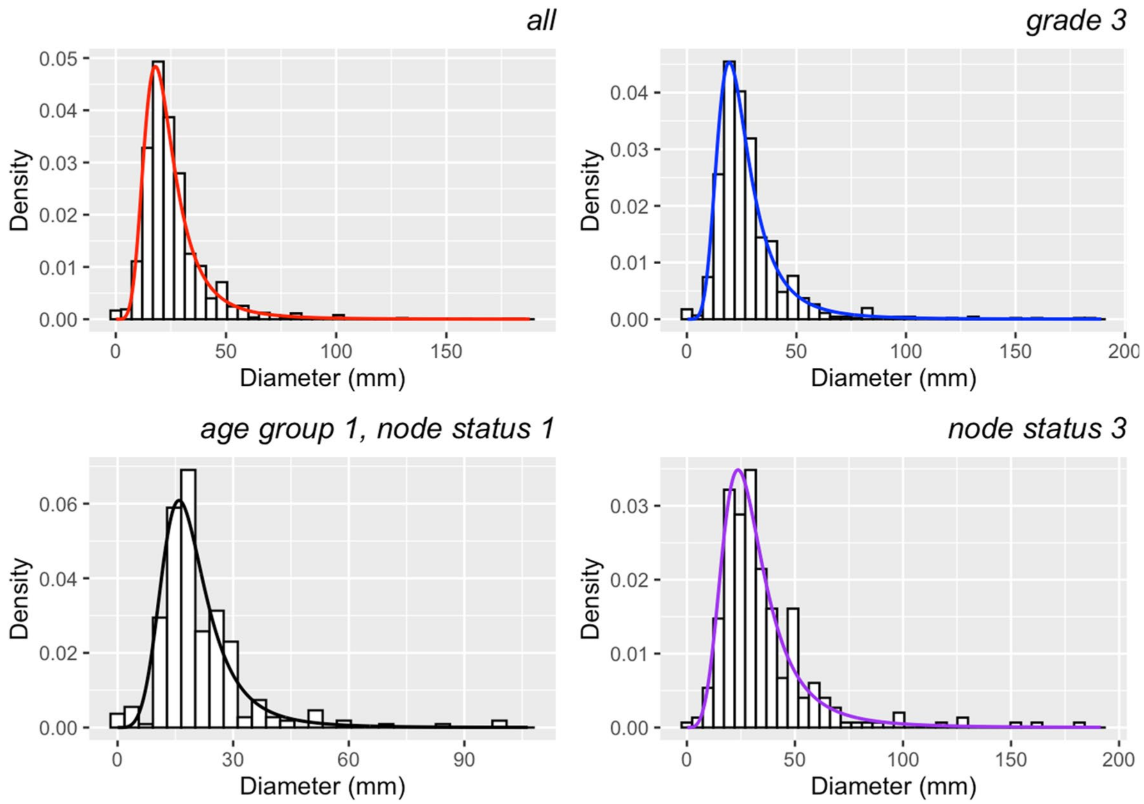

Fig. 4 Density histogram of tumor diameters for four different stratifications and the corresponding best fits with the Dagum probability distribution (see Eq. (1) in text)

diameters into 14 intervals and computed the mean survival for each interval. Figure 7 compares these results (for arbitrarily chosen cases) with predictions from our model survival function (Eq. (2)) using parameters from Table 1. The scatterplots show large fluctuations as a function of diameter, the direct result of a relatively small number of data points within each segment. Additionally, reliable results for larger diameters are absent due to the lack of significant data for $D > 80$ mm. Nonetheless, the consistency of the model prediction is apparent in each case and it provides confidence in the survival function and the parameters derived above.

