Physics in Medicine &
Biology

IPEM
Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT • OPEN ACCESS

Probabilistic clinical target definition with nearest neighbor correlation

To cite this article before publication: Luciano Rivetti *et al* 2025 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/ae2aa1

Probabilistic Clinical Target Definition with Nearest Neighbor Correlation

Rivetti L. *[1,2], Buti G.[2], Amoudruz L.[3], Ajdari A.[2], Sharp G.[2], Studen A.[1,4], Jeraj R[1,5,4], and Bortfeld T.[2]

[1] *Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia*
[2] *Massachusetts General Hospital and Harvard Medical School, Boston, United States of America*
[3] *School of Engineering and Applied Sciences, Harvard University, Cambridge, United States of America*
[4] *Jožef Stefan Institute, Ljubljana, Slovenia*
[5] *University of Wisconsin–Madison, Madison, United States of America*

Abstract

Objective: The delineation of the clinical target volume (CTV) in radiotherapy is fundamentally uncertain due to the invisibility of microscopic disease on medical images. The ICRU 83 report acknowledges this by proposing a probabilistic interpretation of the CTV, but it does not define how to compute the probability of microscopic tumor presence (MTP) in tissue. This work addresses this gap by introducing a novel stochastic model that estimates the probability of MTP at the voxel level based on local spatial correlations in the voxels' neighborhood.

Approach: We developed two first-principles stochastic models to simulate MTP under different assumptions, incorporating spatial correlation between neighboring voxels. The constant marginal probability (CMP) model assumes spatially uniform MTP and is suited for tumors without radial dependence on the distance from the gross tumor volume (GTV). The variable marginal probability (VMP) model introduces radial dependence, modeling decreasing MTP with distance from the GTV. The CMP model was evaluated on prostate cancer data, while the VMP model was assessed using breast and lung cancer data.

Results: Both models accurately reproduced the fraction of times that MTP is present. In the prostate case, the CMP model estimated a marginal probability of MTP of 0.03, consistent with a literature report that indicates an average total microscopic tumor volume of approximately 583 mm^3 across patients. The VMP model successfully replicated the radial distribution of tumor islets, achieving mean absolute errors of 0.01 mm and 0.011 mm for breast and lung cancer distance distributions, respectively. However, not all MTP characteristics could be fully captured by the models, and in some cases discrepancies with population based tumor characteristics remain.

Significance: This work introduces a statistically consistent framework that enables a probabilistic definition of the CTV. The proposed models provide a new way to capture key aspects of microscopic disease spread by introducing local voxel correlations.

*Corresponding author: Luciano.Rivetti@fmf.uni-lj.si

1 Introduction

Accurate delineation of the clinical target volume (CTV) is a critical step in radiation therapy, directly influencing treatment outcomes [21, 11, 27]. CTV delineation is very challenging, largely due to the big uncertainties inherent in identifying microscopic disease spread [29, 30, 20]. Despite advancements in imaging, inconsistencies in CTV contouring remain a persistent issue, especially in complex cases where inter-observer variability can reach up to one centimeter or more [29]. This uncertainty, often an order of magnitude greater than the millimeter-level precision achieved in dose delivery, underscores a fundamental limitation within current CTV practices.

To address the inherent variability of manual CTV delineation, a probability-based definition of the CTV driven by clinical data is desirable. The ICRU 83 report introduced the concept of probabilities in the CTV definition, stating: "The CTV is a volume of tissue that contains a demonstrable gross tumor volume (GTV) and/or subclinical malignant disease with a certain *probability of occurrence* [i.e. presence] considered relevant for therapy." It further specifies that "typically, a *probability of occult disease* higher than 5% to 10% is assumed to require treatment" [16]. However, this definition raises key questions: How should the "probability of subclinical malignant disease occurrence" or "probability of occult disease" in a tissue be statistically defined? On what clinical evidence should these probabilities be based? And how can probabilities assigned to tissues be translated into discretized grids used in treatment planning for dose calculation and optimization?

For instance, consider modeling a tissue as a collection of voxels, where we assume that the probability of a tumor cell being present in each voxel is known. In this scenario, the probability of finding at least one tumor cell in a larger region, composed of multiple voxels, will likely increase with the size of that region. However, this is true if and only if the probabilities of finding tumor cells in the voxels do not directly influence each other. Consider the opposite scenario where the tumor cells found in the voxels within the same region are known to be perfectly influenced by each other—if one voxel contains tumor, then all the others in the region do as well, and vice versa. In this case, the probability of finding a tumor cell in the entire region remains constant, regardless of its size. This simple example underscores the critical role of the assumed level of influence that voxels have on each other in determining whether the larger region composed of those voxels should be targeted for treatment, as outlined in the ICRU report. We refer to this problem as "voxel correlation". It emphasizes the need for a deeper understanding of these definitions, the influence of voxel correlations, and the development of models that accurately reflect clinical data.

The first studies that introduce explicit representations of probabilities into radiation target volumes were presented by authors that aim to create a fully probabilistic definition of the CTV, the so-called clinical target map (CTM). In the CTM, each voxel outside of the GTV is assigned a probability of microscopic tumor presence (MTP) which describes the probability of finding at least one microscopic tumor cell [6, 23]. This MTP probability arises from introducing a binary random variable of MTP to a voxel which takes a positive state if the voxel contains tumor and a zero state otherwise. The MTP probabilities used for the CTM are subsequently obtained as the expected value of the MTP random variable calculated in each voxel. This definition of tumor probability addresses one of the ambiguities in the original ICRU definition; the CTM probabilities are *marginal* probabilities, meaning it expresses the probability of MTP in a voxel, without considering the state of MTP in any other voxel.

