Karolinska Institutet

A Natural History and Copula Based Joint Model for Regional and Distant Breast Cancer Metastasis

Alessandro Gasparini · alessandro.gasparini@ki.se

2022-03-17

Motivation

To fully understand the prognosis of breast cancer, we need information on regional and distant metastasis.

Motivation

To fully understand the prognosis of breast cancer, we need information on regional and distant metastasis.

Past work focussed on regional or distant metastasis alone.

Motivation

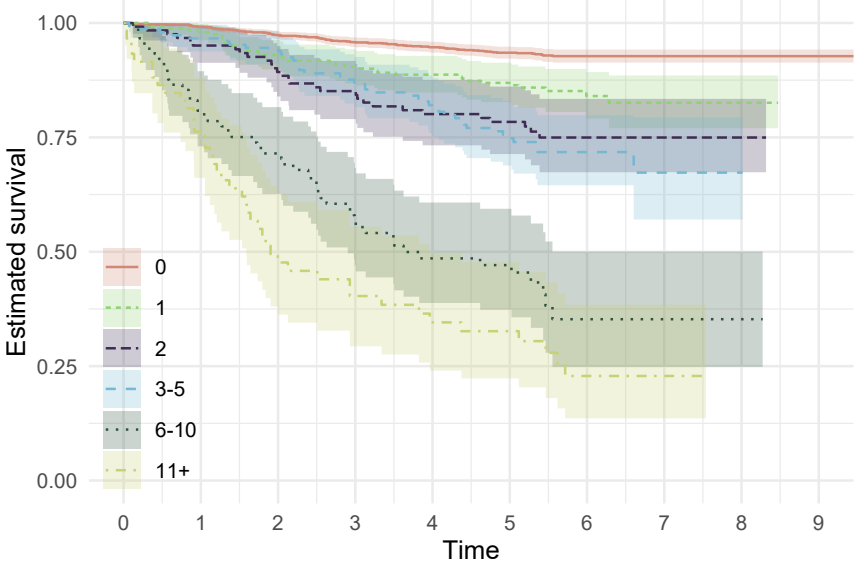
To fully understand the prognosis of breast cancer, we need information on regional and distant metastasis.

Past work focussed on regional or distant metastasis alone.

We want to develop a joint model for the two combined.

This is joint work with Keith Humphreys, who talked about the background of this project in more detail last week.

Time to Metastasis and Affected Lymph Nodes are Correlated



Crude data from CAHRES.

Modelling Tumour Growth

Exponential growth of the tumour:

$$V(t|r) = V_{\text{Cell}} \exp(t/r)$$

Modelling Tumour Growth

Exponential growth of the tumour:

$$V(t|r) = V_{\text{Cell}} \exp(t/r)$$

A random effect on r to allow for heterogeneity:

$$f_R(r) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} r^{\tau_1 - 1} \exp(-\tau_2 r), \; r \geq 0,$$

Modelling Tumour Growth

Exponential growth of the tumour:

$$V(t|r) = V_{\text{Cell}} \exp(t/r)$$

A random effect on r to allow for heterogeneity:

$$f_R(r) = \frac{\tau_2^{\tau_1}}{\Gamma(\tau_1)} r^{\tau_1 - 1} \exp(-\tau_2 r),\ r \geq 0,$$

Finally, in the absence of screening, the rate of symptomatic detection at time $T_{\text{det}} = t$ is proportional to the size of the tumour:

$$P(T_{\text{det}} \in [t, t + dt) | T_{\text{det}} \geq t, R = r) = \eta V(t, r) dt + o(dt),\ t \geq t_0$$

Modelling Spread to the Lymph Nodes (1)

This is based on previous work by Isheden *et al.*

The model for spread to the lymph nodes (seeding) is based on a non-homogeneous Poisson Process with intensity function

$$\lambda(t, r, s^*) = s^* D(t, r)^{k_N} D'(t, r),$$

where $D(t, r)$ is the number of times the cells in the tumour have divided and $D'(t, r)$ is the rate of cell division in the tumour.