Fraction-node-positivity and metastasis rates

To make a quantitative connection between node positivity and metastasis rates, we computed the fraction of patients with at least one positive node (i.e., node status 2 or 3) for small ranges of tumor diameters. The results are shown as data points (open squares) in Fig. 8. If we assume metastasis to occur homogeneously from the tumor surface [19, 20], the rate of metastasis for a tumor of diameter D should be equal to mD^2, where m is a metastasis rate constant. Assuming metastasis to be a Poisson process [21], the total probability of spread to any node during the lifetime of the tumor, i.e., during its entire growth time from size 0 (at the time of inception) to size D (at the time of surgery) is given by:

$$P(\text{node positive}) = 1 - e^{-\int_0^{t_1} mD^2(\tau)d\tau}, \tag{5}$$

where “node positive” means that at least one lymph node has been affected by metastatic spread, $D(\tau)$ is a function that represents how the tumor grows with time, t_1 is the age of the tumor at the time of surgery, and $D(t_1) = D$ is the size of the tumor at the time of surgery. Given that growth rates cannot be obtained from survival data, we need to use growth models developed in the literature. To this end, we considered two different growth models for the primary breast tumor, i.e., a Logistic growth model [22] and a Gompertz growth model [23]. In both these models, the tumor growth rate follows a lognormal distribution with finite standard deviation. Explicitly, the growth models (in terms of diameter) are as follows:

Logistic growth model [22]:

Table 1 Optimized survival parameters for different cofactor sets (fixed $n = 0.8$). These parameters are to be used with Eq. (2) to estimate the survival probability for a given tumor diameter D (in mm) and a given time of survival t (in years)

z	$k(z)$	$\alpha(z)$
All	0.00274	1.196
Node 1	0.00130	1.417
Node 2	0.00299	1.167
Node 3	0.00561	1.137
Age group 1 (younger than 55)	0.00419	0.883
Age group 2 (55 ≤ age < 65)	0.00207	1.248
Age group 3 (65 and over)	0.00208	1.445
Tumor grade 1	0.00076	1.609
Tumor grade 2	0.00112	1.521
Tumor grade 3	0.00474	1.022
Node 1, age group 1	0.00144	1.178
Node 1, age group 2	0.00137	1.353
Node 1, age group 3	0.00100	1.668
Node 2, age group 1	0.00496	0.834
Node 2, age group 2	0.00059	1.713
Node 2, age group 3	0.00245	1.428
Node 3, age group 1	0.00659	1.030
Node 3, age group 2	0.00607	1.072
Node 3, age group 3	0.00482	1.279

$$D(\tau) = D_{\max}\left[1 + \left\{\left(D_{\max}/D_{cell}\right)^{3/4} - 1\right\}e^{-\kappa\tau/4}\right]^{-4/3} \tag{6a}$$

Gompertz growth model [23]:

$$D(\tau) = D_{cell}\left(D_{\max}/D_{cell}\right)^{1-e^{-\kappa\tau}} \tag{6b}$$

In the above, $D_{\max}$ is the theoretical maximum diameter and D_{cell} the diameter of a single cell (tumor size at time 0). In Eq. (6a), the growth parameter κ follows a lognormal distribution with mean 1.07 and standard deviation 1.14 [22], while in Eq. (6b) the growth parameter κ follows a lognormal distribution with mean -2.9 and standard deviation 0.71 [23]. We would like to note that in Norton's original paper [23], time origin $t = 0$ was defined when tumor reaches a size of $N(0) = 4.8 \times 10^9$ cells, while in our definition (Eq. (6b)) time starts when tumor is of size 1 cell. This translational shift in time does not cause any change in the quantitative interpretation of the growth parameter κ (which Norton calls b). In the analysis below, we chose $D_{\max} = 180$ mm and $D_{cell} = 0.0124$ mm (using a spherical cell volume of 10^{-6}mm^3).