For example, the MTP model used in Shusharina et al. states that the state of MTP in one voxel is independent of the state of MTP in any other voxel. While this model is statistically interpretable and computationally simple, it has notable limitations [23]. First, the independence assumption does not fit well with microscopic tumor patterns reported in the literature, such as [15, 31], where it is observed that MTP often manifests as clusters or "islets" of millimeter size around the GTV. This suggests that the presence of tumor in one location conditions the probability of tumor in its neighborhood.

A different MTP model, referred to as *contiguous circumferential growth model*, was proposed by Buti and Bortfeld et al. [7, 5]. Here it was assumed that the state of MTP in a voxel fully determines the state of MTP of all voxels at a distance less than or equal to that voxel. This model was inspired by the way CTV is defined in clinical guidelines; i.e., the CTV margin includes a percentage of *maximum* distances of MTP identified on pre-treated histopathological specimens for a disease site [13, 18]. Buti et al.

modeled the radial dependence of MTP using the maximum microscopic tumor distance probability density function (PDF) reported in histopathologic studies such as [17]. While this model has practical appeal, it cannot reproduce aforementioned microscopic tumor islets observed and therefore overestimates the marginal probability of MTP in a sub-volume. Overestimating MTP probabilities could lead to overly conservative radiotherapy treatment plans that potentially over-dose healthy tissue.

In this study, we present a stochastic model for MTP that quantifies the dependence of MTP in a voxel on its local neighborhood. This dependency is introduced through the use of a statistically well-known concept, the correlation factor, which imposes discrete rules akin to those in the design of a cellular automaton [4]. It will be demonstrated that the correlation parameter regulates the number and size of microscopic tumor islets appearing outside of the GTV. The model parameters are calibrated using MTP characteristics extracted from histo-pathological findings reported for a population of patients of three specific cancer types: prostate, breast, and lung. The model is evaluated based on its ability to replicate observed microscopic tumor features, thereby providing insights into its potential applicability for clinical use in radiotherapy planning.

2 Material and Methods

In this section, we first introduce the terminology and general statistical framework used to model tumor probabilities throughout this study. Second, we show that existing statistical models, i.e. the *independent* model first proposed by Shusharina et al. and the contiguous circumferential growth model proposed by Buti and Bortfeld et al. [7, 5], fail to reproduce the observed clinical data. Third, we propose two new models based on first principles that allow the generation of more realistic patterns. Fourth, we describe the literature data we used to fit our models. Fifth, we describe the simulations we made and the metrics we evaluated.

2.1 Terminology

We follow the terminology introduced by the ICRU by referring to any malignant disease outside of the GTV as "microscopic tumor presence" (MTP). We apply this term irrespective of the fact that the disease extended from the GTV or originated at a specific location. MTP implies that at least one tumor cell is present at that specific location, whereas no MTP means that that location is tumor-free.

2.2 General statistical framework

Consider a 3D grid with N voxels. Each voxel $i \in \{1, \ldots, N\}$ is associated with a binary random variable C_i, where $C_i = 1$ indicates the presence of MTP, and $C_i = 0$ indicates its absence. We use lowercase c_i to denote a specific realization (observed value) of the random variable C_i. The *marginal probability* of MTP in voxel i is denoted p_i, and corresponds to the expected value of the random variable: $\mathbb{E}[C_i] = p_i$.

For the sake of simplicity, we will refer to the state or probability of the MTP random variable in voxel i simply as the state or probability of voxel i. We assume that the probability of all voxels in the GTV is equal to 1, $p_i = 1\ \forall i \in$ GTV. Once the states of all voxels in the grid have been assigned, we can visualize the combined GTV and MTP regions as a single realization, referred to as the *tumor configuration*. The probability of observing a specific tumor configuration is given by the corresponding *joint probability* over the entire grid. This joint probability quantifies the likelihood of the system as a whole adopting a particular configuration, rather than considering voxel states independently.

2.3 Limitations of the current models

The independent model of Shusharina et al. assumes that there is no correlation between the states of any two voxels C_i and C_j, $i \neq j$ [23]. Therefore, tumor configurations can be generated from this model by randomly sampling the state of each voxel i of the grid independently with probability p_i. Simulations in a

circular geometry, assuming a constant marginal probability throughout the prostate, show that a "poppy seed" pattern of MTP is generated (Fig. 1, b-d). This pattern differs from the characteristic islet formations of MTP observed in histopathologic images (Fig. 1a) [15, 31].

The contiguous circumferential growth model from Buti and Bortfeld et al. imposes a direct correlation among the voxels. Specifically, if the state of a voxel at distance d from the GTV is randomly assigned as MTP, then all voxels within a distance smaller or equal to d are also classified as MTP. Figs. 1, e-g show that tumor configurations produce an "onion" pattern with the MTP appearing as being thick, isotropic expansions of the GTV surface, thereby failing to capture the islet formations.

