


# Lymph node metastases in breast cancer: Investigating associations with tumor characteristics, molecular subtypes and polygenic risk score using a continuous growth model

Gabriel Isheden<sup>1</sup>  | Felix Grassmann<sup>1,2</sup> | Kamila Czene<sup>1</sup> | Keith Humphreys<sup>1</sup>

<sup>1</sup>Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

<sup>2</sup>The Institute of Medical Sciences, University of Aberdeen, Aberdeen, Scotland, UK

## Correspondence

Gabriel Isheden, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.  
Email: gabriel.isheden@ki.se

## Funding information

Cancerfonden, Grant/Award Number: 2020/0716; Vetenskapsrådet, Grant/Award Number: 2020-01302

## Abstract

We investigate the association between rate of breast cancer lymph node spread and grade, estrogen receptor (ER) status, progesteron receptor status, decision tree derived PAM50 molecular subtype and a polygenic risk score (PRS), using data on 10 950 women included from two different data sources. Lymph node spread was analyzed using a novel continuous tumor progression model that adjusts for tumor volume in a biologically motivated way and that incorporates covariates of interest. Grades 2 and 3 tumors, respectively, were associated with 1.63 and 2.17 times faster rates of lymph node spread than Grade 1 tumors ( $P < 10^{-16}$ ). ER/PR negative breast cancer was associated with a 1.25/1.19 times faster spread than ER/PR positive breast cancer, respectively ( $P = .0011$  and  $.0012$ ). Among the molecular subtypes luminal A, luminal B, Her2-enriched and basal-like, Her2-enriched breast cancer was associated with 1.53 times faster spread than luminal A cancer ( $P = .00072$ ). PRS was not associated with the rate of lymph node spread. Continuous growth models are useful for quantifying associations between lymph node spread and tumor characteristics. These may be useful for building realistic progression models for microsimulation studies used to design individualized screening programs.

## KEYWORDS

breast cancer, continuous growth model, lymph node metastases, molecular subtype, polygenic risk score

## What's new?

Breast cancer aggressiveness is reflected in the tumour's propensity to spread to the lymph nodes, in many cases a precursory step of distant metastatic spread. Here, the authors apply a novel continuous tumour progression model to estimate the rate of lymph node spread during the tumour's preclinical phase based on grade, oestrogen receptor status, progesterone receptor status, molecular subtype, and polygenic risk score. Combining two datasets with a total of 10,950 women with invasive breast cancer, they show that quantifying tumour aggressiveness using continuous growth models may prove useful in the future era of individualised screening and treatment.

**Abbreviations:** CAHRES, The Cancer and Hormone Replacement Study; CISNET, The Cancer Intervention and Surveillance Network; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; iCOGs, a custom Illumina iSelect genotyping array; PAM50, a 50-gene signature that classifies breast cancer into molecular intrinsic subtypes; PR, progesteron receptor; PRS, polygenic risk score; ST01-08, a cohort of breast cancer cases from the Stockholm-Gotland regional breast cancer register.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *International Journal of Cancer* published by John Wiley & Sons Ltd on behalf of UICC.

## 1 | INTRODUCTION

Breast cancer is a heterogeneous disease. Different subtypes of breast cancer grow and spread at different rates, and they react differently to treatment. Recently, there has been an interest in statistically modeling breast cancer heterogeneity in terms of disease progression.<sup>1,2</sup> Number of lymph node metastases present at diagnosis is associated with long-term breast cancer prognosis.<sup>3,4</sup> It is therefore clinically relevant to understand breast cancer heterogeneity in terms of lymph node metastases at diagnosis. The purpose of this article is to investigate the association between breast cancer tumor characteristics, including molecular subtype, and rate of lymph node spread.

In 2018, the Cancer Intervention and Surveillance Network (CISNET), a consortium of research groups from six different universities, evaluated the contributions of screening and treatment to breast cancer mortality between 2000 and 2012 by molecular subtype, based on estrogen receptor (ER) status and human epidermal growth factor receptor 2 (HER2) status.<sup>1</sup> The group estimated that in 2012, compared to baseline mortality rates, total reduction in mortality rate from interventions was 49%, ranging from 37% for ER−/HER2− breast cancer to 58% for ER+/HER2+ breast cancer. The contributions of screening and treatment differed substantially between molecular subtypes. Screening was estimated to contribute to 31% of the total mortality reduction for ER−/HER2− and to 48% of the reduction for ER+/HER2+ breast cancer. Rueda et al<sup>2</sup> investigated the rate of recurring breast cancer by breast cancer molecular subtypes. They used a semi-Markov model; molecular information was based on PAM50 subtypes<sup>5</sup> and integrative subtypes. After surgery, state transition for local recurrence differed across the PAM50 molecular subtypes: Basal-like breast cancer predominantly recurred within the first 5 years, whereas luminal A breast cancer recurred almost uniformly throughout the 20-year study period. Some of these differences will be due to heterogeneity in rates of breast cancer spread.

Each group in CISNET has developed a breast cancer natural history model.<sup>6</sup> These models are all stage based and include a localized tumor stage, a regionally spread stage and distant metastatic stage. Of these approaches, the University of Wisconsin group uses, arguably, the most sophisticated stage model: a continuous time spread process based on Shwartz.<sup>7</sup> The model assumes that tumor volume follows an exponential Gompertz function with decelerating doubling time, individually assigned doubling times, and that the instantaneous rate of lymph node spread at time  $t$  is equal to  $\lambda(t) = b_1 + b_2V(t) + b_3V'(t)$ , where  $V(t)$  is tumor volume at time  $t$ ,  $V'(t)$  is the rate of growth at time  $t$ , and  $b_1$ ,  $b_2$  and  $b_3$  are constants. In Isheden et al,<sup>8</sup> it was shown that the model of Shwartz suffers from two weaknesses: firstly, it implies that slow growing tumors have a higher degree of lymph node spread compared to fast growing tumors, and, secondly, the model implies either an unrealistically high degree of lymph node spread for large tumors or an unrealistically low degree of lymph node spread for small tumors. Based on two independent data sets, it was shown that lymph node spread following an inhomogeneous Poisson process with rate  $\lambda(t)$  proportional to the number of times the tumor cells have divided,  $D(t)$ , to the power four, and the rate of cell division in the tumor  $D'(t)$ , that is,  $\lambda(t) = \sigma D(t)^4 D'(t)$ , combined with a gamma

distributed random effect for individual spread  $\sigma$ , gives a significantly better model fit compared to the model of Shwartz,<sup>7</sup> and the lymph node spread model of Hanin and Yakovlev.<sup>9</sup> Here, we base our analyses on the lymph node metastases modeling approach of Isheden et al<sup>8</sup> and a recent extension of the model to include a covariate effect on the rate of lymph node spread.<sup>10</sup>

In this article, we use a natural history lymph node spread regression model to quantify the rate of lymph node spread based on grade, ER status, progesteron receptor (PR) status, molecular subtype and polygenic risk score (PRS).