The metastasis rate constant (m) is also expected to vary from person to person, and we assume a lognormal distribution. Past studies indicate a positive correlation between tumor growth and metastasis rates [24, 25]. Thus, we assume that $(\ln(m), \ln(\kappa))$follows a bivariate normal distribution with some positive correlation coefficient ρ, i.e., $(\ln(m), \ln(\kappa)) \sim N(\mu, \Sigma)$, with the mean and covariance matrices given by:

$$\mu = \begin{bmatrix}\mu_m \\ \mu_g\end{bmatrix} \text{ and } \Sigma = \begin{bmatrix}\sigma_m^2 & \rho\sigma_m\sigma_g \\ \rho\sigma_m\sigma_g & \sigma_g^2\end{bmatrix}, \tag{7}$$

where the subscripts "m" and "g" refer to metastasis and growth parameters, respectively.

The growth rate distribution parameters (μ_g, σ_g) were chosen from literature values of marginal lognormal distributions for logistic growth [22] and Gompertz growth [23]. To determine the metastasis rate distribution parameters μ_m, σ_m we use the following strategy:

1. assume a positive value of the growth-metastasis log-log correlation (ρ) and keep it constant;
2. choose a specific value of marginal metastasis parameters (μ_m, σ_m);
3. draw a large number (1000) of random pairs $(\ln(m), \ln(\kappa))$ from a bivariate normal distribution (Eq. (7)), compute P(nodepositive) as a function of D (using Eq. (5)) for each drawn value of m, average over all 1000 drawings, and compare this average P(nodepositive) curve to the open squares of Fig. 8;
4. repeat steps (2) and (3) (for a fixed ρ) until the distribution-averaged P(nodepositive) curve has minimum root-mean-square-error (RMSE) with respect to the open squares of Fig. 8.

For a given value of ρ, we were able to find an optimized pair (μ_m, σ_m), which yields a result quantitatively similar to the dashed curve of Fig. 8. Table 2 lists these optimized metastasis parameters for each of the two growth models for a few different values of the correlation coefficient ρ. Table 2 shows that with increase in ρ the standard deviation σ_m increases, while the mean μ_m remains unchanged. However, the metastasis parameter values are sensitive to the growth model and parameters, which is not unexpected, given noticeable differences between the growth parameters of refs. [22, 23]. Had a tumor growth model existed on the METABRIC population itself, that would have been the most appropriate to use for this dataset.

Finally, we would like to clarify that metastasis rates discussed here pertain only to spread from the primary tumor to the lymph nodes. There are additional processes whereby metastatic spread can occur from the primary tumor and the lymph nodes to distant organs. Extracting such rates from survival data would require complex biological models of direct and indirect spread and assumptions relating a critical level of metastasis to subsequent organ failure and death.

