Inferring rates of metastatic dissemination using stochastic network models

P. Gerlee[1], M. Johansson[2]

[1]*Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden and*

[2]*Department of Oncology, Institute of Clinical Sciences, Sahlgrenska Academy at the University of Gothenburg, Sahlgrenska University Hospital, Gothenburg, Sweden*

(Dated: June 20, 2018)

Abstract

The formation of metastases is driven by the ability of cancer cells to disseminate from the site of the primary tumour to target organs. The process of dissemination is constrained by anatomical features such as the flow of blood and lymph in the circulatory system. We exploit this fact in a stochastic network model of metastasis formation, in which only anatomically feasible routes of dissemination are considered. By fitting this model to two different clinical datasets (tongue & ovarian cancer) we show that incidence data can be modelled using a small number of biologically meaningful parameters. The fitted models reveal site specific rates of dissemination and also allow for patient-specific predictions of metastatic involvement based on primary tumour location and stage. Applied to other data sets this type of model could yield insight about seed-soil effect, and could also be used in a clinical setting to provide personalised predictions about the extent of metastatic spread.

INTRODUCTION

For most forms of cancer, occurrence of distant metastases equals incurable disease. Regional metastases, i.e. positive lymph nodes, also implies an inferior prognosis. In order to diminish the risk of dissemination of the primary tumor, patients commonly receive adjuvant treatment with radiotherapy and/or some kind of medical oncological treatment after surgery. Yet, in many cases, the patient is later on faced with residual or recurrent disease. On the other hand, many patients receive adjuvant treatment with subsequent side effects, even though their illness would never have disseminated. Hence, increased knowledge about the extent of metastasis at diagnosis would improve the care of many cancer patients.

Metastatic spread is known to follow certain disease specific patterns [1, 2], but despite the current state of knowledge, a quantitative understanding of the process of metastatic spread would improve our ability to optimise therapy and reduce over-treatment. Here will we make use of stochastic modelling to quantitatively assess the importance of different routes of dissemination. In addition, this method also makes it possible to estimate the extent of metastatic spread based on tumour stage and location. Our approach is versatile and we highlight this by applying it to two different types of tumours: carcinoma of the oral tongue, which disseminates through the lymphatic system and ovarian carcinoma, which spreads both intraperitoneally, lymphatically and through the blood circulatory system.

Most primary tumour will, given the time, eventually spread to virtually any organ in the body, but the process by which this occurs is a multi-step process. Typically tumour cells detach from the primary tumour, invade the surrounding stroma and find their way to local lymph nodes or blood vessels. Although a late stage tumour can release very large number of circulating tumour cells (CTCs) into the blood stream (up to 4×10^6 cells shed per gram of tumour per day [3]), the low probability of forming metastatic foci [4] coupled with the low probability of passing through capillary beds [5] leads to the conclusion that micrometastases in for example the liver (for gut malignancies) and the lung (for all malignancies) are necessary for further hematogenic dissemination [6–8]. These, often microscopic, lesions release CTCs into the arterial side of the circulation (in the case of the lung), and hence amplify CTC numbers in arterial blood, which otherwise would be low due to the filtration occurring in the lung capillary bed. A similar process is at work during lymphatic spread, where CTCs get trapped in lymph nodes, where they form micrometastases, which shed

CTCs that travel further downstream in the lymphatic system.

This process is known as secondary seeding, and although it is of paramount importance for the metasatic process we still have a limited understanding of the steps involved. Here mathematical modelling can provide a helping hand, since it allows for a quantitative understanding of biological processes that are difficult to measure directly.

Mathematical modelling of metastasis dates back to a series of seminal papers by Liotta and coworkers [9–11]. They considered the release of CTCs from an implanted tumour in mice and the subsequent formation of lung metastases. Using both deterministic and probabilistic methods they could derive predictions of how the number of metastatic foci changes over time and how the probability of being free of metastasis changed [10]. These predictions agreed well with experimental data and highlights the stochastic yet predictable nature of metastatic spread.

Another important contribution was made by Iwata et al. [12] who formulated and analysed a model which accounts for secondary seeding. That model predicts how the size distribution of metastases changes as the disease progresses. Predictions of the model were tested by Baratchart et al. [13] in a murine model of renal carcinoma, and while the model could predict the total metastatic burden, it was unable to describe the size distribution of metastases. However, the model prediction could be improved by assuming interactions between metastatic foci.

The importance of secondary seeding was investigated by Scott et al. [8] in the context of self-seeding, the process whereby a primary tumour can accelerate its own growth by releasing CTCs that return to the site of origin. A careful mathematical treatment of this hypothesis showed that secondary seeding is indeed required for this pathway to contribute to primary growth.

Secondary seeding also has an impact on estimates of metastatic efficiency, since undetected micrometastatic lesions render the apparent spread directly from the primary to target sites more efficient than it actually is [7].