## 2 | METHODS

### 2.1 | Data

We include two independent data sources for our study: the Cancer and Hormone Replacement Study, CAHRES; and breast cancer cases from the Stockholm-Gotland regional breast cancer register, here abbreviated as ST01-08. Ethical approvals were obtained for both data sources.

CAHRES is a case control study, consisting of all Swedish born women between the ages of 50 and 74, who were diagnosed with invasive breast cancer in Sweden from October 1993 to March 1995. The study had a participation rate of 84% ( $n = 3345$ ), and patients were matched to randomly selected controls from the general population based on the expected age frequency distribution of the cases. For the purpose of our study, we use only the cases. Information on tumor size, degree of lymph node spread, grade, ER status and PR status was collected from the Swedish Cancer Registry and the Stockholm-Gotland Breast Cancer Registry. The collection of this data has been described previously by Rosenberg et al<sup>11,12</sup> and Eriksson et al.<sup>13</sup> Tumor size was categorized into millimeter diameter intervals, lymph node involvement categorized according to number of lymph nodes affected by metastases and grade categorized as 1, 2 or 3. Tumors were considered ER or PR positive if they contained at least 0.05 fmol receptor/ $\mu$ g DNA or at least 10 fmol receptor/mg protein. We excluded women if they did not provide written consent, had missing tumor size, missing lymph node status, had a tumor diameter larger than 80 mm or smaller than 1 mm or had more than 30 affected lymph nodes at diagnosis. The total number of women eligible for analysis based on these criteria was 2874, with 1928 having data on grade, 2082 on ER status and 2039 on PR status.

PRSs for a selection of women in CAHRES were available through an extension of the original study.<sup>14</sup> In this extension, 1500 women were randomly selected, together with all women who had taken hormone replacement therapy (191 cases) and all women with self-reported diabetes mellitus (110 cases). These women were contacted by mail and those who consented were given blood sampling kits to be used at their primary health care facility. From all deceased breast cancer cases, attempts were made to retrieve archived tissue samples. Blood samples were collected from 1322 cases and archived tissue was collected for 247 cases (85% of all selected). DNA was isolated from 3 mL of whole blood and from non-malignant cells in the paraffin-embedded tissue samples. DNA samples were genotyped on a custom Illumina iSelect genotyping array (iCOGS).<sup>15</sup>

ST01-08 consists of all women diagnosed with invasive breast cancer in Stockholm from 2001 to 2008. Women were identified through the Stockholm-Gotland Regional Breast cancer quality register, and information was collected on tumor size, lymph node involvement, grade, ER status and PR status.<sup>16</sup> Tumor size, lymph node and grade were categorized in the same way as in CAHRES. ER and PR status were determined using radioimmunoassay or immunohistochemistry (IHC) with cutoff values of more than 10% positive cells for IHC and more than 0 fmol/ $\mu\text{g}$  DNA for radioimmunoassay assays, and categorized as negative or positive. We excluded women if they had missing tumor size, missing lymph node status, a tumor diameter larger than 80 mm or smaller than 1 mm, or if they had more than 30 affected lymph nodes at diagnosis. This left a total of 8076 women eligible for analysis. Less than 2% of patients had missing data on tumor size and lymph node involvement. Twenty percent of patients had missing data for ER and PR status. Grade was included in the register from 2004, with 7% of patients having missing data. After exclusions, the final numbers of available women with data on grade, ER status and PR status were 5227, 6518 and 6385, respectively.

All women in the Stockholm-Gotland Regional Breast Cancer quality register still alive in 2009, diagnosed with invasive breast cancer between 2001 and 2008, and younger than age 80 at diagnosis were invited to participate in a study named Libro-1. Invitations were mailed out in 2009, and 62% ( $n = 5715$ ) consented to take part in the study. These women gave blood specimens for genetic analysis. Of these, 5125 were successfully genotyped in a large-scale genotyping study on breast cancer risk.<sup>17</sup> Five thousand one hundred and twenty-two had enough remaining DNA for mutation testing using targeted sequencing. For the women in the Libro-1 study, data on molecular markers were retrieved in 2015 and 2016, from medical and pathology records at treating hospitals. From these, molecular subtype was assigned based on age at diagnosis, ER, PR, HER2 and Ki67 status using a random forest algorithm.<sup>18</sup> After applying the exclusion criteria, we were left with 1749 patients with data on molecular subtype. Our study was carried out with informed consent and ethical approvals from the Swedish ethical review board.

## 2.2 | Polygenic risk score

We constructed a PRS based on 158 single-nucleotide polymorphisms (SNPs) that were genotyped, or that could be imputed based on neighboring SNPs, in both studies. SNPs were chosen based on published studies on breast cancer risk. The PRS was constructed by summing the number of alleles of each SNP, weighted by per-allele odds ratios for breast cancer. Per allele odds ratios were taken from published studies, for example, Michailidou et al.<sup>19</sup> A PRS was thus calculated for each individual using the formula

$$\text{PRS} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n, \quad (1)$$

Where  $\beta_k$  is the per-allele log odds ratio for breast cancer associated with the minor allele for SNP  $k$ ,  $x_k = 0, 1$  or  $2$  is the number of minor

alleles for the same SNP, and  $n = 158$  is the total number of SNPs. After exclusions based on tumor and lymph node data, the PRS could be calculated for 1119 of the available cases in CAHRES, and 4150 of the available cases in Libro-1/ST01-08.

## 2.3 | Statistical models

To model the effect of breast cancer characteristics on rate of lymph node spread, we use a continuous tumor growth model, which includes a sub-model for lymph node spread. Under our modeling assumptions, the number of affected lymph nodes can be expressed as a direct function of current tumor characteristics. The approach was developed by Isheden et al.<sup>8</sup> and was recently extended to include covariate effects in the rate of lymph node spread.<sup>10</sup> A detailed description of the modeling approach is given in the Appendix S1. In short, the model assumes that tumor growth follows an exponential function with gamma distributed inverse growth rates, that time to symptomatic detection follows a failure time model with rate proportional to the current tumor volume and that the rate of breast cancer lymph node spread follows an inhomogeneous Poisson process with intensity function proportional to the growth rate of the tumor and the fourth power of number of times the cells in the tumor has divided. In summary, we assume that tumor cells spread to the lymph nodes as the primary tumor grows, according to an inhomogeneous Poisson process with intensity function given by