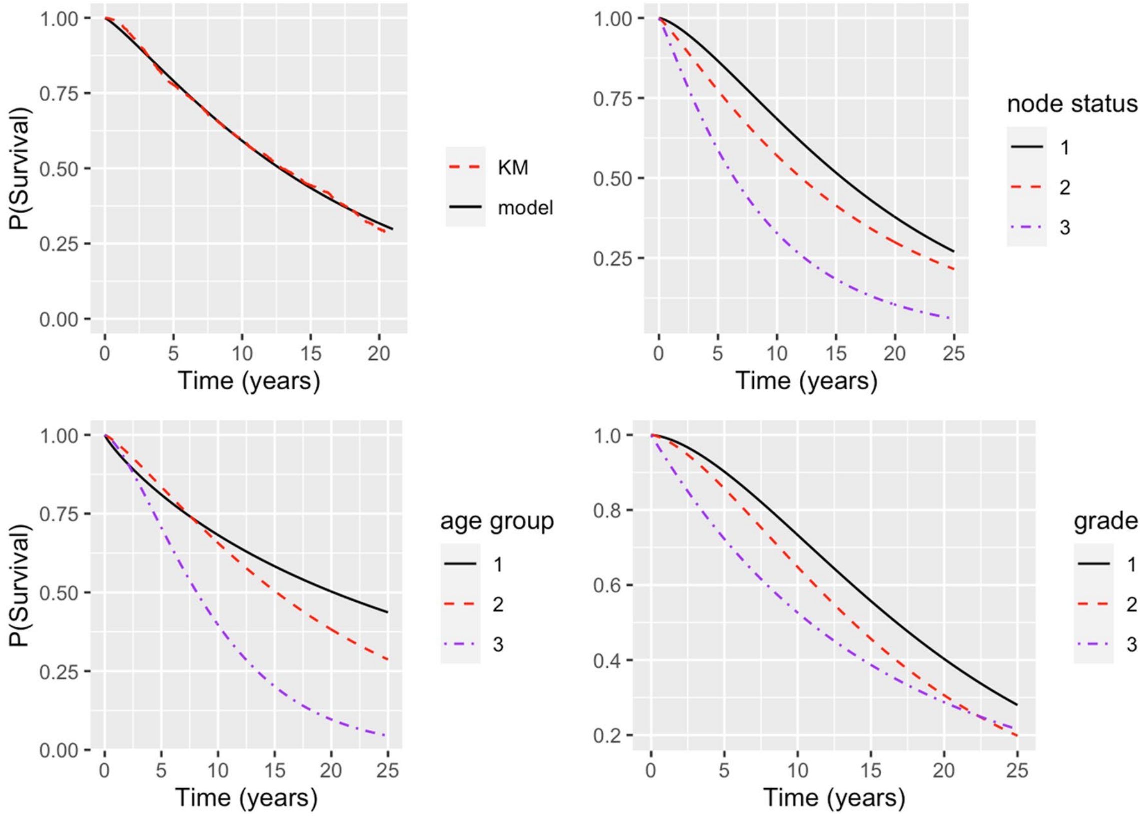

Fig. 5 Model-predicted mean survival probability using Eqs. (2) and (4) for the four cases corresponding to Fig. 2. **Top left** whole dataset, **top right** for each node status, **bottom left** for each age group, **bottom right** for each tumor grade. For direct comparison, **top left** also includes the KM mean survival curve from Fig. 2a

Summary

The aim in this work was to develop a model for predicting survival probability as a function of continuous time and tumor diameter for different cofactors such as node status, patient age group, and tumor grade. In order to overcome data sparsity shortcomings within small diameter ranges in a moderately-sized dataset like the METABRIC dataset, we adopted the strategy of representing the tumor diameter distribution among patients with a Dagum probability distribution [8], and then optimizing the model parameters to best match the corresponding KM mean survival curve. By analyzing the METABRIC dataset [9] we observe that the Cox proportional hazard model is not applicable for the cofactors of interest, i.e., the hazard ratio between different node statuses, age groups, or tumor grade levels do not remain constant as a function of time. Our model survival function intrinsically takes this into account by incorporating cofactor-dependent exponents (Eq. (2)), along with marginalization over diameter distribution (Eq. (4)). These parameters can be readily used to estimate the survival probability of a patient for any specified length of time beyond surgery. Such an approach was found to have accurate predictive power for mean survival probability for different cofactor combinations and was able to flexibly reproduce unexpected features in the data, e.g., the reversal of survival probabilities between age groups 1 and 2 as a function of time.

Finally, by studying the fraction of node-positive patients as a function of tumor diameter, we show how to decipher