The idea of metastatic spread occurring on a network, where the nodes represent organs and the links correspond to routes of spread was first described by Scott et al. [14], and later modelled by Newton et al. [15, 16]. They considered a stochastic model where the dissemination of cancer cells is modeled as an ensemble of random walkers on the network. The dynamics of the model is determined by a transition matrix, which was obtained by

fitting the model to a large autopsy data set [17]. The entries of this transition matrix gives information about rates of dissemination between different organs.

Here we build on the work of Newton et al., but with secondary seeding in mind, make modifications which alleviate the problem underdetermination, which plagued their work. In order to parametrise their model they had to assume that the observed patterns of metastasis correspond to a steady-state distribution of the model. We instead make use of primary tumour stage to create a model which is temporal and considerably easier to parametrise.

Focusing on two primary tumours we show that our model provides highly accurate estimates of dissemination rates, and crucially allows us to estimate rates between sites, which are inaccessible if one simply analyses incidence data.

RESULTS

Our aim is to quantify the rate of metastatic spreading by applying a stochastic network model to clinical data. We will consider two different data sets: (i) A cohort of 141 patients diagnosed with tongue cancer where metastases occur in the head and neck region and (ii) data from patients with ovarian cancer obtained from the SEER-database where metastases occur both regionally and to some extent in distant organs. Regional spread typically occurs via the lymphatic system and metastases appear in lymph nodes (LN), whereas distant metastases are mediated by the blood circulatory system and appear in other organs such as the liver or the lung. Although the two processes are different in some respects they also share many commonalities. Firstly, the spread of the disseminated tumour cells is constrained by anatomical structures, i.e. the lymphatic system and the blood circulatory system. Secondly, the spread is directed since it is subject to the flow in the system, and lastly, the formation of metastases at secondary sites affects the downstream rate of formation since metastases, like the primary tumour, disseminate tumour cells as they grow. These similarities makes it possible to formulate a general mathematical model, which can be tailored to describe both data sets.

The model consists of a network with sites and directed links [14]. The nodes correspond to lymph node stations or organs, and be in either of two states: negative (i.e. containing no metastases) or positive (containing one or more metastases). The state of each node is a

random variable, and the probability of it being positive depends on the state of the other nodes, and the rates of shedding between the nodes.

We assume that a positive site sheds CTCs that flow according to the links of the network. This applies both to the primary tumour and subsequent metastases. For simplicity we do not consider the size of metastases in this model, since this would require more parameters [18], and would make it more difficult to fit the model to clinical data. This implies that the state of each site X_i is a binary variable taking the value 0 if the site is empty and 1 if it contains a metastasis.

Each cancer cell which is disseminated from a site has the same probability of forming a metastasis at a downstream site. Since it is known that filtration rates are high (only 1 in 10 000 cells pass through a capillary bed) [8] we assume that CTCs only flow to neighbouring sites, e.g. CTCs released on the venous side of the circulatory system can only give rise to metastases in the lung, since this is the first capillary bed they encounter. This is also a good approximation for lymphatic spread, where the occurence of skip metastases, in which intermediary LNs are negative, is rare [19]. We aggregate the rate of release of CTCs, the survival probability in the circulatory system and the probability of forming a metastases in a downstream site into a single rate parameter, which we call λ, when flow is from the primary tumour, and ϕ, when flow is from metastatic sites. This parameter then corresponds to the rate at which a downstream site becomes positive given a positive site upstream.

We include a temporal dimension into the model by using the primary tumour stage as a proxy for time. This means that we consider the stage as informative about the total number of CTCs that have been disseminated from the initiation of the tumour up until diagnosis. This is of course a crude approximation since it is known that tumour progress at different rates and preferentially disseminate to different sites depending on for example which mutations it harbours [20]. Although tumours might progress at different rates they still have to pass all the intermediary stages, and very rarely regress to a lower stage. Since we do not have a mapping from real time to tumour stage, the rates that we infer from clinical data are only relative, but given a primary tumour of a certain stage the model can still predict the probability of metastases at different sites. The relative magnitude of the rates also inform us about the risk of developing new metastases given a certain metastatic burden.

We now move on to discuss the models specific to tongue cancer and ovarian cancer.

Lymphatic spread of tongue cancer

The drainage of the lymphatic system in the head and neck area can be described by a network where the nodes correspond to lymph node stations and the links represent flow between the stations. The location of the primary tumour determines how much it drains into the different stations, Tumours of the tongue mainly drain into station I and II. Station I drains into II, which in turn drains into III, which drains into IV. The dissemination network is shown in fig. 1, where the directed links show how CTCs flow in the system.