$$\lambda(t, r, s^*) = s^* D(t, r)^4 D'(t, r), \quad (2)$$

Where  $s^*$  is a gamma distributed random effect,  $D(t, r)$  is the number of times the cells in the tumor has divided and  $D'(t, r)$  is the rate of cell division in the tumor—both at time  $t$ , assuming an inverse growth rate  $r$ . In Isheden et al.<sup>8</sup> it was shown that these models lead to there being, at any time point, a negative binomial distribution, for the number of affected lymph nodes  $N$ , such that the probability of  $n$  affected lymph nodes, conditional on current tumor volume  $V$ , follows the functional form

$$P(N = n | V) = \frac{\Gamma(\gamma_1 + n) \gamma_2^{\gamma_1} \left( \left( \log \frac{V}{V_0} \right)^5 \right)^n}{\Gamma(\gamma_1 + n) n! \left( \left( \log \frac{V}{V_0} \right)^5 + \gamma_2 \right)^{\gamma_1 + n}}, \quad (3)$$

where  $V_0$  is the minimal volume of a detectible lymph node metastasis (here assumed to be 0.5 mm),  $\Gamma(\cdot)$  represents the gamma function, and  $\gamma_1$  and  $\gamma_2$  are the parameters of the gamma distributed random effect. The authors further showed<sup>10</sup> that the association between a covariate  $X$  and breast cancer lymph node spread can be modeled by assuming that the rate of lymph node spread in the underlying dynamic model of spread, during the pre-clinical phase, is amplified (or decreased) by a factor  $e^{\beta X}$

$$\lambda(t, r, s^*, X) = s^* e^{\beta X} D(t, r)^4 D'(t, r), \quad (4)$$

where  $\beta = (\beta_1, \beta_2, \dots, \beta_n)$  is a vector with the values of the covariate effects. When combined with the other assumed models, this leads to a negative binomial distribution, with number of affected lymph nodes  $N = n$ , given tumor volume  $V = v$  and covariate  $X$ , following

$$P(N = n|V) = \frac{\Gamma(\gamma_1 + n) \gamma_2^{\gamma_1} \left( e^{\beta X} \left( \log \frac{V}{V_0} \right)^5 \right)^n}{\Gamma(\gamma_1 + n) n! \left( e^{\beta X} \left( \log \frac{V}{V_0} \right)^5 + \gamma_2 \right)^{\gamma_1 + n}}, \quad (5)$$

Lymph node spread at diagnosis can also be affected by the tumor growth rate of the tumor. A faster growing tumor will result in a larger tumor volume at diagnosis, consequently leading to more lymph node spread at diagnosis. This can be accounted for by making a regression of tumor characteristics on the growth rate of the tumor, as was done in Isheden et al.<sup>10</sup> However, this requires screening data that we do not have for all of our study population. In our study, we therefore focus on the contribution to the rate of breast cancer lymph node spread. In the following sections, we use this likelihood to make inference on the effects of tumor characteristics on rates of breast cancer lymph node spread using data on tumor characteristics, tumor volume and number of lymph node metastases at diagnosis.

In addition to describing our approach in the Appendix S1, we also include a table (Table 1), summarizing the key characteristics/assumptions of our approach and the data used to make inference.

## 2.4 | Point estimates and confidence intervals

Point estimates are calculated using maximum likelihood estimation, where the likelihood is based on the probability of  $N = n$  affected lymph nodes, conditional on the tumor volume  $V = v$  and the covariate of interest  $X = x$ :  $p_n = P(N = n|V = v, X = x)$ .  $p_n$  is calculated using Equation (5). 95% confidence intervals are estimated from 2000 bootstrap replicates using the percentile method.

We model the effect of grade on rate of lymph node spread in two different ways: firstly by modeling the effect of grade as an ordinal variable, so that the rate of lymph node spread is amplified by the factor  $e^{\beta g}$ , where  $\beta$  is the log effect and  $g$  is the grade; and secondly, as a discrete variable, so that the rate of lymph node spread is amplified by the factor  $e^{\beta_2 g_2 + \beta_3 g_3}$ , where Grade 1 is the reference,  $\beta_2$  and  $\beta_3$  correspond to the log effects of Grades 2 and 3, respectively, and  $g_2, g_3$  are grade indicator variables. We model the effect of the PRS as a continuous variable, so that the rate of lymph node spread is amplified by a factor  $e^{\beta \cdot \text{PRS}}$ , where  $\beta$  is the log effect and PRS is the polygenic risk score. We model the effect of molecular subtype as a discrete variable  $e^{\beta_2 \text{LumB} + \beta_3 \text{HER2} + \beta_4 \text{Basal}}$ , where luminal A is the reference,  $\beta_2, \beta_3, \beta_4$  correspond to the log effects, and LumB, HER2, Basal are indicator variables. The effects of the remaining tumor characteristics are modeled with the amplification factor  $e^{\beta X}$ , where  $\beta$  is the log effect and  $X$  is the indicator variable of interest.

**TABLE 1** A summary of the model components, assumptions and data used

Submodel/assumptions	Data/comments
<i>Tumor growth:</i> Primary tumors are assumed to grow exponentially, with variability in growth rates across tumors, accounted for using a random effect	Under the modeling assumptions, size of the primary tumor (at diagnosis) does not need to be modeled in order to obtain estimates of lymph node spread during the preclinical phase of the primary tumor
<i>Seeding and detection of lymph node metastases:</i> During a tumor's preclinical phase, lymph node metastatic seeding occurs as a non-homogeneous Poisson process with rate proportional to the (unobserved) number of cell divisions, approximated as a function of tumor size and rate of growth of the primary tumor (additional variability in rates is accounted for, using a random effect). Lymph node metastases grow at the rate of growth of the primary tumor and are detectable once they reach a fixed size	Under the modeling assumptions, number of affected lymph nodes at diagnosis of the primary tumor is a direct function of tumor size (independent of growth rate of the primary tumor) and can therefore be modeled as a function of tumor size at diagnosis, from which rate of the underlying distribution of rates of spread to the lymph nodes during the preclinical phase of the primary tumor can be directly estimated. As well as being modeled with inter-patient variability (random effect), systematic variation in the rate of spread is allowed for as a function of covariates/tumor characteristics (eg, grade, PRS); Equation (5)
<i>Detection of the primary tumor:</i> No assumptions made	-

The rate ratio, that is, the ratio of the rate of lymph node spread (at all points in time during the cancer's preclinical phase) between two different tumors with different covariate levels  $X = x_1$  and  $X = x_2$ , assuming the same tumor volume  $V$ , inverse growth rate  $R$  and spread parameter  $s^*$ , is calculated as

$$\text{RR} = \frac{\lambda(t, r, s^*, x_1)}{\lambda(t, r, s^*, x_2)} = \frac{s^* e^{\beta x_1} D(t, r)^4 D'(t, r)}{s^* e^{\beta x_2} D(t, r)^4 D'(t, r)} = e^{\beta(x_1 - x_2)}, \quad (6)$$

$e^{\beta}$  can be interpreted as the rate ratio when we compare two tumors at covariate levels  $x_1 = 0$  and  $x_2 = 1$ . We calculate  $P$ -values using the log likelihood test statistic from a reduced data set where outliers have been removed. Simulations performed by the authors have shown that the likelihood ratio test gives over-inflation of low  $P$ -values when they are estimated based on the full data set. This is caused by the existence of outliers, which makes asymptotic convergence slow. To remedy this, we removed outliers for the  $P$ -value calculations. Removal of outliers was done by estimating the model without any covariates on the full data set, and then removing the 1-percentile of the data with smallest log-likelihood values. In the combined data set, this corresponds to removing data points with a log likelihood value smaller than  $-6.2$ .