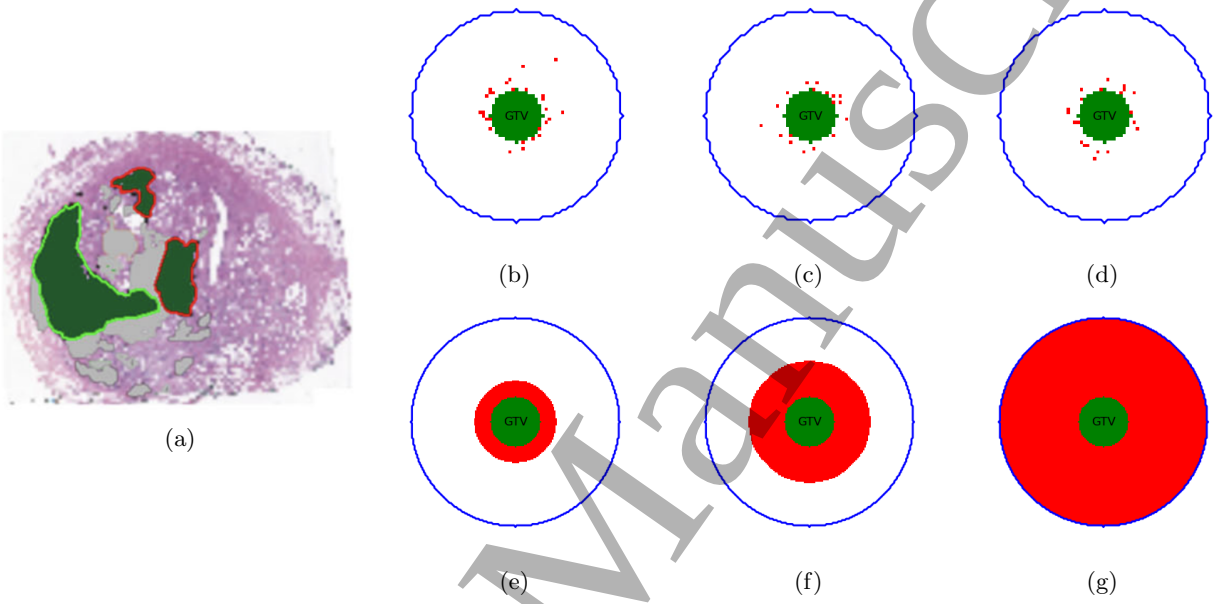


Figure 1: Histopathology and simulation of MTP. The green contours (or circle) indicates the GTV, the red contours/structures represent MRI-invisible tumor lesions (i.e., MTP in our framework), and the blue contours depict an anatomical barrier impermeable to tumor cells, such as the prostate wall. (a) Real histopathology slice from a prostate cancer patient, adapted from van Houdt et al. [15] with their permission. (b–d) 2D simulations generated using the independent model proposed by Shusharina et al. [23], assuming a constant marginal probability throughout the prostate (2 shell model), with a circular GTV. Each dot represents a voxel. (e–g) 2D simulations performed using the contiguous circumferential growth model proposed by Buti et al. [7], also with a circular GTV.

It is evident that these two models represent extreme cases with rigid assumptions: either complete voxel independence or full correlation. We propose that a more realistic model can be achieved by introducing a parameter to model the degree of voxel correlation, bridging the gap between these extremes and better reflecting observed MTP patterns.

2.4 Nearest neighbor correlation with constant marginal probability of MTP

Our approach differs from the existing methods by introducing pairwise correlations between voxels and by keeping the marginal probability in every voxel within a specified volume the same, here denoted as p. Similar to the design of a cellular automaton, fixed rules are defined that determine the state of each voxel based on the voxels in its neighborhood. The algorithm begins by selecting a random voxel $i = 1$ from the grid and sampling its state $c_1 \sim \text{Bernoulli}(p)$. Then, the states of its nearest neighbor (NN) voxels are sequentially sampled according to a predefined visiting order. The state c_i of a newly visited voxel i is determined in one of two ways:

1. with probability q, its state is copied from a randomly selected previously visited nearest neighbor $k \in \mathrm{vNN}_i$, i.e., $c_i = c_k$, where $\mathrm{vNN}_i = \{k \mid k \text{ was visited before } i\}$;

2. with probability $1 - q$, a new value is sampled independently: $c_i \sim \mathrm{Bernoulli}(p)$.

This process ensures that the state of each newly visited voxel i is partially inherited from its visited nearest neighbors.

The visiting algorithm follows a structured traversal pattern to systematically explore the voxel grid. The process starts with a randomly selected voxel, after which its NN voxels are visited in a cross-like pattern, moving sequentially up, left, down, and right (Fig. 2b–e). Once all first-order NNs of the initial voxel have been visited, the algorithm continues with the NNs of the next visited voxel, ensuring that no voxel is revisited (Fig. 2f–h). This structured approach is iteratively applied until all voxels in the grid have been assigned a state.

Following these rules, the *conditional probability* of C_i given its visited nearest neighbors can be written as:

$$\mathbb{P}\big(C_i = 0 \;\big|\; \{C_k = c_k\}_{k\in \mathrm{vNN}_i}\big) = q\left(1 - \frac{1}{n}\sum_{k\in \mathrm{vNN}_i} c_k\right) + (1-q)\,(1-p), \quad n = \big|\mathrm{vNN}_i\big|, \tag{1}$$

Using the conditional and marginal probabilities defined above, we can calculate the joint probability of a given state using Bayes' theorem:

$$\mathbb{P}(C_i, \{C_k\}_{k\in \mathrm{vNN}_i}) = \mathbb{P}(C_i \mid \{C_k\}_{k\in \mathrm{vNN}_i})\mathbb{P}(\{C_k\}_{k\in \mathrm{vNN}_i}). \tag{2}$$

Thus, the probability of the configuration b) in figure 2 is equal to $p(q + (1-q)p)$. With this algorithm, if we assign $\mathbb{E}[C_0] = p$ to the first voxel, it is easy to show that

$$\mathbb{E}[C_i] = \mathbb{E}[C_k] = p \quad \forall\, i, k, \tag{3}$$

as intended. This result holds for any value of the correlation coefficient q, highlighting that p and q are independent parameters of the model..