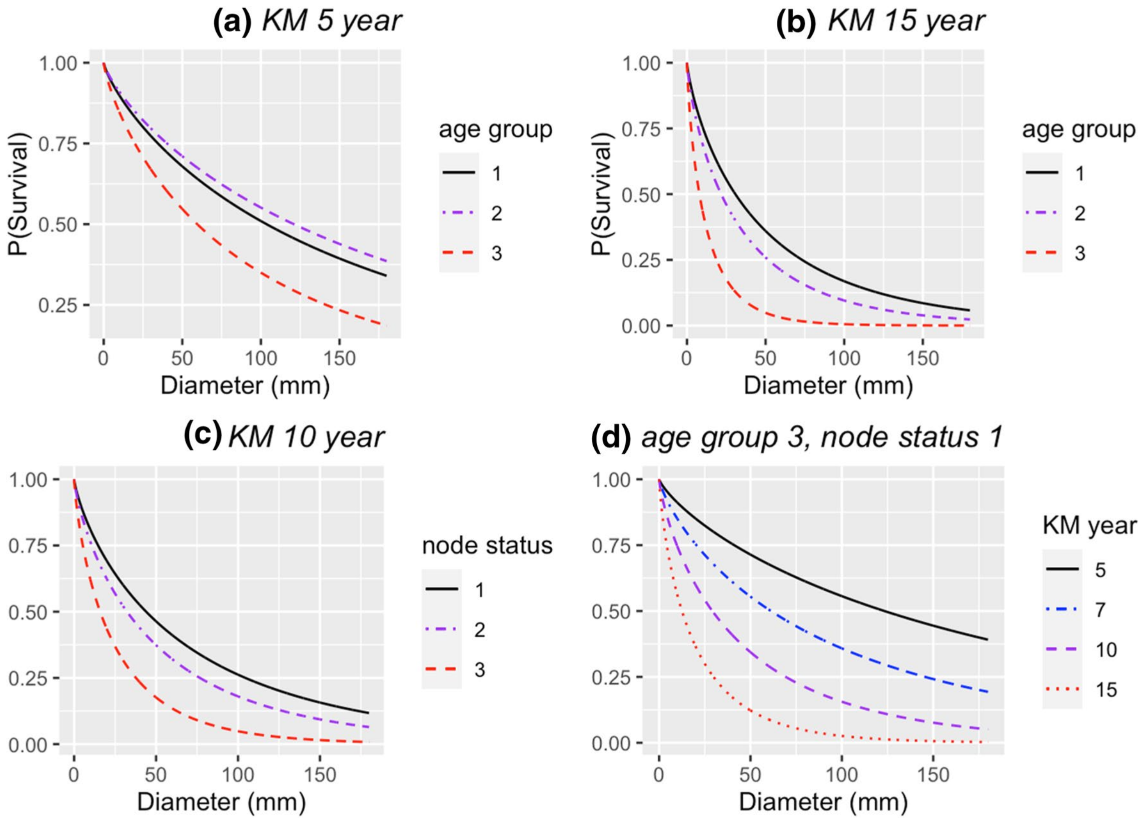

Fig. 6 Model-predicted survival probability as a function of tumor diameter using Eq. (2): **a** KM5 curves for three age groups, **b** KM15 curves for three age groups, **c** KM10 curves for the three nodes, **d** KM5, KM7, KM10, KM15 curves for node 1, age group 3

metastasis rates from the primary tumor surface to the lymph nodes (prior to tumor removal via surgery). More specifically, assuming known models for tumor growth rate κ from the literature [22, 23], assuming a metastasis rate m proportional to the tumor surface area, and assuming a bivariate lognormal distribution of (m, κ), we determine the marginal patient-to-patient distribution of m (see Table 2).

A knowledge of the mean survival probability (along with uncertainty bounds) as a function of the primary tumor's size, grade, node status, and patient age can be of immense help in managing the patient's quality of life beyond surgery. Such knowledge can aid doctor's recommendation and patient's choice of postoperative treatment options. Additionally, knowing mean metastasis rates (and its variability) could enable doctors make more informed assessment on the progression of the disease, especially in cases where the metastatic tumors are too small to be detectable by current clinical means.