### 3 | RESULTS

Table 2 shows descriptive data for CAHRES, ST01-08 and the combined data set, indicating the number of women with data on tumor size, lymph node metastases, grade, ER status, PR status, molecular subtype and PRS, together with the observed frequency and distribution of each tumor characteristic and genetic variable.

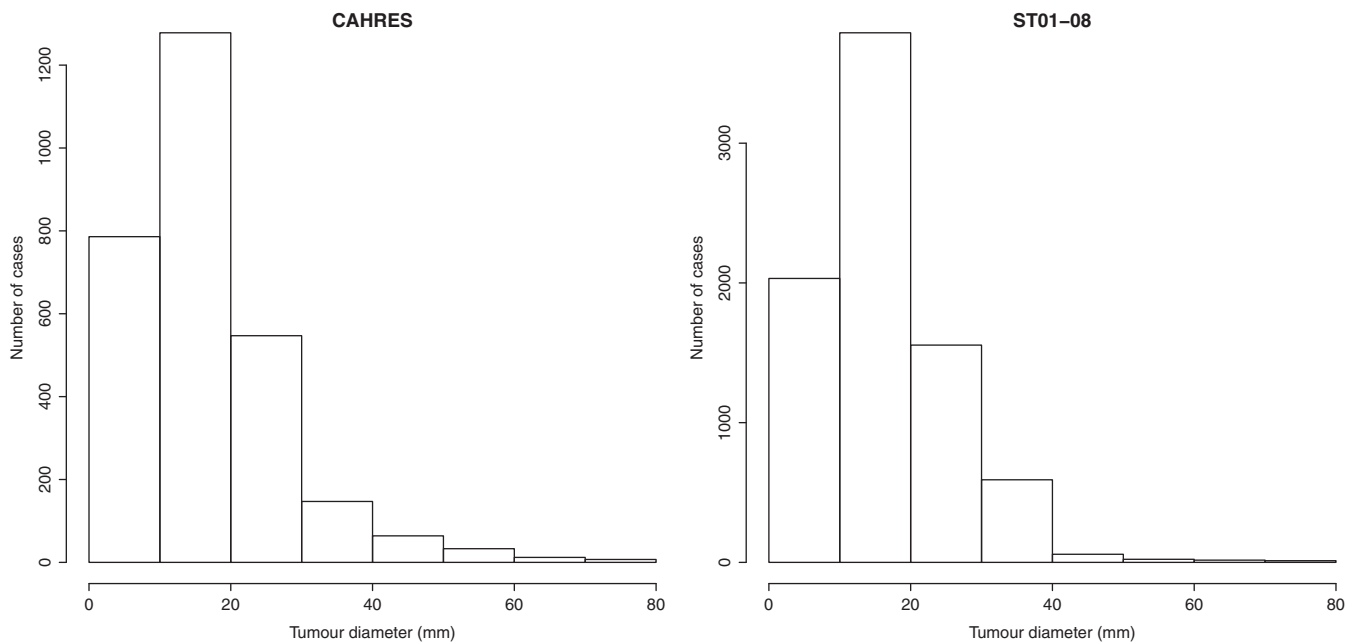
#### 3.1 | Tumor size, number of affected lymph nodes and lymph node positivity

We first calculated the fractions of patients with tumor diameters less than 10, 10 to 19, 20 to 29, and more than 30 mm. In the CAHRES data set, these fractions were 19%, 45%, 22% and 14%, and in the ST01-08 data set, they were 18%, 46%, 23% and 13%. The percentage (unit) difference between the two data sets is less than 1% for all four size

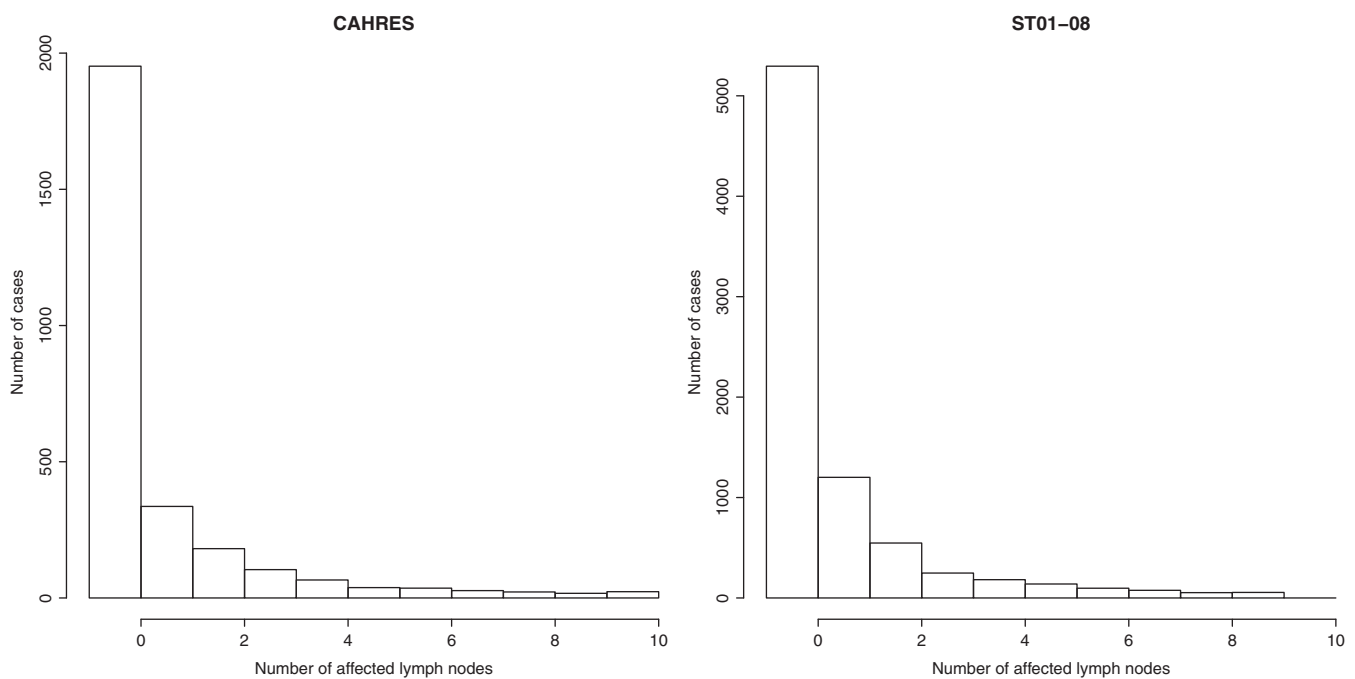
categories. In Figure 1, we show histograms of tumor diameters, divided into 10 mm intervals, for CAHRES and ST01-08. The fractions of patients with no affected lymph nodes, one affected lymph node, two affected lymph nodes and three or more affected lymph nodes were 68%, 12%, 6% and 14% in the CAHRES data set and 65%, 15%, 7% and 13% in the ST01-08 data set. In Figure 2, we show histograms of number of lymph nodes affected, from 0 to 10, for CAHRES and ST01-08. We next examined the proportion of patients with lymph node positive breast cancer. This was done for tumor size intervals 1 to 10, 11 to 20, 21 to 30, up to 71 to 80 mm for the CAHRES data, for ST01-08 and for the combined data. In Figure 3, these proportions are plotted for each data set as circles, with bootstrapped 95% confidence intervals intersecting each circle. Sopik and Norad<sup>20</sup> recently presented the distribution of number of affected lymph nodes using a large number of patients included in the Surveillance, Epidemiology and End Results (SEER) program database. We note that the pattern of association observed in that large study is very similar to that displayed in Figure 3.