The statistical interpretation of q can be clarified by evaluating the Pearson correlation coefficient ρ between a random variable C_i and its nearest neighbor C_k on a 1D grid. The correlation coefficient ρ is given by:

$$\rho_{C_i,C_k} = \frac{\mathrm{COV}_{C_i,C_k}}{\sigma_{C_i}\sigma_{C_k}} = \frac{\mathbb{E}[C_iC_k] - \mathbb{E}[C_i]\mathbb{E}[C_k]}{\sigma_{C_i}\sigma_{C_k}}. \tag{4}$$

Since the expected values of C_i and C_k are equal to p, and we are dealing with binary random variables, the standard deviations are $\sigma_{C_i} = \sigma_{C_k} = \sqrt{p - p^2}$. Because in a 1D grid there is only one NN, we have $\mathbb{E}[C_iC_k] = qp + (1-q)p^2$. This yields

$$\rho = \frac{qp + (1-q)p^2 - p^2}{p - p^2} = q. \tag{5}$$

Therefore, in the case where the marginal probability is the same in the whole grid, the design parameter q corresponds to the Pearson correlation coefficient ρ between two voxels in a 1D grid. However, in higher-dimensional grids (2D or 3D), the Pearson correlation coefficient is given by $\rho = \sum_{n=1}^{\infty} a_n q^{2n-1}$ where the coefficients a_n satisfy $\sum_{n=1}^{\infty} a_n = 1$ (Appendix A). The values of n and a_n depend on the dimension and size of the grid.

An important question is the probability of finding no tumor cells in a group of voxels. We can calculate the probability of having no tumor cells in an entire grid with N voxels using equation 1 and 2 as:

$$\mathbb{P}(C_0 = 0, \ldots, C_N = 0) = P_0(p, q, N) = (1-p)\Big(q + (1-q)(1-p)\Big)^{N-1}. \tag{6}$$

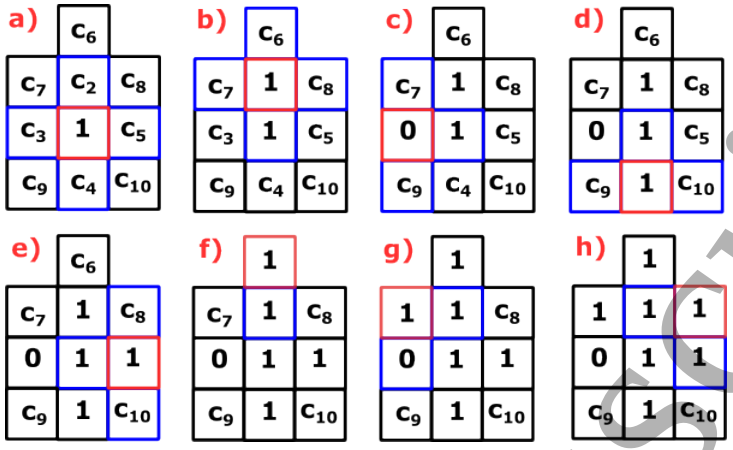


Figure 2: Illustration of the sampling algorithm in a 2D grid. The red squares indicate the last visited voxel at each step and the blue squares shows their NN. (a) The algorithm begins by randomly selecting a voxel in the grid (C_1) and sampling its state with probability p. (b-e) Then, all the NN of the first sampled voxel are visited incorporating correlation effects for each neighboring voxel. (f-h) Once all NN of C_1 are visited, the algorithm proceeds by sampling voxels adjacent to the second visited voxel (C_2) without revisiting voxels.

However, this result is specific to the entire grid and cannot be directly applied to subsets of voxels. In subsets, it is not guaranteed that every voxel has a nearest neighbor within the subset, which could affect the calculation.

Additionally, P_0 can be interpreted as the fraction of patients who do not exhibit MTP. To correctly reproduce P_0, the parameters p and q must satisfy equation 6. Since $q \in [0, 1]$, this relationship constrains the admissible range of the marginal probability p: its maximum value, $1 - P_0$, occurs when $q = 1$, while its minimum value, $1 - P_0^{1/N}$, occurs when $q = 0$.

Interpreting how p varies with changes in P_0, q, or N is challenging based on equation 6, especially since N appears as an exponent in the equation. To provide a clearer understanding of equation 6, we approximated this equation using Taylor series, assuming a large number of voxels in the grid (Appendix B). By discarding the quadratic and higher-order terms, we simplify equation 6 to:

$$\frac{P_0}{1-p} \approx e^{-N(1-q)p} = e^{-\frac{p}{T(q,N)}} \tag{7}$$

where $T(q, N) = \frac{1}{N(1-q)}$. Equation 7 shows that the marginal probability of the voxels p only depends on P_0 and $T(q, N)$. Fig 3 shows the relationship between p and $T(q, N)$ for different values of P_0 and using a fixed $N = 75735$. This value corresponds to the number of voxels in the sampling space of the prostate patient that we will describe in the evaluation subsection. We observe that $T(q, N)$ serves as a "transition factor" between two tumor configuration states, fully correlated and uncorrelated. As we will later show, when $T \to \infty$, the tumor configuration exhibits clustering, meaning the voxel states become more correlated. In contrast, when $T = \frac{1}{N} \approx 0$, the tumor configuration generates isolated islets, where voxel states are essentially uncorrelated. From figure 3, P_0 can be interpreted as influencing the slope of the transition between different tumor configurations. The definition of $T(q, N)$ ensures that our model remains resolution-independent: for a given N, we can adjust q to achieve a desired value of T, thereby making the model adaptable to different grid resolutions.