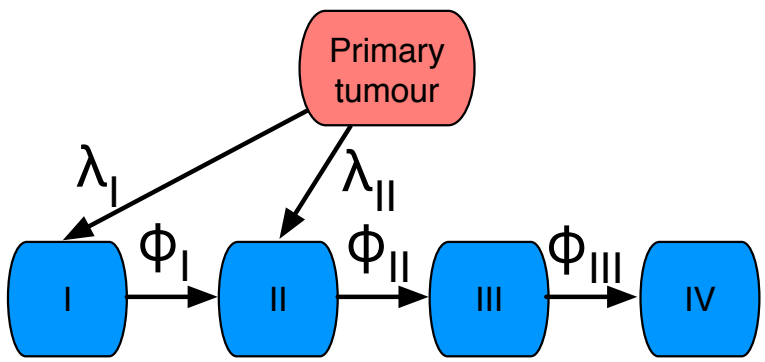

FIG. 1. Schematic of the flow of metastatic cells in the case of tongue cancer. The cancer cells flow from the primary tumour to lymph node station I and II with rates λ_I and λ_{II} respectively. The flow between lymph node stations is defined by the rates ϕ_I, ϕ_{II} and ϕ_{III}.

Given a primary tumour metastases will eventually form in downstream stations, but the dynamics of the model crucially depend on the parameters $(\lambda_I, \lambda_{II}, \phi_I, \phi_{II}, \phi_{III})$. The aim is therefore to estimate these from clinical data. Our data set contains 141 patients with carcinoma of the oral tongue diagnosed between 2004 and 2014 and treated at the department of Oncology at Sahlgrenska University Hospital in Gothenburg. For each patient we have information about the stage of the primary tumour and the presence/absence of metastasis in LN station I-IV.

First we transform the data so that the model can be fit to it. Let us first focus on station I and let N_t^I denote the number of patients with stage t primary tumours that are positive in station I. Here t can take the values $1, 2, 3, 4$ corresponding to primary tumour stage T1, T2, T3 and T4. Let M_t denote the total number of patients with stage t disease. The fraction of patients with stage t tumour with positive lymph node at station I is then

given by

$$p_t^I = \frac{N_t^I}{M_t}. \tag{1}$$

The same procedure is applied to stations II-IV yielding stage-dependent fractions.

The fraction p_t^i can be interpreted as the probability of finding a patient with stage t disease with a metastasis in LN station i. This quantity can readily be calculated from the model (see Methods for details). We now fit the model (i.e. find the rates $\lambda_I, \lambda_{II}, \phi_I, \phi_{II}, \phi_{III}$) using a maximum likelihood estimation of the rates in order to fit the model to the probabilities p_t^i for each lymph node station (see Method for details).

A comparison of the clinical data and the model fit is shown in figure 2, where each panel correspond to a LN station (I-IV). The estimated dissemination rates are shown in table I together with 95 % confidence intervals obtained using parametric bootstrap (see Methods). It is worth noting that the rate of dissemination from station I to II are close to zero, suggesting the flow of CTCs between these stations is negligible. However, the confidence intervals obtained from bootstrapping shows that the variability in flow from the primary to site II is large, ranging from practically zero to 1.22, the largest of all rates. By plotting the indirect rate (ϕ_I) against the direct rate (λ_{II}) for every iteration of the bootstrap procedure we observe an interesting pattern (see fig. 3). We see that we either obtain a model with large direct rate (and small indirect rate) or a model with large indirect rate (and small direct rate). Hence, with the current amount of the data we are unable to ascertain the importance of flow from station I to II.

TABLE I. The parameter estimates for lymphatic spread of tongue cancer.

Parameter	direct metastatic rate λ	95 % CI
λ_I	0.09	[0.06, 0.12]
λ_{II}	0.20	[0.06, 0.24
ϕ_I	5.7×10^{-10}	[3.6×10^{-11} , 1.22]
ϕ_{II}	0.30	[0.21, 0.43]
ϕ_{III}	0.13	[0.04, 0.24]