Characteristic	CAHRES	ST01-08	Combined
Tumor size	2874	8076	10 950
Up to 9 mm	536 (19%)	1431 (18%)	1967 (18%)
10 to 19 mm	1303 (45%)	3706 (46%)	5009 (46%)
20 to 29 mm	644 (22%)	1899 (23%)	2543 (23%)
30 mm or more	391 (14%)	1040 (13%)	1431 (13%)
Number of affected lymph nodes	2874	8076	10 950
No affected lymph nodes	1952 (68%)	5295 (65%)	7247 (66%)
1 affected lymph node	334 (12%)	1201 (15%)	1535 (14%)
2 affected lymph nodes	181 (6%)	547 (6%)	728 (7%)
3 or more affected lymph nodes	405 (14%)	1099 (13%)	1504 (14%)
Grade	1928	5227	7155
Grade 1	299 (15%)	982 (19%)	1281 (18%)
Grade 2	805 (42%)	2662 (51%)	3467 (48%)
Grade 3	824 (43%)	1583 (30%)	2407 (34%)
Estrogen receptor status	2082	6518	8600
ER+	1628 (78%)	5532 (85%)	7160 (83%)
ER–	454 (22%)	986 (15%)	1440 (17%)
Progesteron receptor status	2039	6385	8424
PR+	1393 (68%)	4370 (68%)	5763 (68%)
PR–	646 (32%)	2015 (32%)	2661 (32%)
Molecular subtype	–	1749	–
Luminal A	–	1253 (72%)	–
Luminal B	–	174 (10%)	–
HER2-enriched	–	207 (12%)	–
Basal-like	–	115 (6%)	–
Polygenic risk score	1119	4150	5269
Lower quartile	129.1	129.1	129.0
Median	134.4	134.2	134.1
Upper quartile	138.8	139.2	139.1

**TABLE 2** Number of patients and descriptive statistics of tumor characteristics and genetic variables in CAHRES, ST01-08 and the combined data



**FIGURE 1** Histograms of tumor diameters divided into 10 mm intervals for CAHRES (left) and ST01-08 (right)



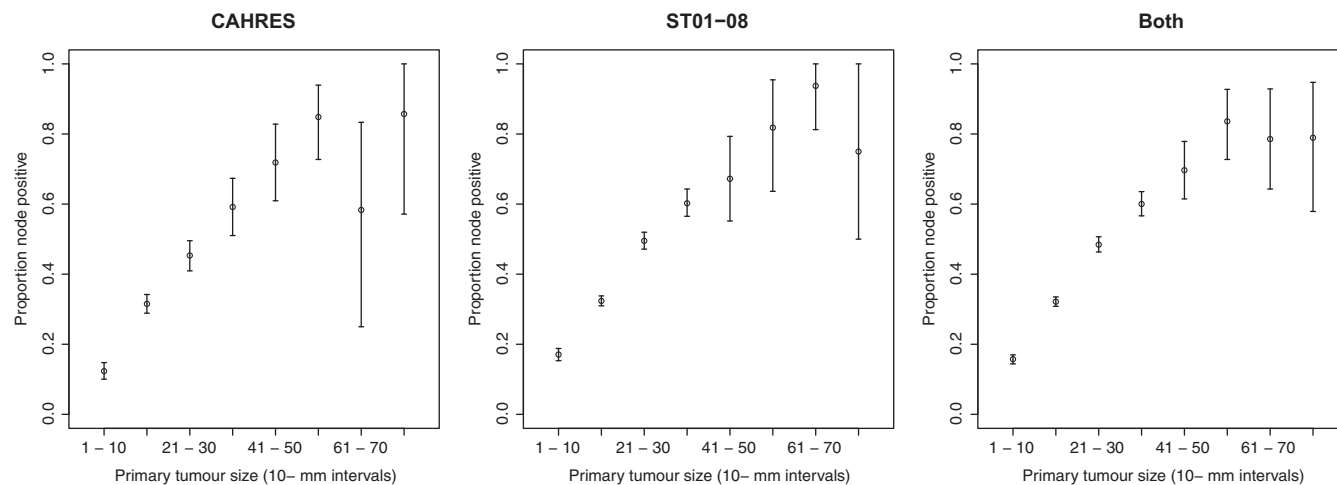
**FIGURE 2** Histograms of number of lymph nodes affected, for CAHRES (left) and ST01-08 (right)

### 3.2 | Association between lymph node spread and grade, hormone receptor status, molecular subtypes and PRS

In Table 3, we present point estimates and 95% confidence intervals from our analyzes of associations between rate of lymph node spread and grade, hormone receptor status, molecular subtype and PRS. *P*-values based on the data sets with outliers removed are presented in Table 4.