Not only p is bounded in this model but also the number of unconnected components or tumor islets. If we consider the case when $T \approx 0$, the voxels are uncorrelated and the number of unconnected components islets are maximal. Then, because in this regime $p \approx 1 - P_0^{1/N} \approx 0$, we can assume that the probability of having two tumor voxels connected is low and the maximum average number of tumor islets can be calculated

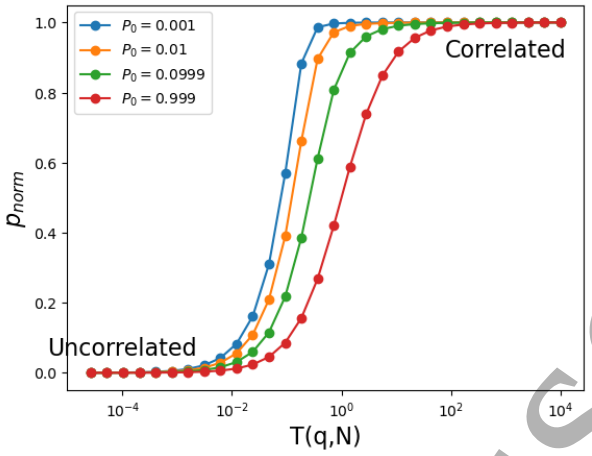


Figure 3: Normalized marginal probability versus transition factor $T(q, N)$ calculated for $N = 75735$ at different P_0. The probability p was normalized such that $p_{\text{norm}} = \frac{p - p_{\min}}{p_{\max} - p_{\min}}$, where $p_{\min} = 1 - P_0^{1/N}$ and $p_{\max} = 1 - P_0$. The plot illustrates how the system transitions from the correlated state to the uncorrelated state, with different colors representing transitions for various values of P_0. The minimum marginal probability is achieved when the system is uncorrelated ($T \to 0$ or $q = 0$), while the maximum occurs when the system is fully correlated ($T \to \infty$ or $q = 1$). The x-axis is displayed in logarithmic scale.

as Np. Using equation 7 we can write:

$$Np \approx \ln\left(\frac{1}{P_0}\right) \tag{8}$$

Conversely, in the opposite limit when $T \to \infty$, the voxels become fully correlated. In this regime the system can only generate two possible outcomes: either the entire admissible volume is tumor, or no tumor is present at all. Since the probability of no tumor is P_0, the complementary probability of having tumor is $1 - P_0$. In this case, at most a single connected islet can be formed with probability $1 - P_0$, and therefore

$$\mathbb{E}[\#\text{islets}] = \bar{N}_I = 1 \cdot (1 - P_0) = 1 - P_0.$$

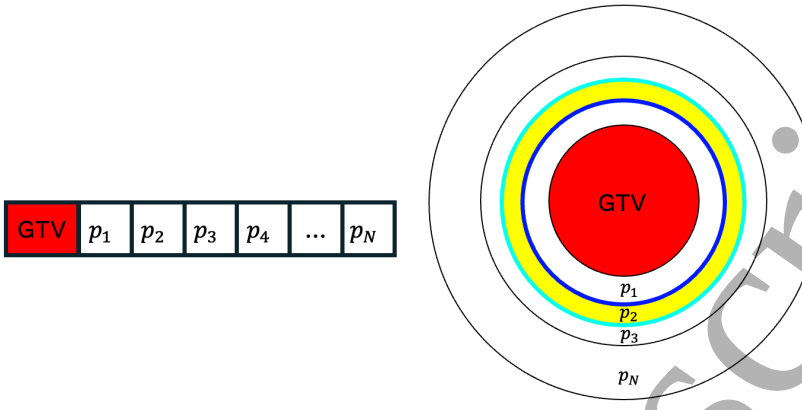
This expression defines the lower bound for the expected number of islets.

2.5 Nearest neighbor correlation with variable marginal probability of MTP

Several studies have shown that in certain tumors, the MTP decreases as the distance from the GTV border increases [25, 12, 8, 22, 28]. Our previous model cannot capture this behavior, as it assumes constant marginal probabilities across voxels. To address this limitation and incorporate radial dependence, we introduce variable marginal probabilities.

The visiting voxel strategy in this model is the same as the one described in the previous section. However, the main differences between the models is how we calculate the posterior probability of a new voxel given its visited nearest neighbor states. Because in this model the two NN voxels can have different marginal probabilities we should contemplate this case in the equations. To clarify this point, lets first consider a 1D grid (Fig 4a).

Inspired by the model proposed in our previous work [5, 6, 7], we adopted the *no-tunneling assumption*, which requires that in the fully correlated case, tumor configurations form continuous regions—i.e., isolated tumor islets do not appear.



(a) 1D grid with variable marginal probabilities (b) Shell representation of a circular GTV

Figure 4: Variable marginal probability model representation. a) Shows that the marginal probabilities in this model are higher closer to the GTV. b) Shows the shell representation used in Shusharina et al. for 2D or 3D simulations. The second shell is outlined in blue and the third shell in cyan; the yellow region between them corresponds to the second layer. This model assumes that all the voxels in the same layer have the same marginal probability.