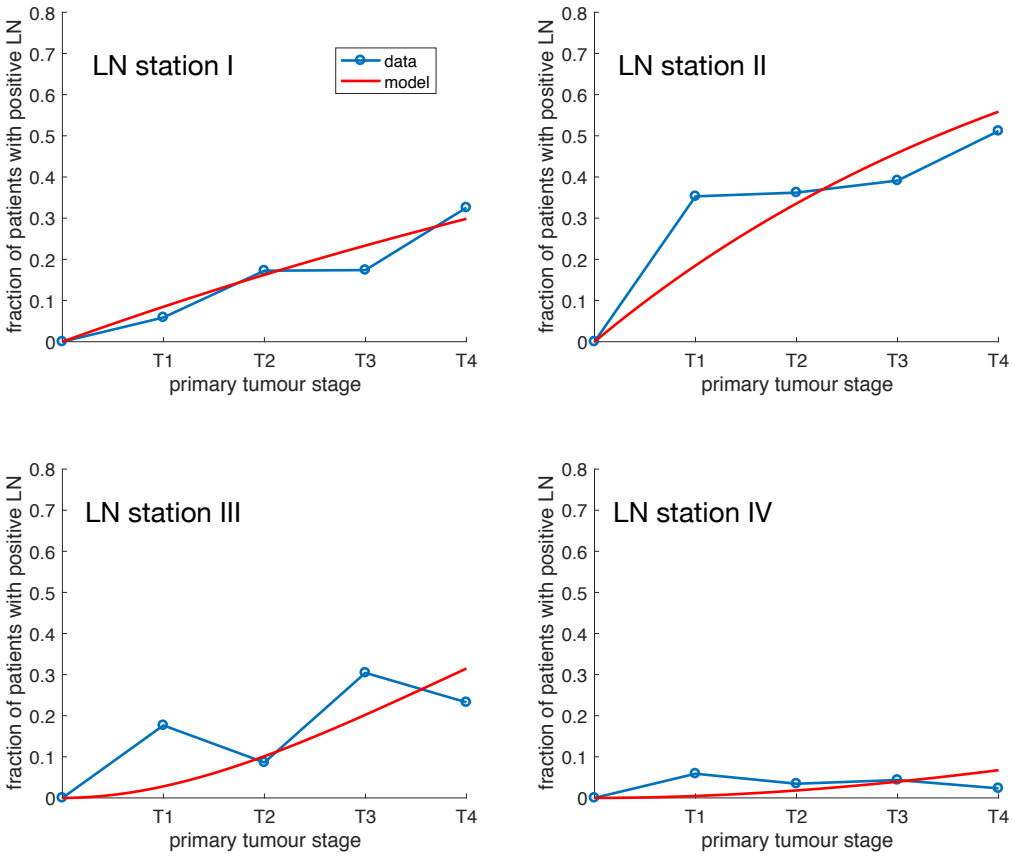

FIG. 2. The data and model predictions for metastatic spread to lymph node station I-IV for primary tongue cancer. Each panel corresponds to a lymph node station and shows the probability of finding a metastases in the lymph node as a function of the primary tumour stage. The parameters of the model are given in table I.

Metastasis in ovarian cancer

Ovarian cancer has the highest mortality rate of the gynecological cancers and a majority of patients are diagnosed in an advanced stage [21]. Ovarian cancer predominantly metastasises within the peritoneal cavity and through the pelvic lymph nodes [22]. In the peritoneal cavity cancer cells metastasize through a process commonly described as transcoelomic dissemination, where the cancer cells loose cell-cell contact and exfoliate into the peritoneal cavity. They float in the peritoneal fluid and are spread across the peritoneal cavity, where they attach to the peritoneal organs and form a metastatic tumour [23]. Ascites produced in

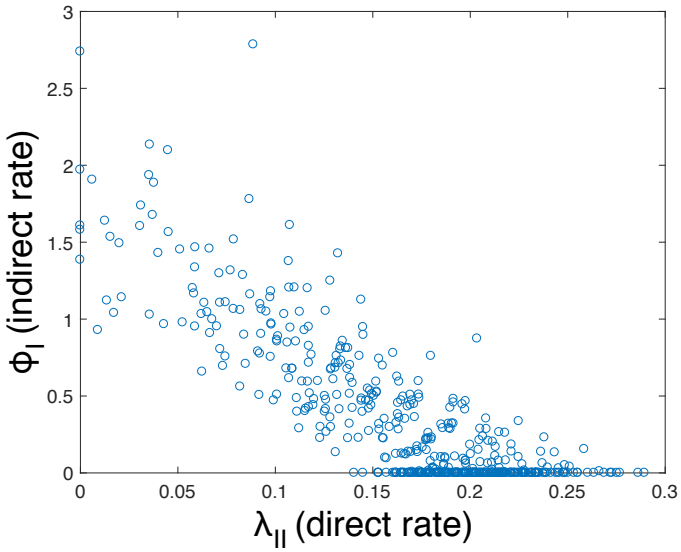

FIG. 3. Scatter plot for parameter pairs (λ_{II}, ϕ_I) determining the inflow into station II obtained from bootstrapping procedure. When performing the bootstrapping two distinct clusters of parameter pairs appear: either a large indirect rate ϕ_I (and small direct rate λ_{II}) or vice versa.

the peritoneal cavity is drained through lymph vessels in the diaphragm [24], enabling cancer cells to enter into the blood circulation. Historically, hematogenous metastasis has been regarded as occurring only in late stages of ovarian cancer. Recent work however, suggest that this mode of dissemination may be more common than previously thought [25]. In the case of distant metastases, the most common sites are liver, pleura, lung, central nervous system and skin [26].

The data set on ovarian cancer was obtained from the SEER-database (see Methods). As of 2010 SEER contains information about the presence of absence of metastases at diagnosis in liver, lung, brain and bone. The status of regional LNs (including the pelvis and diaphragm) is also available.

Given the available data we did not try to model transcoelomic dissemination, and instead focused on dissemination to local LNs and hematogenic spread to the organs represented in the SEER database. From a primary tumour located in the ovaries cancer cells can thus either spread to regional LNs or via the venous blood vessels to the lungs (see fig. 4). Metastases in regional LNs also allow for dissemination to the lungs, and from there metastatic lesions shed CTCs to all organs of the body, including the liver, bone and brain

(for which we have data).