Modelling Spread to the Lymph Nodes (1)

This is based on previous work by Isheden *et al.*

The model for spread to the lymph nodes (seeding) is based on a non-homogeneous Poisson Process with intensity function

$$\lambda(t, r, s^*) = s^* D(t, r)^{k_N} D'(t, r),$$

where $D(t, r)$ is the number of times the cells in the tumour have divided and $D'(t, r)$ is the rate of cell division in the tumour.

Under the assumption of a time to clinical detectability of t_0, the corresponding cumulative intensity function for detectable lymph node metastases is

$$\Lambda(t - t_0, r, s) = s \left[\log \left(\frac{V(t, r)}{V_0} \right) \right]^{k_N + 1}, t \geq t_0 \tag{1}$$

with $s = s^*/[(k_N + 1)(\log 2)^{k_N + 1}]$.

Modelling Spread to the Lymph Nodes (2)

Assuming a Gamma(γ_1, γ_2) random effect on s to allow for heterogeneity in spread, Isheden *et al.* showed that the probability of $N = n$ clinically detectable lymph nodes is independent of both S and R.

Modelling Spread to the Lymph Nodes (2)

Assuming a Gamma(γ_1, γ_2) random effect on s to allow for heterogeneity in spread, Isheden *et al.* showed that the probability of $N = n$ clinically detectable lymph nodes is independent of both S and R.

This follows a negative binomial distribution $NB(l, p)$ with size $l = \gamma_1$ and probability $p = 1 - [(\log(v/V_0)^{k_N+1}]/[(\log(v/V_0))^{k_N+1} + \gamma_2]$.

Modelling Spread to the Lymph Nodes (2)

Assuming a Gamma(γ_1, γ_2) random effect on s to allow for heterogeneity in spread, Isheden *et al.* showed that the probability of $N = n$ clinically detectable lymph nodes is independent of both S and R.

This follows a negative binomial distribution $NB(l, p)$ with size $l = \gamma_1$ and probability $p = 1 - [(\log(v/V_0)^{k_N+1}]/[(\log(v/V_0))^{k_N+1} + \gamma_2]$.

The probability of having $N = n$ affected lymph nodes given a tumour volume $V = v$ is:

$$P(N = n|V = v) = \frac{\Gamma(n + l)}{\Gamma(l)n!} p^l (1 - p)^n,$$

Modelling Distant Metastatic Spread (1)

The model for time to distant metastatic spread is also based on a similar non-homogeneous Poisson process but with parameters σ^* and k_W. In previous work we derived a survival model for time to detection of distant metastasis; here, we extend that model to allow for between-subject heterogeneity.

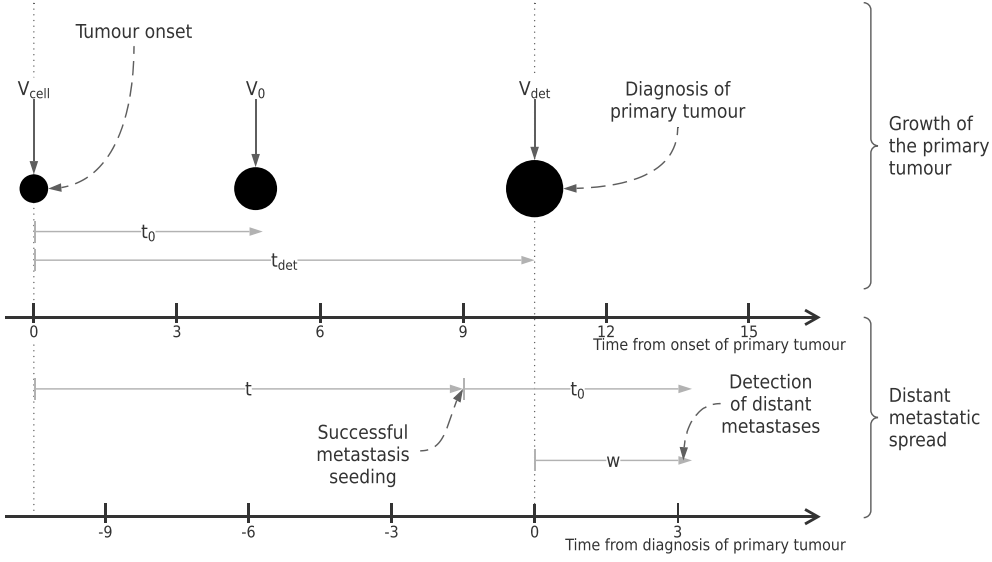
Modelling Distant Metastatic Spread (1)

The model for time to distant metastatic spread is also based on a similar non-homogeneous Poisson process but with parameters σ^* and k_W. In previous work we derived a survival model for time to detection of distant metastasis; here, we extend that model to allow for between-subject heterogeneity.