Let C_i and C_k be two neighboring voxels such that $p_i \geqslant p_k$. Under the non-tunneling assumption, the following conditional probabilities must hold:

$$\begin{aligned} \mathbb{P}(C_k = 1 \mid C_i = 0) &= (1-q)p_k \\ \mathbb{P}(C_i = 1 \mid C_k = 1) &= q + (1-q)p_i. \end{aligned} \tag{9}$$

These expressions enforce spatial continuity when $q = 1$. Specifically, if voxel i is in the tumor state, then all voxels from i up to the GTV boundary must also be in the tumor state. This is ensured when: $\mathbb{P}(C_i = 1 \mid C_k = 1) = 1$. Conversely, if voxel i is non-tumor, all subsequent voxels up to the outer boundary N must also be non-tumor. This is captured by the condition: $\mathbb{P}(C_k = 1 \mid C_i = 0) = 0$.

To ensure consistency with the prescribed marginal probabilities $\mathbb{E}[C_i] = p_i$ and $\mathbb{E}[C_k] = p_k$, two additional conditions must be satisfied:

$$\begin{aligned} \mathbb{P}(C_k = 1 \mid C_i = 1) &= \frac{p_k}{p_i} q + (1-q)p_k \\ \mathbb{P}(C_i = 1 \mid C_k = 0) &= \frac{p_i - p_k}{1-p_k} q + (1-q)p_i. \end{aligned} \tag{10}$$

These expressions ensure that the model reproduces the correct marginal expectations. For the mathematical derivation look at Appendix C. Note that, when $p_i = p_k$, the previous equations simplifies to Equation (1), recovering the constant marginal probability model described in the previous section.

To extend this model to 2D and 3D, we adopted the shell representation proposed by Shusharina et al. [23]. In this framework, the CTM is modeled using a finite set of iso-probability surfaces, called *shells*, each associated with a probability value (Fig 4b). This probability represents the likelihood that the MTP extends beyond the corresponding shell within a given patient population. The innermost shell coincides with the GTV contour, while the outermost shell delineates the region beyond which tumor presence is considered negligible. The region between two adjacent shells is defined as a *layer*. This model also assumes that all voxels within the same layer share the same marginal probability.

The state of a new voxel C_i can be computed by considering the states of its visited nearest neighbors vNN_i along with their marginal probabilities. Assume voxel i has n visited neighbors, where the first l neighbors satisfy $p_k \geqslant p_i$ for all $k \in \{1, 2, \ldots, l\}$, and the remaining $n - l$ neighbors satisfy $p_k < p_i$ for all $k \in \{l+1, \ldots, n\}$. Then, using equations 9 and 10, we can express the conditional probability of voxel i being in state 1, given the neighboring realizations $\{c_k\}_{k \in \mathrm{vNN}_i}$, as:

$$\mathbb{P}(C_i = 1 \mid \{C_k = c_k\}_{k \in \mathrm{vNN}_i}) = \frac{q}{n}\left(\sum_{k=1}^{l} \frac{p_i}{p_k} c_k + \sum_{k=l+1}^{n} \left[c_k + (1 - c_k)\frac{p_i - p_k}{1 - p_k}\right]\right) + (1 - q)p_i, \quad n = |\mathrm{vNN}_i|\,. \tag{11}$$

In this model, we define N' marginal probabilities, denoted as $(p_1, p_2, \ldots, p_{N'})$, where N' represents the number of distinct voxel classes given by the number of layers. To fully determine these $N' + 1$ parameters, we require a corresponding set of $N' + 1$ independent equations.

Empirical studies have reported the probability of finding MTP beyond a given distance r from the nearest point on the GTV, denoted here as $\mathrm{w}(r)$ (see Section 2.6 for more details) [25, 22, 28] . We used these empirical constraints to estimate the marginal probabilities as follows:

$$1 - \mu_i(q, p_1, \ldots, p_{N'}) = \mathrm{w}(r_i), \quad \forall i \in \{1, \ldots, N\}, \tag{12}$$

where r_i is the distance to the i-th shell and $\mu_i(q, p_1, \ldots, p_{N'})$ is the probability of finding no tumor in all the voxels outside the i-th shell. Note that equation 12 must hold for any selected value of q in the model.

However, obtaining an analytical expression for $\mu_i(q, p_1, \ldots, p_{N'})$ is generally intractable. Therefore, we estimate this value using Monte Carlo simulations. Specifically, we approximate $\mu_i(q, p_1, \ldots, p_{N'})$ as the fraction of realizations that do not contain tumor voxels beyond the i-th shell. Given a fixed value of q, we can determine the optimal values $\{\hat{p}_1, \ldots, \hat{p}_{N'}\}$ by solving the following optimization problem:

$$\{\hat{p}_1, \ldots, \hat{p}_{N'}\} = \underset{\{p_1, \ldots, p_{N'}\}}{\operatorname{argmin}} \sum_{i=1}^{N} \big(1 - \mu_i(q, p_1, \ldots, p_{N'}) - \mathrm{w}(r_i)\big)^2. \tag{13}$$

2.6 Data

Population-based clinical data from the literature were used to fit the parameters of the proposed models for three different cancer types: prostate, breast, and lung. These datasets provide information on microscopic disease patterns, enabling a quantitative evaluation of the model's flexibility in capturing real-world tumor behavior. For each cancer type, relevant studies reporting histopathological findings were selected to extract parameters such as the probability of MTP presence, average islet volume, and MTP radial dependence. The data are summarized in Table 1. Definitions of all model parameters and metrics used in this study are provided in Table 2.