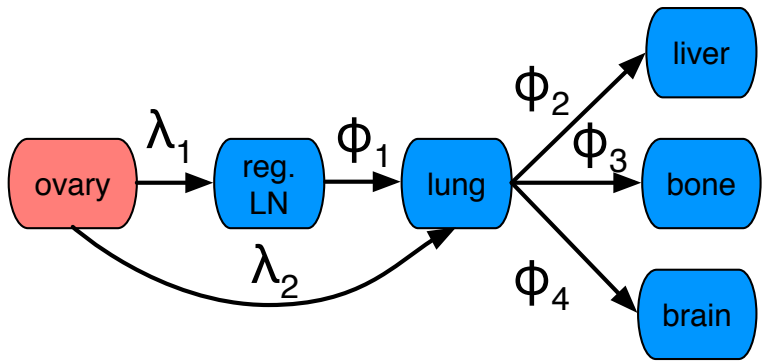

FIG. 4. The dissemination network for ovarian cancer. The tumour cells spread either via regional lymph nodes or directly to the lung where they form metastases. From there further dissemination to the liver, bone and brain occurs.

Again our aim is to estimate the dissemination parameters $(\lambda_1, \lambda_2, \phi_1, \phi_2, \phi_3, \phi_4)$ from clinical data. In this case the data set is considerably larger containing 16 055 patients diagnosed with ovarian cancer. The primary tumour stage for this data is more refined and each T-stage is divided into three substages a,b and c, giving us in total 9 different stages (T1a-T3c). We calculate the probability of finding a metastasis in each of the sites considered (regional LN, lung, liver, brain and bone) as a function of the primary tumour stage, using equation (1).In order to fit the parameters we apply a similar maximum likelihood methods as for the previous data set (see Methods). A comparison between the data and the model is shown in fig. 5, which shows that the model is able to recapitulate the overall behaviour of the data. Due to the low incidence for bone and brain (in total 65 and 16 cases respectively) the data is quite noisy and the model represents a poor fit for those sites. The parameter values are collected in table II together with 95 % confidence intervals obtained using parametric bootstrap (see Methods). Again we note large confidence intervals, in this case for the dissemination rate from regional LN to lung. This occurs because there are two parallel routes of spread and the data can be explained by many combinations of direct and indirect dissemination.

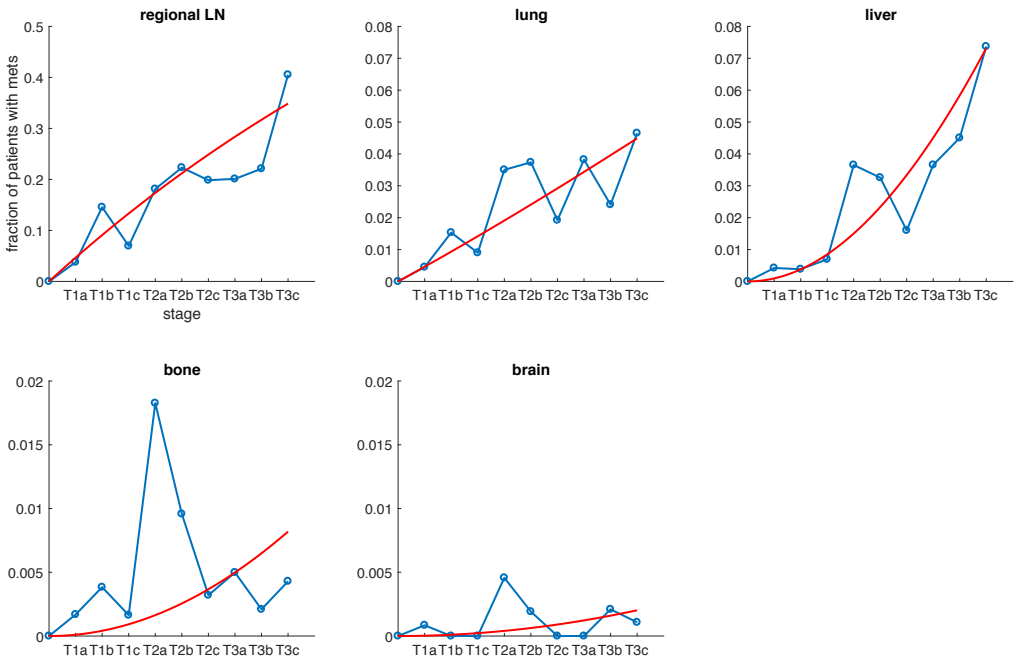

FIG. 5. The data and model predictions for metastatic spread to regional lymph nodes and distant sites for primary ovarian cancer. Each panel corresponds to a site/organ and shows the probability of finding a metastases at the site as a function of the primary tumour stage. The parameters of the model are given in table II.

TABLE II. The parameter estimates for the metastatic dissemination of ovarian cancer.