Modeling the association between lymph node spread and grade on a continuous scale, we estimated rate ratios, with corresponding 95% confidence intervals, when comparing grade 0/1 tumor to grade 1/2 tumors, to be 1.68 (1.41, 2.02), 1.44 (1.23, 1.66) and 1.51 (1.34, 1.69) based on CAHRES, ST01-08 and the combined data, respectively. When outliers were removed, the estimated rate ratios and corresponding 95% confidence intervals were 1.61 (1.41, 1.87), 1.36 (1.25, 1.47) and 1.43 (1.34, 1.54). The corresponding *P*-values were all smaller than  $10^{-10}$ . Modeling the association with the discrete model,





**FIGURE 3** Proportion of lymph node positive cases in different tumor size categories (circles) alongside bootstrapped 95% confidence intervals (lines) for CAHRES, ST01-08 and the combined data

using Grade 1 as reference, rate ratio and 95% confidence intervals for Grade 2 tumors were 1.68 (1.11, 2.59), 1.56 (1.07, 2.11) and 1.59 (1.20, 2.06), and the corresponding estimates for Grade 3 tumors were 2.83 (1.89, 4.42), 2.15 (1.47, 2.94) and 2.32 (1.73, 2.99). These estimates were similar when outliers were removed, and the corresponding  $P$ -values were all smaller than  $10^{-9}$ . All analyses of grade were consistent in that increasing grade implied increasing rate of lymph node spread.

Compared to ER negative breast cancer, ER positive breast cancer was associated with rate ratios and corresponding 95% confidence intervals of 0.60 (0.40, 0.84), 0.61 (0.44, 0.81) and 0.61 (0.47, 0.76), based on CAHRES, ST01-08 and the combined data, respectively. When outliers were removed, the corresponding estimates were 0.79 (0.64, 0.96), 0.82 (0.72, 0.95) and 0.80 (0.72, 0.90), and the estimated  $P$ -values were  $2.2 \times 10^{-2}$ ,  $3.8 \times 10^{-3}$  and  $1.1 \times 10^{-4}$ . Similarly for PR positive breast cancer, rate ratios and corresponding 95% confidence intervals were estimated as 0.71 (0.53, 0.91), 0.63 (0.48, 0.77) and 0.65 (0.52, 0.78), based on CAHRES, ST01-08 and the combined data, respectively. Corresponding estimates in the data set with outliers removed were 0.79 (0.66, 0.94), 0.86 (0.77, 0.95) and 0.84 (0.76, 0.92). The  $P$ -values were estimated as  $1.1 \times 10^{-2}$ ,  $3.8 \times 10^{-3}$  and  $1.2 \times 10^{-4}$ . For both the ER positive and ER negative breast cancer estimates, rate ratios were consistently estimated as negative and statistically significant at  $\alpha = .05$ . In the analysis of molecular subtypes, HER2-enriched breast cancer was associated with a rate ratio and corresponding 95% confidence interval of 1.83 (1.05, 4.18), compared to luminal A breast cancer. When removing outliers, the corresponding estimate was 1.53 (1.15, 1.99). The  $P$ -value for association between molecular subtype and rate of lymph node spread, based on a test with 3 degrees of freedom, was  $7.2 \times 10^{-4}$ .

We note that, for both studies, information on grade, ER status and PR status was not as complete as it was for tumor size and lymph node status. While distributions of the latter two characteristics were similar for the two studies, the proportions of high grade and ER–

tumors differed. All statistical analyses based on grade and ER status, conditions on these characteristics (grade and ER status are included as covariates), therefore non-random missingness on these characteristics will not introduce bias in the analyses presented here.

We did not find any convincing evidence that the PRS is associated with the rate of lymph node spread.

As an illustration of our results, we display graphically the observed and estimated model-based relationship between lymph node spread and grade. Under our modeling assumptions, the number of affected lymph nodes can be expressed as a direct function of current tumor characteristics, see Equation (5). In Figure 4, we plot the expected number of affected lymph nodes as a continuous function of tumor volume for the different grades, under the model where grade was treated as an ordinal variable (using  $\beta = .41$ ). We note that the model treating grade on the ordinal scale gave a very similar fit to the data as the model treating grade as a discrete covariate ( $P$ -value = .45, testing for a difference using a likelihood ratio test).

## 4 | DISCUSSION

In this article, we have investigated the association between rate of lymph node spread and tumor characteristics, molecular subtype and genetic factors, using data from two large and independent observational studies comprising a total of 10 950 women. The data sets were largely in concordance in terms of tumor characteristics: tumor size distributions differed by at most 1% for the four considered tumor size categories, lymph node spread distributions differed by at most 3%, the percentage PR-positive breast cancers was the same and distributions of PRS were very similar. They differed most in terms of grade, with a 13% difference in number of Grade 3 cases, and ER status, with a 7% difference in number of ER-positive cases. Both cohorts were analyzed on a stand-alone basis, and as one big data set, using a novel continuous growth model that adjusts for tumor volume

**TABLE 3** Estimates and confidence intervals of breast lymph node spread (rate ratios) for different tumor characteristics

Characteristic	CAHRES	ST01-08	Combined
<i>Outliers included</i>			
Grade (continuous)	1.68 (1.41, 2.02)	1.44 (1.23, 1.66)	1.51 (1.34, 1.69)
Grade 1	Ref	Ref	Ref
Grade 2	1.68 (1.11, 2.59)	1.56 (1.07, 2.11)	1.59 (1.20, 2.06)
Grade 3	2.83 (1.89, 4.42)	2.15 (1.47, 2.94)	2.32 (1.73, 2.99)
Estrogen receptor status			
ER+	0.60 (0.40, 0.84)	0.61 (0.44, 0.81)	0.61 (0.47, 0.76)
ER–	Ref	Ref	Ref
Progesteron receptor status			
PR+	0.71 (0.53, 0.91)	0.63 (0.48, 0.77)	0.65 (0.52, 0.78)
PR–	Ref	Ref	Ref
Molecular subtype			
Luminal A	–	Ref	–
Luminal B	–	1.34 (0.77, 1.74)	–
HER2-enriched	–	1.83 (1.05, 4.18)	–
Basal-like	–	0.70 (0.35, 1.18)	–
Polygenic risk score <sup>a</sup>	0.99 (0.89, 1.11)	0.95 (0.87, 1.02)	0.96 (0.89, 1.01)
<i>Outliers removed</i>			
Grade (continuous)	1.61 (1.41, 1.87)	1.36 (1.25, 1.47)	1.43 (1.34, 1.54)
Grade 1	Ref	Ref	Ref
Grade 2	1.76 (1.25, 2.59)	1.59 (1.33, 1.91)	1.63 (1.39, 1.91)
Grade 3	2.73 (2.00, 3.89)	1.96 (1.64, 2.35)	2.17 (1.86, 2.57)
Estrogen receptor status			
ER+	0.79 (0.64, 0.96)	0.82 (0.72, 0.95)	0.80 (0.72, 0.90)
ER–	Ref	Ref	Ref
Progesteron receptor status			
PR+	0.79 (0.66, 0.94)	0.86 (0.77, 0.95)	0.84 (0.76, 0.92)
PR–	Ref	Ref	Ref
Molecular subtype			
Luminal A	–	Ref	–
Luminal B	–	1.31 (0.97, 1.67)	–
HER2-enriched	–	1.53 (1.15, 1.99)	–
Basal-like	–	0.70 (0.42, 1.04)	–
Polygenic risk score <sup>a</sup>	0.98 (0.89, 1.09)	1.01 (0.97, 1.05)	1.00 (0.96, 1.05)

<sup>a</sup>Median polygenic risk score compared to lower quartile.