Prostate Data

Clinical data for prostate cancer were obtained from the article written by Bajgira et al., encompassing 518 patients who underwent multiparametric MRI followed by histological examination post-resection [19]. The study found that 76% of patients had multifocal microscopic islets detectable only through histopathology (not in MRI), corresponding to $P_0 = 0.24$. The average number of islets $\bar{N}_I$ considering the 518 patients was 0.87 per patient. The reported average islet radius was $7.82 \pm 5.74\,\mathrm{mm}$. Assuming a truncated Gaussian distribution for the radius and spherical islet shape, the estimated average islet volume ($\bar{V}_I$) was $513\,\mathrm{mm}^3$, with the 5th and 95th percentiles at $1\,\mathrm{mm}^3$ and $2024\,\mathrm{mm}^3$, respectively. The average total microscopic tumor volume ($\bar{V}_T$) was estimated by multiplying $\bar{V}_I$ by the average number of islets considering the patients which presented MTP. This resulted in an $\bar{V}_T$ equal to 583 mm^3, with the 5th and 95th percentiles at $1\,\mathrm{mm}^3$ and $2300\,\mathrm{mm}^3$. A detailed description of how these data and calculations were obtained is provided in Appendix D.

Breast Data

Breast cancer data were sourced from the publication by Stroom et al., where 60 patients underwent contrast-enhanced MRI followed by histopathological examination [25]. The study reported that 72% of patients had multifocal islets detectable only through histopathology ($P_0 = 0.28$). Among those patients with islets, the reported average number of islets was 4.2. When averaged over the entire cohort of 60 patients, including those without islets, this corresponds to an average number of islets of approximately 3 per patient. The mean islet radius was 2.1 mm, and assuming spherical islets, $\bar{V}_I$ can be approximated to 39 mm^3. Similarly $\bar{V}_T$ can be approximated to 162 mm^3. Additionally, the authors analyzed the spatial distribution of islets relative to the GTV, reporting that it follows a normal distribution with a mean of zero and a variance of 7 mm^2.

Additionally, a related study [22] provided probability data for detecting MTP beyond a given distance r from the closest point on the GTV, denoted as w(r). Note that w(0) = $1 - P_0$ and in this case this value was 0.72 (Fig 5). Since the study does not report the distribution of MTP in relation to polar coordinates, we assume an isotropic MTP distribution.

Lung Data

The lung cancer data were obtained from the publication written by Stroom et al., comprising 34 patients who underwent CT imaging followed by histopathological analysis post-resection [25]. Multifocal islets, detectable only through histopathology, were identified in 50% of cases. The reported $\bar{N}_I$ calculated over the entire cohort of 34 patients was 2.47. The mean islet radius was reported as 1.6 mm, and assuming spherical islets, the estimated $\bar{V}_I$ was 18 mm^3 and $\bar{V}_T$ was estimated to be 85 mm^3. Additionally, the authors analyzed the spatial distribution of islets relative to the GTV, reporting that it follows a normal distribution with a mean of zero and a variance of 10 mm^2.

A related study [28] provided w(r), with w(0) = $1 - P_0 = 0.5$ (Fig 5). Since the MTP distribution data in polar coordinates were not reported, an isotropic distribution was assumed.

Cancer Type	P_0	$\bar{\mathbf{N}}_\mathbf{I}$	$\bar{\mathbf{V}}_\mathbf{I}(\mathbf{mm}^3)$	$\bar{\mathbf{V}}_\mathbf{T}$ $(\mathbf{mm}^3)$	**w(r)**
Prostate	0.24	0.87	513	583	No
Breast	0.28	3	39	162	Yes
Lung	0.50	2.47	18	85	Yes

Table 1: Summary of microscopic tumor parameters for prostate, breast, and lung cancers. P_0: proportion of patients with MTP; $\bar{N}_I$: average number of islets per patient; $\bar{V}_I$: average islet volume per patient; $\bar{V}_T$: average total microscopic tumor volume among patients with microscopic tumor presence; w(r): indicates whether data on the cumulative probability distribution of detecting MTP outside a given distance from the GTV are available in the dataset (Yes/No).

Parameter/Metric	**Description**
q	Correlation factor representing the probability of copying a neighboring state
p	Marginal probability of a voxel
T	Transition factor controlling the change between MTP states
P_0	Proportion of patients without MTP
w(r)	Cumulative probability of finding MTP beyond distance r
$\bar{N}_I$	Mean number of MTP islets per patient
$\bar{V}_I$	Mean volume of MTP islets per patient (mm^3)
$\bar{V}_T$	Mean total MTP volume per patient (mm^3)

Table 2: Summary of model parameters and metrics used in this study.

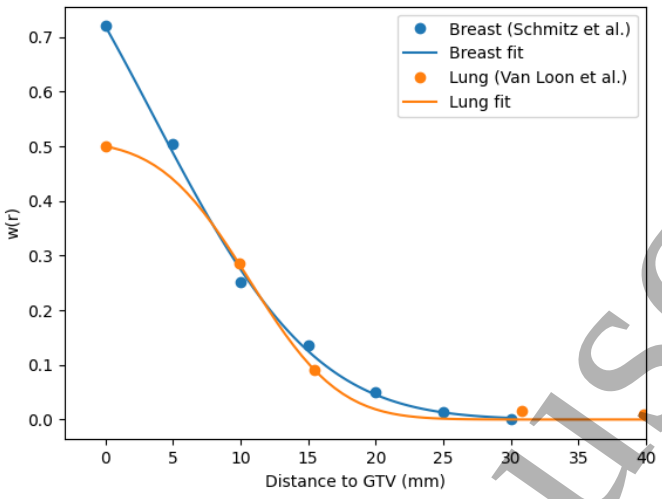


Figure 5: Probability of finding MTP beyond a given distance r from the nearest point on the GTV denoted as $\mathbf{w}(r)$. The blue points represent the data extracted from Schmitz et al.[22] for a breast cancer population and the line shows the best fit. The orange points represent the data extracted from Van Loon et al.[28] for a lung cancer population and the line shows the best fit.