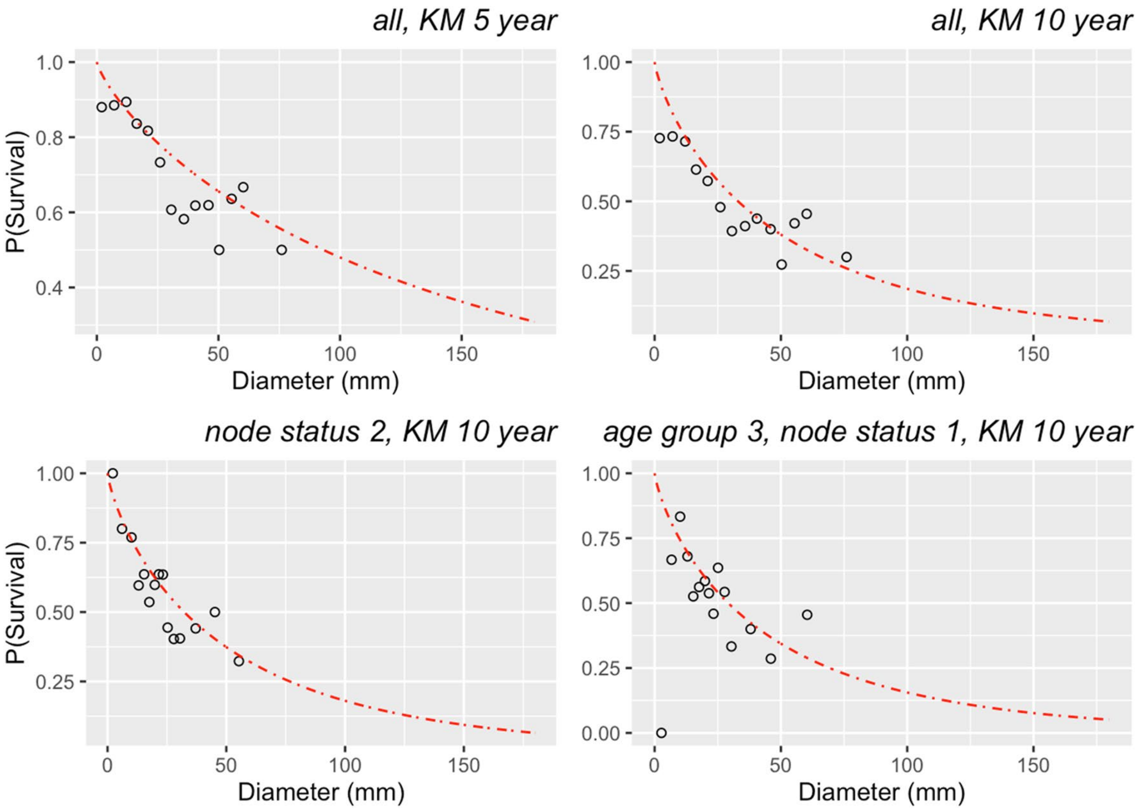

Fig. 7 Results from direct calculation of mean KM survival probability for various diameter segmentations within specified cofactor categories (open circles) compared with model-predicted probability (dashed line) using Eq. (2). The four cases have been arbitrarily chosen for illustration purposes. **Top left** whole dataset for KM 5 year, **top right** whole dataset for KM 10 year, **bottom left** node status 2 for KM 10 year, **bottom right** node status 1 and age group 3 for KM 10 year

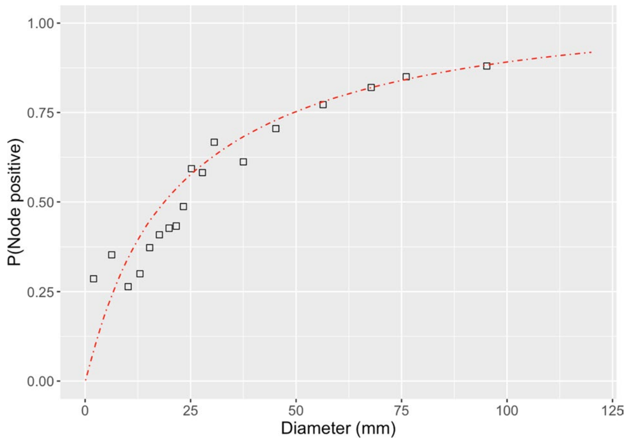

Fig. 8 Fraction-node-positive results for the whole METABRIC dataset. (open squares) Fraction of patients with at least one node affected (i.e. node status 2 or 3) for different diameter ranges directly computed from the METABRIC dataset; (dashed line) best fit using Eq. (5) that was averaged over the joint lognormal distribution of tumor growth and metastasis rates (Eq. (7)). The optimized metastasis parameters are listed in Table 2