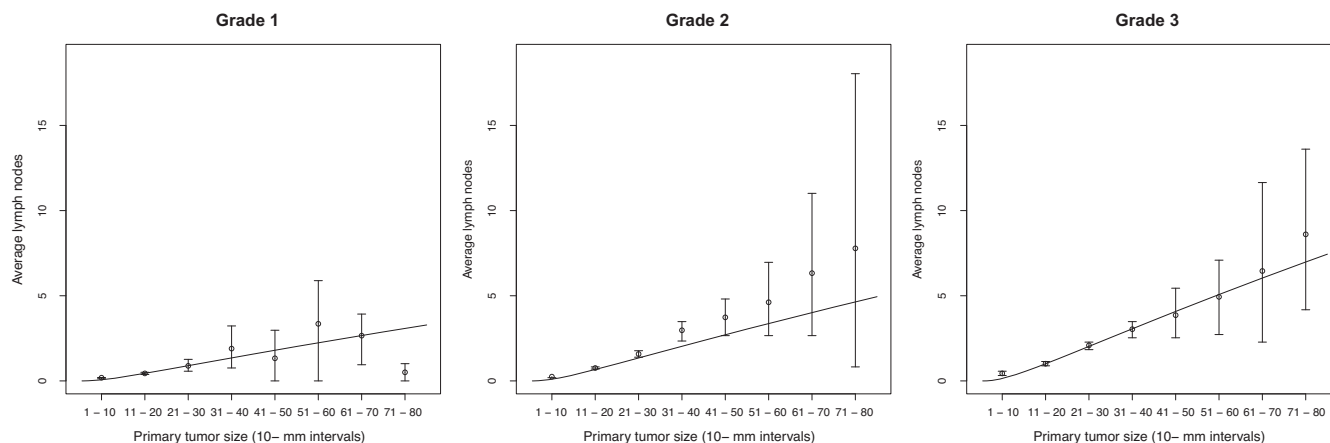
**TABLE 4** P-values of association with rate of lymph node spread, calculated based on CAHRES, ST01-08 and the combined data

Characteristic	CAHRES	ST01-08	Combined
Grade (continuous)	$2.3 \times 10^{-11}$	$8.0 \times 10^{-13}$	$p < 1 \times 10^{-16}$
Grade (discrete)	$2.0 \times 10^{-10}$	$7.6 \times 10^{-13}$	$p < 1 \times 10^{-16}$
Estrogen receptor status	$2.2 \times 10^{-2}$	$3.8 \times 10^{-3}$	$1.1 \times 10^{-4}$
Progesteron receptor status	$1.1 \times 10^{-2}$	$3.8 \times 10^{-3}$	$1.2 \times 10^{-4}$
Molecular subtype	–	$7.2 \times 10^{-4}$	–
Polygenic risk score	$6.9 \times 10^{-1}$	$6.9 \times 10^{-1}$	$8.6 \times 10^{-1}$

in a biologically motivated way. The model assumes that the rate of lymph node spread follows a continuous time Poisson process, and that the association with tumor characteristics is through the rate

coefficient. Though our model fits data better than previously suggested lymph node spread models from the literature,<sup>7-9</sup> as well as Poisson and negative binomial regression techniques,<sup>10</sup> our model is





**FIGURE 4** Model-based estimates of expected number of lymph nodes affected as a continuous function of tumor volume for Grade 1, Grade 2 and Grade 3 tumors ( $\beta = .41$  from the ordinal model for grade), alongside observed means (circles) and bootstrapped 95% confidence intervals (bars intersecting circles) from the combined data set

not likely to fully capture the biological intricacies of the tumor spread process. For example, we make the simplifying assumption that tumor characteristics do not change over the tumor growth process. If characteristics do “switch” over time, we would expect that the estimates for the more benign characteristics are less likely to be affected by “switching,” than the more malignant tumor characteristics (such as Grade 3), since in these cases, tumors would have mutated from a less spreading state to a more spreading state, causing an underestimation of rate of lymph node spread. In theory, switching could be accommodated in our (continuous growth) model, but this would involve a non-trivial extension. In simpler multi-state models, this has been done in the breast cancer screening data modeling literature for grade,<sup>21</sup> based on a mover-stayer model, which is an extension of the Markov chain, where the population of tumors is assumed to consist of two unobserved groups, a stayer group consisting of tumors with a zero probability of change and a mover group following an ordinary Markov process.

The strongest evidence of association was found between grade and rate of lymph node spread. Higher grade breast cancer is generally less differentiated, more invasive and more proliferative. We therefore expected higher rates of lymph node spread for higher grade tumors. In our data, Grade 1 tumor had least lymph node spread across all tumor sizes; see Figure 4. When modeling the association between grade and rate of lymph node spread, both the linear and discrete relationships were highly significant. All  $P$ -values were less than  $5 \times 10^{-10}$ , with the average rate ratio between a grade  $x$  tumor and a grade  $x + 1$  tumor being 1.49. Our estimates are consistent with Nouh et al and Gann et al<sup>22,23</sup> who found statistically significant relationships between lymph node positivity and grade.

ER and PR positive breast cancers were associated with a reduced rate of lymph node spread compared to ER and PR negative breast cancers. Molecular subtype was significantly associated with rate of lymph node spread. In our model, luminal A was the reference subtype. This subtype is generally ER positive and of low grade.

Compared to luminal A, luminal B generally has a higher grade at diagnosis. In our study, luminal B was associated with a higher rate of lymph node spread compared to luminal A breast cancer. The HER2-enriched subtype was associated with an even higher rate of lymph node spread than luminal B. Basal-like breast cancer had the lowest rate of breast cancer lymph node spread. Basal-like breast cancer is often triple-negative. These cancers are more likely to forego spreading to the lymph nodes<sup>20</sup> and instead form distant metastases. The associations that we found for the molecular subtypes are consistent with the results of Liu et al,<sup>24</sup> who investigated lymph node positivity for luminal A, luminal B, HER2-positive and triple-negative breast cancer.