Parameter	direct metastatic rate λ	95 % CI
λ_1	0.05	$[0.0465, 0.0488]$
λ_2	0.0046	$[0.0029, 0.0053]$
ϕ_1	0.01	$[1.5 \times 10^{-10}, 0.0356]$
ϕ_2	0.38	$[0.3397, 0.4171]$
ϕ_3	0.04	$[0.0329, 0.0505]$
ϕ_4	0.01	$[0.0059, 0.0144]$

DISCUSSION

We have presented a novel method for inferring the rates of metastatic dissemination, and shown that that one can obtain reliable estimates for both lymphatic and hematogeneous spread. Our work builds on previous mathematical models [15], but with additional biologi-

cal knowledge we simplify the model structure and make parameter estimates more reliable. Firstly, we make use of the fact that high filtration rates imply that secondary seeding is responsible for metastatic spread beyond the first capillary bed/lymph node. This implies that we do not need to consider all possible links between the sites (e.g. no direct link between ovary and liver). Secondly, we make use of known anatomical structures and flow directions to further prune the network (e.g. the flow in the lymphatic system dictates the topology of the tongue cancer network (see fig. 1)). Lastly, we make use of primary tumour stage as a proxy for time, which means that we can resolve the data temporally. This implies that we do not have to rely on assumptions about stationarity of the underlying metastatic process, and instead fit the parameters for the time-dependent problem.

Our method makes it possible to infer dissemination rates with high accuracy (cf. the small CI for nearly all parameters). The model does not fit the data perfectly, but this rather due to noise in the data, than to deficiencies in the model. Traditionally this type of data is analysed by looking at incidence rates [27]. Our analysis goes beyond this by disentangling incidence rates into dissemination rates between different lymph node stations/organs. This means that we can quantify processes that at a first glance seem inaccessible, but which appropriate assumptions and modelling techniques can help us reveal.

We did encounter one issue when it comes to parameter identifiability. Both in the case of tongue and ovary cancer our model was unable to accurately identify the rate of dissemination into a site which has flow from both the primary tumour and an upstream site (station II and lung). This problem is not due to the choice of model, but a more fundamental problem where given a site with two inlets (i.e. for lung we have inflow from the primary and regional LN) the total inflow can be obtained with infinitely many combinations of flow from the two inlets. A possible solution to this problem would be to use known flow rates to constrain the possible dissemination rates in the model.

The results presented here could be followed in many interesting directions. For example, it would be intriguing to see how well the model generalises to other tumour types, and how the rates of dissemination differ between tumour from different primary sites. This type of analysis would make it possible to quantify 'seed-soil' effects, which would appear as a variation in dissemination rates. Further, it would be possible to simulate the effect of treatment, which would reveal potential benefits of treating metastatic sites.

From a clinical point of view, this work could be of importance, by contributing to

an increased possibility to predict the risk of future regional and/or distant metastases. Especially so in the current era, with new treatment modalities emerging and a current development towards more individualised treatment programs.

In conclusion we believe that this framework for analysing metastatic spread, which incorporates known anatomical constraints and a temporal dimension, allows for novel insights and will hopefully be of assistance to both cancer biologists and clinicians in the future.

METHODS

Data

Tongue cancer. After approval from the the Regional ethical review board in Gothenburg, data was obtained from the register of the department of Oncology at Sahlgrenska University Hospital in Gothenburg. Data for all patients diagnosed with carcinoma of the oral tongue between 2004 and 2014 were extracted, a total number of 141 cases. Information about primary tumour, LN metastasis and distant metastis was registered according to the TNM classification system (AJCC 6th edition until 2009 and 7th edition from 2010 on). T-stage varied from T1 to T4, 78 patients had regional LN metastes (N1 - N3) and only one patient had a distant metastasis (M1), in the lung. For patients with positive nodal status, we collected information about in which lymph node levels (I-V) metastases were present. Only one patient exhibited metastasis in lymph node level V, and therefore this level was excluded from the analysis. Information about involved LN levels was in approximately 50 % of the cases obtained from radiological examinations (CT scans or magnetic resonance imaging), while for the remainig part this information was obtained from the pathology reports performed after primary surgery. For each patient we had the following information:

- T-stage (derived AJCC 6th and 7th edition)
- Presence/Absence of regional LN metastases (calculated from N-stage)
- Localization of regional LN metastases (lymph node levels I-VI)

Ovarian cancer. The data was obtained from the case listing database *Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2016 Sub (2000-*

2014). We extracted all cases of ovarian cancer from 2010-2014 with recorded metastases in liver, lung, brain and bone. From this set of patients we selected those with a T-stage (derived AJCC 7th edition) which was in the range T1a to T3c. Information about the presence or absence of regional lymph node metastases was obtained by looking at the N-stage (derived AJCC 7th edition). We considered patients with N0 to be free of regional LN metastases while all other stages to be indicative of positive LN. The total number of cases that met the above critera was 16 055 and for each case we had the following information:

- T-stage (derived AJCC 7th edition)
- Presence/Absence of regional LN metastases (calculated from N-stage)
- Presence/Absence of metastases in lung (CS-mets at DX-lung)
- Presence/Absence of metastases in liver (CS-mets at DX-liver)
- Presence/Absence of metastases in bone (CS-mets at DX-bone)
- Presence/Absence of metastases in brain (CS-mets at DX-brain)

Mathematical model

The model consists of N nodes each representing a specific site/organ. To each node we associate a random variable $X_i(t)$, $i = 1, ..., N$, which takes the value 0 at time t if the site is void of metastases and 1 if the site contains one or more metastases. The state of each site evolves according to a continuous-time Markov chain with state space $\{0, 1\}$ and the rate at which site i becomes positive (transitions from $0 \to 1$) is given by

$$\Lambda_i(t) = \sum_{i=1}^{N} \beta_{ij} X_j(t),$$

where β_{ij} is the rate of CTC flow from site j to i. We assume that once a site has become positive it will remain so for all future times, i.e. the transition rate from $1 \to 0$ is zero for all sites. The initial condition is $X_1(0) = 1$, where 1 is the site of the primary tumour, and $X_k(0) = 0$ for $k > 1$.