Some key model assumptions:

- Metastatic seeding completely stops at diagnosis of the primary;
- Already seeded, successful colonies are not affected by surgery following diagnosis/treatment;
- Times from seeding to detection are the individual specific times t_0.

Modelling Distant Metastatic Spread (2)



Modelling Distant Metastatic Spread (3)

We can derive the following density and survival functions for time to detection of distant metastasis:

$$f_{W|V=v,R=r}(w) = \frac{k_W+1}{r}\left(\frac{w}{r}+\log\frac{v}{V_0}\right)^{k_W}\frac{\omega_1{\omega_2}^{\omega_1}}{\left[\omega_2+\left(\frac{w}{r}+\log\frac{v}{V_0}\right)^{k_W+1}\right]^{\omega_1+1}},$$

$\forall\ 0 \leq w \leq r\log(V_0/V_{\text{Cell}})$.

$$S_{W|V=v,R=r}(w) = \begin{cases} \left\{\omega_2/\left[\omega_2+\left(\frac{w}{r}+\log\frac{v}{V_0}\right)^{k_W+1}\right]\right\}^{\omega_1} & \text{if } 0 \leq w \leq r\log(V_0/V_{\text{Cell}}) \\ \left\{\omega_2/\left[\omega_2+\left(\log\frac{v}{V_{\text{Cell}}}\right)^{k_W+1}\right]\right\}^{\omega_1} & \text{if } w > r\log(V_0/V_{\text{Cell}}) \end{cases}$$

Joint Modelling

First, we need to define the joint distribution of the number of affected lymph nodes $N = n$ and the time to first detected distant metastasis $W = w$, given tumour size at detection $V = v$ and inverse growth rate $R = r$:

$$f_{N,W|V=v,R=r}(n, w)$$

There are several ways to connect the two processes. For instance, we could specify correlated random effects for the spread rates; however, this is computationally demanding.

Joint Modelling

First, we need to define the joint distribution of the number of affected lymph nodes $N = n$ and the time to first detected distant metastasis $W = w$, given tumour size at detection $V = v$ and inverse growth rate $R = r$:

$$f_{N,W|V=v,R=r}(n, w)$$

There are several ways to connect the two processes. For instance, we could specify correlated random effects for the spread rates; however, this is computationally demanding.

Instead, we take a copula modelling approach:

- We have already specified the marginal distributions of N and W,
- It is reasonable in the absence of a clear underlying biological model.

Copula

A copula is defined as a multivariate cumulative distribution function (CDF) for which the marginal probability distributions are uniform on the interval $[0, 1]$.

Formally, if F is a bivariate CDF with univariate CDF margins F_1, F_2 then, according to Sklar's theorem, for every bivariate distribution there exists a copula representation such that

$$F(x_1, x_2|\theta) = C(F_1(x_1), F_2(x_2); \theta)$$

for a certain parameter (or vector of parameters) θ.

Joint Copula Modelling

Let C be a bivariate copula and $F_{N|V=v,R=r}(n)$ and $F_{W|V=v,R=r}(w)$ be the cumulative distribution functions of affected lymph nodes at detection and time to distant metastasis, respectively.

The joint bivariate cumulative distribution can therefore be defined using the copula C as

$$F_{N,W|V=v,R=r}(n,w) = C(F_{N|V=v,R=r}(n), F_{W|V=v,R=r}(w))$$

The joint bivariate density function follows as:

$$f_{N,W|V=v,R=r}(n,w) = \frac{\partial^2 \, C(F_{N|V=v,R=r}(n), F_{W|V=v,R=r}(w))}{\partial n \, \partial w}$$

Possible Copula Formulations

We focus on Achimedean copulae:

Name of Copula	Bivariate Copula $C(u, v; \theta)$	Domain of θ	Possible Correlation τ
Ali–Mikhail–Haq	$\frac{uv}{1-\theta(1-u)(1-v)}$	$\theta \in [-1, 1]$	$\tau \in [-0.18, 0.33]$
Clayton	$\left[\max\left\{u^{-\theta} + v^{-\theta} - 1; 0\right\}\right]^{-1/\theta}$	$\theta \in [-1, \infty) \backslash \{0\}$	$\tau \in [-1, 1] \backslash 0$
Frank	$-\frac{1}{\theta}\log\left[1 + \frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right]$	$\theta \in \mathbb{R} \backslash \{0\}$	$\tau \in [-1, 1] \backslash 0$
Gumbel	$\exp\left[-\left((-\log(u))^{\theta} + (-\log(v))^{\theta}\right)^{1/\theta}\right]$	$\theta \in [1, \infty)$	$\tau \in [0, 1]$
Product	uv	—	$\tau = 0$
Joe	$1 - \left[(1-u)^{\theta} + (1-v)^{\theta} - (1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta}$	$\theta \in [1, \infty)$	$\tau \in [0, 1]$

Another alternative is the Gaussian copula:

$$C_{\text{Gaussian}}(u, v; \theta) = \Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \theta\right),$$

Likelihood Function

In the absence of screening:

$$L^{\text{No Screening}} = f_{V_{\text{det}}}(v) \int_R P(N = n, W = w | V_{\text{det}} = v, R = r) f_{R|V_{\text{det}}=v}(r)\, dr$$

Likelihood Function

In the absence of screening:

$$L^{\text{No Screening}} = f_{V_{\text{det}}}(v) \int_R P(N = n, W = w | V_{\text{det}} = v, R = r) f_{R|V_{\text{det}}=v}(r)\, dr$$

For a screened population:

$$L^{\text{Screen Detection}} \propto P(B_0|V = v)P(V = v, N = n, W = w|A)P(B^c|A, V = v, N = n, W = w)$$

$$L^{\text{Symptomatic Detection}} \propto P(V_{\text{det}} = v, N = n, W = w|A)P(B^c|A, V_{\text{det}} = v, N = n, W = w)$$

Likelihood Function

In the absence of screening:

$$L^{\text{No Screening}} = f_{V_{\text{det}}}(v) \int_R P(N = n, W = w | V_{\text{det}} = v, R = r) f_{R|V_{\text{det}}=v}(r)\, dr$$

For a screened population:

$$L^{\text{Screen Detection}} \propto P(B_0|V = v)P(V = v, N = n, W = w|A)P(B^c|A, V = v, N = n, W = w)$$

$$L^{\text{Symptomatic Detection}} \propto P(V_{\text{det}} = v, N = n, W = w|A)P(B^c|A, V_{\text{det}} = v, N = n, W = w)$$

I will skip the details here, but please come talk to us if interested!

Model-Based Predictions

After fitting the joint copula model we can obtain a variety of predictions. Among others:

- Probability of having detected distant metastases at diagnosis of the primary tumour given size of the tumour and number of affected lymph nodes;

Model-Based Predictions

After fitting the joint copula model we can obtain a variety of predictions. Among others:

- Probability of having detected distant metastases at diagnosis of the primary tumour given size of the tumour and number of affected lymph nodes;
- Probability of having latent/undiagnosed distant metastases given size of the tumour and number of affected lymph nodes at diagnosis of the primary tumour;

Model-Based Predictions

After fitting the joint copula model we can obtain a variety of predictions. Among others:

- Probability of having detected distant metastases at diagnosis of the primary tumour given size of the tumour and number of affected lymph nodes;
- Probability of having latent/undiagnosed distant metastases given size of the tumour and number of affected lymph nodes at diagnosis of the primary tumour;
- Survival probability at any time $w^* > 0$ for the event of distant metastasis, conditional on characteristics observed at diagnosis and on being free of distant metastasis at that time;

Model-Based Predictions

After fitting the joint copula model we can obtain a variety of predictions. Among others:

- Probability of having detected distant metastases at diagnosis of the primary tumour given size of the tumour and number of affected lymph nodes;
- Probability of having latent/undiagnosed distant metastases given size of the tumour and number of affected lymph nodes at diagnosis of the primary tumour;
- Survival probability at any time $w^* > 0$ for the event of distant metastasis, conditional on characteristics observed at diagnosis and on being free of distant metastasis at that time;
- More *standard* quantities such as tumour doubling time, etc.