Table 2 Optimized parameters for the marginal lognormal distribution of metastasis rate$m$obtained by fitting the observed fraction of node-positive patients in the METABRIC dataset as a function of tumor diameter$D$(open squares in Fig. 8)

Tumor growth model	ρ	(μ_m, σ_m)
Logistic [22]	0.1	(− 6.1, 2.45)
	0.3	(− 6.1, 2.7)
	0.5	(− 6.1, 3.05)
	0.7	(− 6.1, 3.25)
Gompertz [23]	0.1	(− 7.5, 3.5)
	0.3	(− 7.5, 3.7)
	0.5	(− 7.5, 3.85)
	0.7	(− 7.5, 4.0)

Results for two different growth models are indicated separately. The various parameters are defined in the text (Eq. (7))

Acknowledgements The author would like to sincerely thank Dr. J. S. Michaelson (Pathology, Harvard Medical School) for stimulating discussions at the beginning stages of this project.

Funding The author has no financial disclosures, conflicts of interest, or any funding sources to report for the manuscript.

Data availability The METABRIC clinical data is publicly available, see: https://www.cbioportal.org/study/summary?id=brca_metabric.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. See Wikipedia page: https://en.wikipedia.org/wiki/Epidemiology_of_breast_cancer
2. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
3. Michaelson JS et al (2002) Predicting the survival of patients with breast carcinoma using tumor size. Cancer 95:713–723
4. Michaelson JS et al (2003) The effect of tumor size and lymph node status on breast carcinoma lethality. Cancer 98:2133–2143
5. Chen LL et al (2009) The impact of primary tumor size, nodal status, and other prognostic factors on the risk of cancer death. Cancer 115:5071–5083
6. Michaelson JS et al (2011) Improved web-based calculators for predicting breast carcinoma outcomes. Breast Cancer Res Treat 128:827–835
7. Wang R et al (2019) The clinicopathological features and survival outcomes of patients with different metastatic sites in stage IV breast cancer. BMC Cancer 19:1091
8. Dagum C (1977) A new model of personal income distribution: specification and estimation. Econ Appl 30:413–437. Also see the following Wikipedia page https://en.wikipedia.org/wiki/Dagum_distribution
9. Curtis C et al (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486:346–352
10. Fletcher MNC et al (2013) Master regulators of FGFR2 signalling and breast cancer risk. Nat Commun 4:2464
11. Boughorbel S et al (2016) Model comparison for breast cancer prognosis based on clinical data. PLoS ONE. https://doi.org/10.1371/journal.pone.0146413
12. Chen R, Goodison S, Sun Y (2019) Molecular profiles of matched primary and metastatic tumor samples support an evolutionary model of breast cancer. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-19-2296
13. Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. Biometrika 69:553–566
14. Cox DR (1972) Regression models and life tables. J R Stat Soc B 34:187–220
15. Cox DR (1975) Partial likelihood. Biometrika 62:269–276
16. Grambsch P, Therneau T (1994) Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81:515–526
17. Miller RG (1977) Survival analysis. Wiley, New York
18. Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data. Springer, New York
19. Iwata K, Kawasaki K, Shigesada N (2000) A dynamical model for the growth and size distribution of multiple metastatic tumors. J Theor Biol 203:177–186
20. Maiti E (2012) Monte Carlo simulation-based approach to model the size distribution of metastatic tumors. Phys Rev E 85:012901
21. Cox DR, Isham V (1980) Point processes. Chapman & Hall, London
22. Weedon-Fekjær H et al (2008) Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res 10:R41
23. Norton LA (1988) Gompertzian model of human breast cancer growth. Cancer Res 48:7067–7071
24. Klein CA (2010) Tumour cell dissemination and growth of metastasis. Nat Rev Cancer 10:156
25. Yoo T-K et al (2015) In vivo tumor growth rate measured by us in preoperative period and long-term disease outcome in breast cancer patients. PLoS ONE. https://doi.org/10.1371/journal.pone.0144144