Due to the sparseness of the network (most β_{ij}'s are zero), and for notational convenience we denote flow rates from the primary with λ_i and flow rates between sites with ϕ_i. We now move on to specify the models for tongue cancer and ovarian cancer respectively.

Tongue cancer. We let $P_i(t)$ denote the probability that LN station i is negative (i.e. $P_i(t) = \Pr(X_i(t) = 0)$), where $i = I, II, III, IV$. Then for station I we have that since $\Lambda_I = \lambda_I$ is a constant, $P_I(t)$ obeys the ordinary differential equation

$$\frac{dP_I(t)}{dt} = -\lambda_I P_I(t),$$

where λ_I is the dissemination rate from the primary tumour. This equation, with the initial condition $P_I(0) = 0$ (the station is always negative at tumour initiation), has the solution $P_I(t) = e^{-\lambda_I t}$. The probabilities for the remaining stations can be written

$$\begin{aligned}
\frac{dP_{II}(t)}{dt} &= -\lambda_{II} P_{II}(t) - \phi_I(1 - e^{-\lambda_I t}) P_{II}(t) \\
\frac{dP_{III}(t)}{dt} &= -\phi_{II} P_{II}(t) P_{III}(t) \\
\frac{dP_{IV}(t)}{dt} &= -\phi_{III} P_{III}(t) P_{IV}(t),
\end{aligned}$$

where the dissemination rates are given in fig. 1. The equation for $P_{II}(t)$ can be solved explicitly to yield

$$P_{II}(t; \lambda_I, \lambda_{II}, \phi_I) = e^{\phi_I/\lambda_I - (\phi_I + \lambda_{II})t - \phi_I e^{-\lambda_I t}/\lambda_I}.$$

For $P_{III}(t)$ and $P_{IV}(t)$ it is not possible to get closed form expression and we therefore have to rely on numerical solutions.

To estimate the parameters we make use of a maximum likelihood method. Preferably we would like to have the time until onset of metastasis for each station in each patient, but we only have access to data obtained at diagnosis. This information can be used in the following way: Assume that we observe a patient with a stage τ tumour that has a positive lymph node at station I . The probability of this occurring is equal to the onset of metastasis at station I being smaller than τ, i.e.

$$\Pr(\text{lymph node is positive}) = \Pr(\text{metastasis appeared at } t < \tau) = 1 - e^{-\lambda\tau}. \tag{2}$$

On the other hand, the probability of the patient being node negative is simply given by

$$\Pr(\text{lymph node is negative}) = e^{-\lambda\tau}. \tag{3}$$

If we let M_τ denote the total number of patients at stage τ, and N_τ^I the number of patients with stage τ and positive lymph nodes at station I, then the likelihood of the observations

in our data is given by:

$$L(\lambda_I) = \prod_{\tau=1}^{4} \left(1 - e^{-\tau\lambda_I}\right)^{N_\tau^I} \left(e^{-\tau\lambda_I}\right)^{M_\tau - N_\tau^I}, \tag{4}$$

which depends on the direct dissemination rate λ_I. In order to find the λ_I that best fits the data we maximise $L(\lambda_I)$ by solving the equation

$$\frac{\partial(\log L)}{\partial \lambda_I} = 0 \tag{5}$$

for λ_I. The same procedure is then applied to station II-IV to obtain all the dissemination rates in the model. Confidence intervals are calculated using parametric bootstrapping with 500 samples.

Ovary cancer. Again we let $P_i(t)$ denote the probability that site i is negative, where $i =$ LN, lung, liver, bone and brain. $P_{LN}(t)$ obeys the ordinary differential equation

$$\frac{dP_{LN}(t)}{dt} = -\lambda_1 P_{LN}(t),$$

where λ_1 is the dissemination rate from the primary tumour. This equation, with the initial condition $P_{LN}(0) = 0$ (the LN are always negative at tumour initiation), has the solution $P_{LN}(t) = e^{-\lambda_{LN} t}$. The probabilities for the remaining sites can be written

$$\begin{aligned}
\frac{dP_{lung}(t)}{dt} &= -\lambda_2 P_{lung}(t) - \phi_1(1 - e^{-\lambda_1 t})P_{lung}(t) \\
\frac{dP_{liver}(t)}{dt} &= -\phi_2 P_{lung}(t) P_{liver}(t) \\
\frac{dP_{bone}(t)}{dt} &= -\phi_3 P_{lung}(t) P_{bone}(t), \\
\frac{dP_{brain}(t)}{dt} &= -\phi_4 P_{lung}(t) P_{brain}(t),
\end{aligned}$$

where the dissemination rates are given in fig. 4. The equation for $P_{lung}(t)$ can be solved explicitly to yield

$$P_{lung}(t; \lambda_1, \lambda_2, \phi_1) = e^{\phi_1/\lambda_1 - (\phi_1+\lambda_2)t - \phi_1 e^{-\lambda_1 t}/\lambda_1}.$$

For the remaining sites it is not possible to get closed form expression and we therefore have to rely on numerical solutions.