Application: Data

We analyse data from CAHRES, which consists of incident cases of postmenopausal breast cancer recorded in a case-control setting:

- Women born and residing in Sweden,
- Aged 50 – 74,
- Diagnosed with an incident primary invasive breast cancer between October 1st 1993 and March 31st 1995.

Furthermore,

- This was linked to data from the Swedish Cancer Registry and the Stockholm-Gotland Breast Cancer Registry, and
- An extension of the original case-control study collected mammographic images and screening histories from mammography screening units and radiology departments.

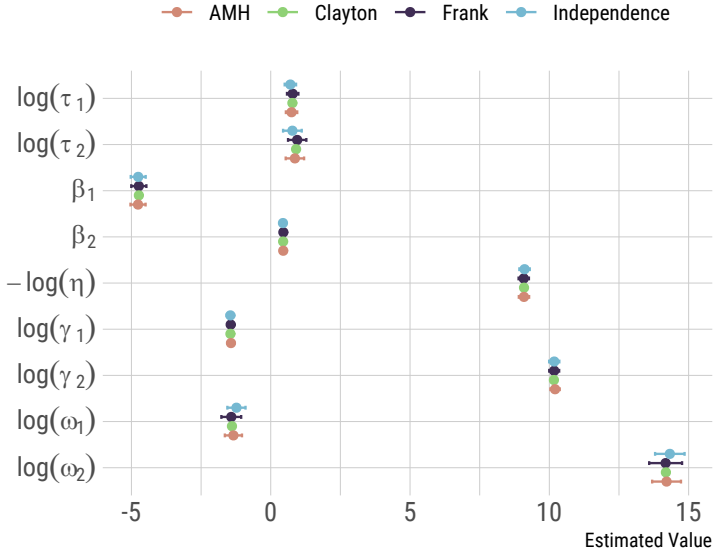
Application: Some Statistics

- 1581 women, of which:
  - 1019 (64.4%) detected through screening
  - 562 (35.6%) detected symptomatically
- Median tumour diameter at detection of 15 mm (I.Q.I. 10 – 22 mm);
- 1091 women (69.0%) had no affected lymph nodes at detection, 170 (10.8%) had one, 91 (5.8%) had two, 229 women (14.4%) had three or more;
- One woman had detected distant metastasis at the time of diagnosis of the primary tumour. During follow-up, 288 more women (18.2%) were diagnosed with distant metastasis;
- Median follow-up time was 5.50 years (95% C.I.: 5.41 – 5.59 years);
- Kendall's τ correlation between the lymph nodes and the times to distant metastasis was -0.15 (if discretising time: -0.17).

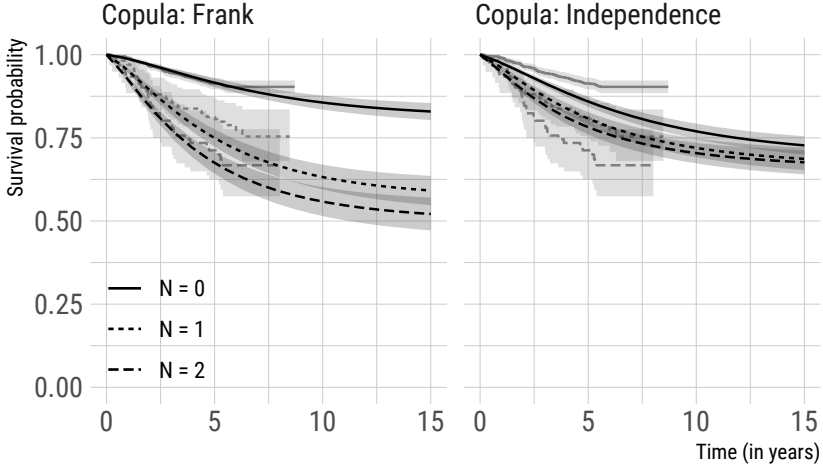
Application: Choice of the Copula Function