While Figure 3 partly captures the clinical significance of the rate ratio estimates (for grade), using our model, with sufficient information, it would also be possible to describe clinical significance of the rate ratio estimates in other, perhaps even more meaningful, ways. Evaluating time to LN spread would, though, need to incorporate information on the relationship between growth rate of the primary tumor and the tumor characteristic (or eg, PRS). We do not have such information. However, if we did, then we could calculate, for example, expected times to first affected lymph node, for each tumor characteristic, and we could even study the impact of delayed detection for tumors with different characteristics. For a description of how the latter can be done, see Isheden et al.<sup>10</sup>

Our study of genetic factors and rates of lymph node spread may suffer from survivorship bias. In Libro-1, blood samples were collected after 2008, and in CAHRES, they were collected after 1997. Some women had already died before blood samples were collected, which means that the association between genetic factors and rate of lymph node spread in our study may be underestimated. In any case, we found no significant association between our PRS and rate of lymph node spread. Using Libro-1 data, Li et al<sup>25</sup> did not find a significant relationship between PRS and lymph node status, or between PRS and survival. Furthermore, we are not aware of any study that has

found a significant association between PRS and lymph node status at diagnosis.

## 5 | CONCLUSIONS

Survival benefits of screening and treatment vary across different breast cancer molecular subtypes.<sup>1</sup> Some cancers are more aggressive and some cancers are less aggressive. In part, this is reflected in a tumors propensity to spread to the lymph nodes, which in many cases is a precursory step of distant metastatic spread. In our current study, we have quantified tumor aggressiveness in terms of rate of lymph node spread based on genetic markers, tumor characteristics and for different molecular subtypes. Quantifying tumor aggressivity may prove useful in the future era of individualized treatment and screening.

### CONFLICT OF INTEREST

The authors declare no conflict of interest.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author on request, after ethical approvals have been obtained from the Swedish ethical review board.

### ORCID

Gabriel Isheden  <https://orcid.org/0000-0003-2536-2051>

### REFERENCES

- Plevritis SK, Munoz D, Kurian AW, et al. Association of screening and treatment with breast cancer mortality by molecular subtype in US women, 2000-2012. *J Am Med Assoc.* 2018;319(2):154-164.
- Rueda OM, Sammut SJ, Seoane JA, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. *Nature.* 2019;567(7748):399.
- Colzani E, Liljegren A, Johansson AL, et al. Prognosis of patients with breast cancer: causes of death and effects of time since diagnosis, age, and tumor characteristics. *J Clin Oncol.* 2011;29(30):4014-4021.
- Ullah I, Karthik GM, Alkods A, Kjällquist U, et al. Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. *J Clin Invest.* 2018;128(4):1355-1370.
- Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160.
- Berry DA, Cronin KA, Plevritis SK, et al. Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med.* 2005;353(17):1784-1792.
- Shwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. *Cancer.* 1978;41(4):1550-1564.
- Isheden G, Abrahamsson L, Andersson T, Czene K, Humphreys K. Joint models of tumor size and lymph node spread for incident breast cancer cases in the presence of screening. *Stat Methods Med Res.* 2019;28(12):3822-3842.
- Hanin L, Yakovlev A. Multivariate distributions of clinical covariates at the time of cancer detection. *Stat Methods Med Res.* 2004;13(6):457-489.
- Isheden G, Czene K, Humphreys K. Random effects models of lymph node metastases in breast cancer: quantifying the roles of covariates and screening using a continuous growth model. *Biometrics.* 2021. <https://doi.org/10.1111/biom.13430>.
- Rosenberg LU, Magnusson C, Lindström E, Wedrén S, Hall P, Dickman PW. Menopausal hormone therapy and other breast cancer risk factors in relation to the risk of different histological subtypes of breast cancer: a case-control study. *Breast Cancer Res.* 2006;8(1):R11.
- Rosenberg LU, Granath F, Dickman PW, et al. Menopausal hormone therapy in relation to breast cancer characteristics and prognosis: a cohort study. *Breast Cancer Res.* 2008;10(5):R78.
- Eriksson L, Czene K, Rosenberg L, Humphreys K, Hall P. The influence of mammographic density on breast tumor characteristics. *Breast Cancer Res Treat.* 2012;134(2):859-866.
- Wedrén S, Lovmar L, Humphreys K, et al. Oestrogen receptor  $\alpha$  gene haplotype and postmenopausal breast cancer risk: a case control study. *Breast Cancer Res.* 2004;6(4):R437.
- Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet.* 2015;47(4):373.
- Holm J, Humphreys K, Li J, et al. Risk factors and tumor characteristics of interval cancers by mammographic density. *J Clin Oncol.* 2015;33(9):1030-1037.
- Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.* 2013;45(4):353.
- Holm J, Eriksson L, Ploner A, et al. Assessment of breast cancer risk factors reveals subtype heterogeneity. *Cancer Res.* 2017;77(13):3708-3717.
- Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature.* 2017;551(7678):92.
- Sopik V, Narod SA. The relationship between tumor size, nodal status and distant metastases: on the origins of breast cancer. *Breast Cancer Res Treat.* 2018;170:647-656.
- Chen HH, Duffy SW, Tabar L. A mover-stayer mixture of Markov chain models for the assessment of dedifferentiation and tumor progression in breast cancer. *J Appl Stat.* 1997;24(3):265-278.
- Nouh MA, Ismail H, El-Din N, El-Bolkainy MN. Lymph node metastasis in breast carcinoma: clinico-pathological correlations in 3747 patients. *J Egypt Natl Canc Inst.* 2004;16(1):50-56.
- Gann PH, Colilla SA, Gapstur SM, Winchester DJ, Winchester DP. Factors associated with axillary lymph node metastasis from breast carcinoma: descriptive and predictive analyses. *Cancer.* 1999;86(8):1511-1519.
- Liu N, Yang Z, Liu X, Niu Y. Lymph node status in different molecular subtype of breast cancer: triple negative tumors are more likely lymph node negative. *Oncotarget.* 2017;8(33):55534.
- Li J, Ugalde-Morales E, Wen WX, et al. Differential burden of rare and common variants on tumor characteristics, survival, and mode of detection in breast cancer. *Cancer Res.* 2018;78(21):6329-6338.

### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Isheden G, Grassmann F, Czene K, Humphreys K. Lymph node metastases in breast cancer: Investigating associations with tumor characteristics, molecular subtypes and polygenic risk score using a continuous growth model. *Int. J. Cancer.* 2021;149(6):1348-1357. <https://doi.org/10.1002/ijc.33704>