In order to find the most likely parameters that describe the data we again make use of the maximum likelihood method described above.

ACKNOWLEDGMENTS

Philip Gerlee was supported by the Swedish Foundation for Strategic Research (grant no. AM13-0046) and Vetenskapsradet (grant no. 2014-6095).

[1] L. Weiss, K. Haydock, J. Pickren, and W. Lane, The American journal of pathology **101**, 101 (1980).

[2] S. K. Mukherji, D. Armao, and V. M. Joshi, Head & Neck **23**, 995 (2001).

[3] T. P. Butler and P. M. Gullino, Cancer research **35**, 512 (1975).

[4] M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, and M. F. Clarke, Proceedings of the National Academy of Sciences **100**, 3983 (2003).

[5] Y. Okumura, F. Tanaka, K. Yoneda, M. Hashimoto, T. Takuwa, N. Kondo, and S. Hasegawa, The Annals of thoracic surgery **87**, 1669 (2009).

[6] L. Weiss, Clinical and Experimental Metastasis **10**, 191 (1992).

[7] J. G. Scott, A. G. Fletcher, P. K. Maini, A. R. Anderson, and P. Gerlee, European journal of cancer **50**, 3068 (2014).

[8] Scott, D. Basanta, A. Anderson, and P. Gerlee, Journal of The Royal Society Interface **10**, 1 (2013).

[9] L. Liotta, G. Saidel, and J. Kleinerman, Biometrics , 535 (1976).

[10] L. Liotta, C. Delisi, G. Saidel, and J. Kleinerman, Cancer Letters **3**, 203 (1977).

[11] G. Saidel, L. Liotta, and J. Kleinerman, Journal Of Theoretical Biology **56**, 417 (1976).

[12] K. Iwata, K. Kawasaki, and N. Shigesada, Journal Of Theoretical Biology **203**, 177 (2000).

[13] E. Baratchart, S. Benzekry, A. Bikfalvi, T. Colin, L. S. Cooley, R. Pineau, E. J. Ribot, O. Saut, and W. Souleyreau, PLoS Computational Biology **11**, e1004626 (2015).

[14] J. Scott, P. Kuhn, and A. R. A. Anderson, Nature Reviews Cancer , 1 (2012).

[15] P. K. Newton, J. Mason, K. Bethel, L. A. Bazhenova, J. Nieva, and P. Kuhn, PLoS ONE **7**, e34637 (2012).

[16] P. K. Newton, J. Mason, K. Bethel, L. Bazhenova, J. Nieva, L. Norton, and P. Kuhn, Cancer Research **73**, 2760 (2013).

[17] G. Disibio and S. W. French, Archives of pathology & laboratory medicine **132**, 931 (2008).

[18] P. Gerlee, Cancer research **73**, 2407 (2013).

[19] R. M. Byers, R. S. Weber, T. Andrews, D. McGill, R. Kare, and P. Wolf, Head & neck **19**, 14 (1997).

[20] H. Wang, C. Zhang, J. Zhang, L. Kong, H. Zhu, and J. Yu, Oncotarget **8**, 26368 (2017).

[21] R. L. Siegel, K. D. Miller, and A. Jemal, CA: a cancer journal for clinicians **65**, 5 (2015).

[22] N. Tsuruchi, T. Kamura, N. Tsukamoto, K. Akazawa, T. Saito, T. Kaku, N. To, and H. Nakano, Gynecologic oncology **49**, 51 (1993).

[23] A. K. Mitra, in *Tumor Metastasis* (InTech, 2016).

[24] D. S. Tan, R. Agarwal, and S. B. Kaye, The lancet oncology **7**, 925 (2006).

[25] S. Pradeep, S. W. Kim, S. Y. Wu, M. Nishimura, P. Chaluvally-Raghavan, T. Miyake, C. V. Pecot, S.-J. Kim, H. J. Choi, F. Z. Bischoff, *et al.*, Cancer cell **26**, 77 (2014).

[26] G. Cormio, C. Rossi, A. Cazzolla, L. Resta, G. Loverro, P. Greco, and L. Selvaggi, International Journal of Gynecological Cancer **13**, 125 (2003).

[27] M. Qiu, J. Hu, D. Yang, D. P. Cosgrove, and R. Xu, Oncotarget **6**, 38658 (2015).