	Frank	Clayton	AMH	Independence
Log-likelihood	-6,380.31	-6,417.57	-6,394.91	-6,443.43
Kendall's τ	-0.33	-0.09	-0.18	—

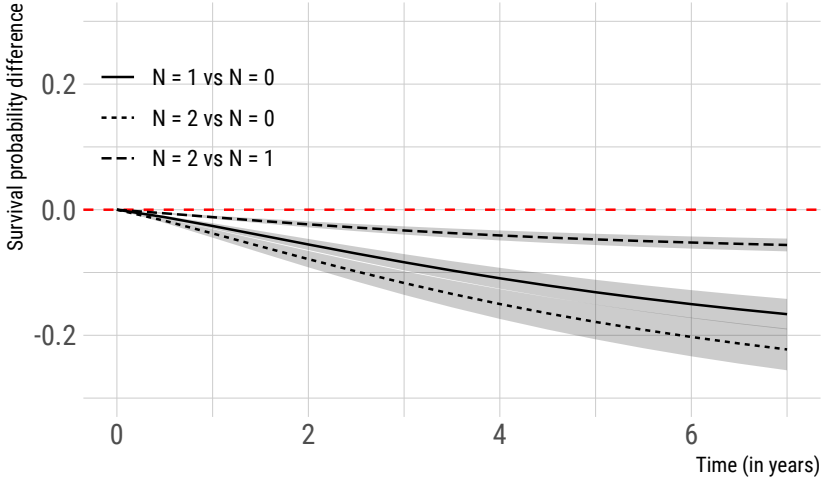
Application: Comparing Copulae



Application: Time to Distant Metastasis Predictions



Application: Standardised Survival Difference



Application: Cured Fraction

- Marginally over the overall observed covariates distribution: 0.697
- Marginally over number of affected lymph nodes:
  - Zero lymph nodes: 0.805
  - One lymph node: 0.553
  - Two lymph nodes: 0.479

This estimate is similar to that reported by Dal Maso et al. from the EUROCARE-5 study: 0.66 for breast cancers diagnosed in 2000.

Application: Microsimulation

Finally, we use the joint copula model to showcase its potential for microsimulation purposes, as it can connect the latent natural history of a tumour with the risk of future events.

For this purpose, we simulate 10 million tumours from the best fitting model (i.e., assuming a Frank copula) and we assess *what the 5-years risk of distant metastasis would be* in the counterfactual scenario of early detection.

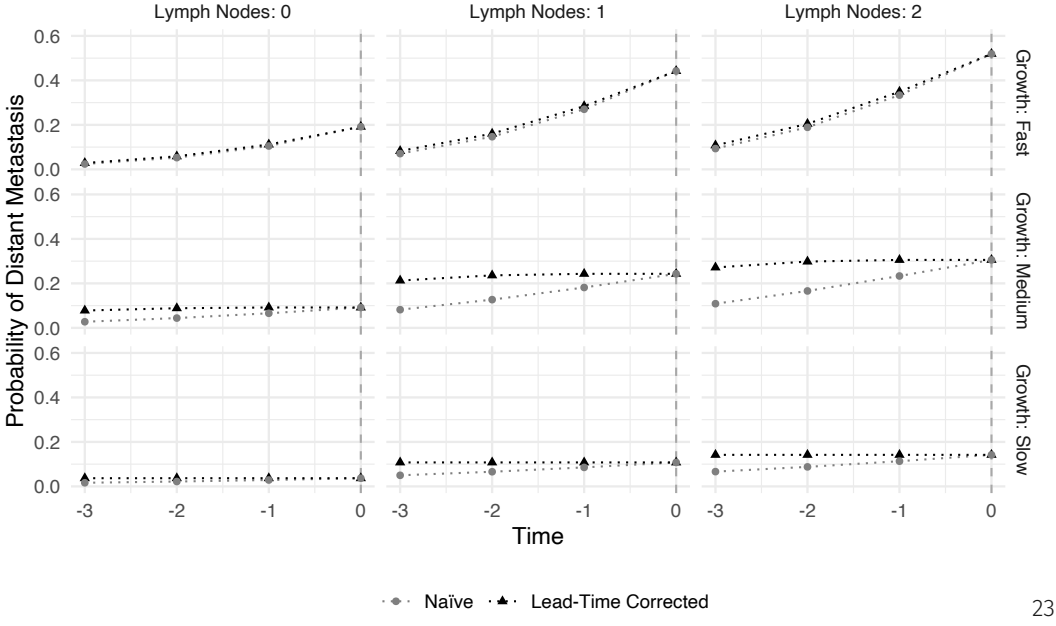
Application: Microsimulation

Finally, we use the joint copula model to showcase its potential for microsimulation purposes, as it can connect the latent natural history of a tumour with the risk of future events.

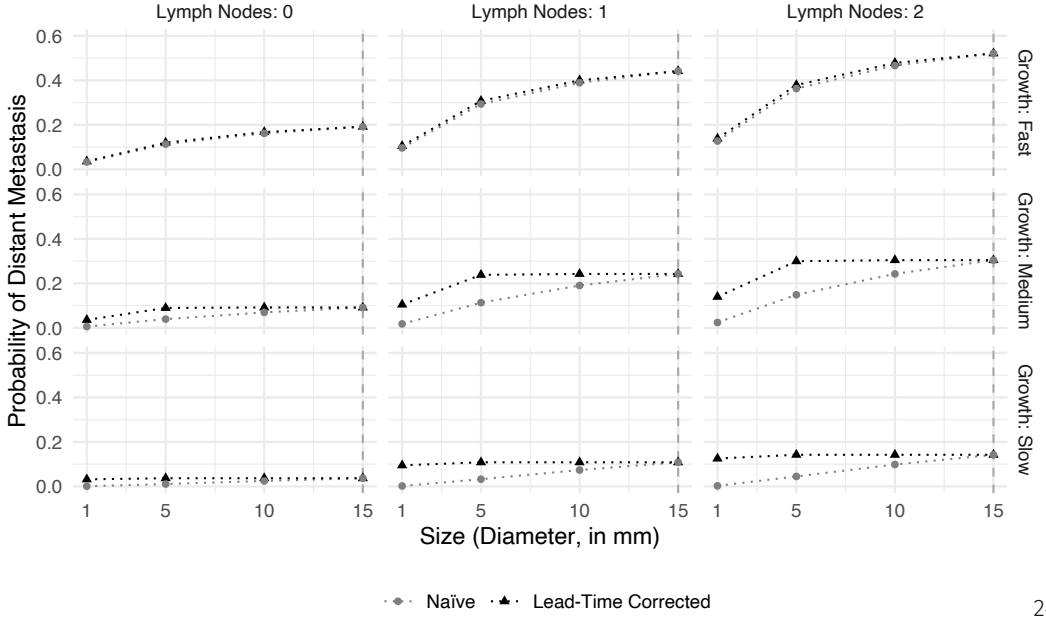
For this purpose, we simulate 10 million tumours from the best fitting model (i.e., assuming a Frank copula) and we assess *what the 5-years risk of distant metastasis would be* in the counterfactual scenario of early detection.

This quantity is likely affected by lead-time bias, but given that we know the counterfactuals, we can provide a *lead-time corrected estimate* as well.

Application: Early Detection



Application: Detecting Smaller Cancers



Wrap Up

1. We have introduced a joint, copula-based model for the latent growth of breast cancer, detection, spread to the lymph nodes, and distant metastatic spread.

Wrap Up

1. We have introduced a joint, copula-based model for the latent growth of breast cancer, detection, spread to the lymph nodes, and distant metastatic spread.
2. We have shown that this model was able to capture relevant patterns in data.

Wrap Up

1. We have introduced a joint, copula-based model for the latent growth of breast cancer, detection, spread to the lymph nodes, and distant metastatic spread.
2. We have shown that this model was able to capture relevant patterns in data.
3. We have demonstrated how a model of this kind could be used in microsimulation studies of breast cancer.

Wrap Up

1. We have introduced a joint, copula-based model for the latent growth of breast cancer, detection, spread to the lymph nodes, and distant metastatic spread.
2. We have shown that this model was able to capture relevant patterns in data.
3. We have demonstrated how a model of this kind could be used in microsimulation studies of breast cancer.
4. The model is of course not perfect, but it provides solid building blocks on which we could develop and extend upon, e.g., by directly modelling cancer-specific death within a unified framework.