2.7 Simulation design

Several simulations were ran on patient-specific anatomical data to study the properties of the models developed in this work and to find the optimal parameters that fit the literature data.

Prostate Simulations

The prostate simulation was conducted using the constant marginal probability model described in Section 2.4. This model was chosen because, outside the GTV, it assumes that the marginal probability of MTP is uniform throughout the prostate tissue, meaning that it does not vary with distance from the visible lesion. This assumption was adopted because to the best of our knowledge there is no evidence in the literature suggesting a spatial correlation between invisible lesions and the GTV. Additionally, in clinical practice, the entire prostate is typically included within the CTV and irradiated uniformly outside the visible lesion. This widespread treatment approach further supports the notion that there is limited information on the spatial correlation between the GTV and microscopic tumor presence.

The simulations were performed on the anatomy of the third patient from the publicly available QIN PROSTATE dataset, which includes physician-delineated contours of the GTV and prostate [10]. The simulation space was resampled to a 1 mm resolution and defined as the volume outside the GTV but within the prostate (Fig. 8a).

The study was conducted by generating 5000 simulations for each value of $T(q, N)$, which was varied from 1×10^{-4} to 1×10^{3}, while grid resolutions were adjusted between 0.5 mm and 2 mm. The parameter P_0 was fixed to match the proportion of patients without MTP reported on Bajgira et al [19]. From these simulations, $\bar{N}_I$, $\bar{V}_I$ and $\bar{V}_T$ were computed. Each islet was defined as the independent components of the simulated MTP. Additionally, the distribution of islet volumes for different T values was analyzed, and the 3D MTP marginal probability, along with different MTP realizations, were visually examined.

Breast and Lung Simulations

The breast and lung simulations were performed using the variable marginal probability model described in Section 2.5. The breast simulations were based on the anatomy of the fifth patient from the Advanced-MRI-Breast-Lesions dataset, which contains physician-delineated contours of the GTV [9]. The lung simulations were conducted using the anatomy of the first patient from the NSCLC-Radiomics dataset, which includes physician-delineated contours of the GTV and lungs [1].

For both cancer types, the simulation space was resampled to a 2 mm resolution grid and defined as the volume outside the GTV, extending up to the distance where the probability of detecting MTP falls below 1%, while accounting for anatomical barriers (Figs. 11a and 12a).

The sampling space for both cancer types was discretized into 12 layers that were equally spaced considering $\mathrm{w}(r)$. The distance d_i to i-th shell is equal to $d_i = \mathrm{w}^{-1}(\mathrm{w}_1(13-i)/12))$, where $\mathrm{w}_1 = 1 - P_0$ is the probability of finding MTP outside the GTV. The parameter space was explored for different values of q ranging from 0 to 1. The parameters $p_1, \ldots, p_{12}$ were determined such that $1-\mu_i(q, p_1, \ldots, p_{12}) = \mathrm{w}_i$ for each value of q. This was achieved by solving the optimization problem defined in Equation 13 using Bayesian optimization as implemented in the `scikit-optimize` library [14]. At each step of the optimization, 10,000 simulations were conducted to approximate μ_i.

Using the optimal model parameters, we evaluated $\bar{N}_I$, $\bar{V}_I$, and $\bar{V}_T$. Additionally, we analyzed the radial distribution of the simulated islets relative to the GTV boundary. A visual assessment of the CTM, the isoprobability shell 0.05 derived from the simulations, and representative MTP realizations was also performed.

2.8 Implementation

The constant and variable marginal probability models were implemented in Python. Simulations of the MTP were accelerated using parallel computing techniques to handle the high computational load. In particular, we used CUDA library to distribute independent simulations across available GPU, significantly reducing runtime [*].

3 Results

In this section, we investigate the behavior of the model under different parameter settings. We first explore how the model responds to changes in key parameters, analyzing its statistical properties and consistency with theoretical expectations. We then assess the model's ability to fit empirical MTP data reported in the literature and evaluate its capacity to reproduce observed patterns.

3.1 Prostate Simulations

The simulations generated in the prostate patient showed that $\bar{N}_I$ transition with $T(q, N)$. Notably, this characteristic does not depend on the grid resolution selected for the simulation (Fig 6).

The maximum $\bar{N}_I$ obtained from the simulations occurs when voxel states are nearly uncorrelated ($T \approx 0$), aligning with the theoretical value of $\ln(1/P_0) = 1.43$. Conversely, the minimum $\bar{N}_I$ is observed when voxel states are almost fully correlated ($T \to \infty$), yielding $1 - P_0 = 0.76$. Notably, $T(q, N) = 0.2$ appears to define the midpoint of this transition, while the $\bar{N}_I$ reported in the literature closely matches our model's predictions at $T(q, N) = 0.6$.

The distribution of islet volumes for three different values of T exhibited distinct patterns (Fig 7c). When T is in the region where voxel states are nearly uncorrelated ($T \leqslant 10^{-4}$), the generated islets remain small ($\sim 100\,\mathrm{mm}^3$) as the model fails to cluster many voxels together. As T approaches the middle of the transition ($T \approx 10^{-1}$), voxel clustering increases, leading to larger islets. In this regime, a fraction of the islets can cluster to fill the entire sampling space, corresponding to the volume of the prostate. When T approximates

[*] The source code related to this project is available at `https://github.com/RivettiLuciano/MicroTumorModel`

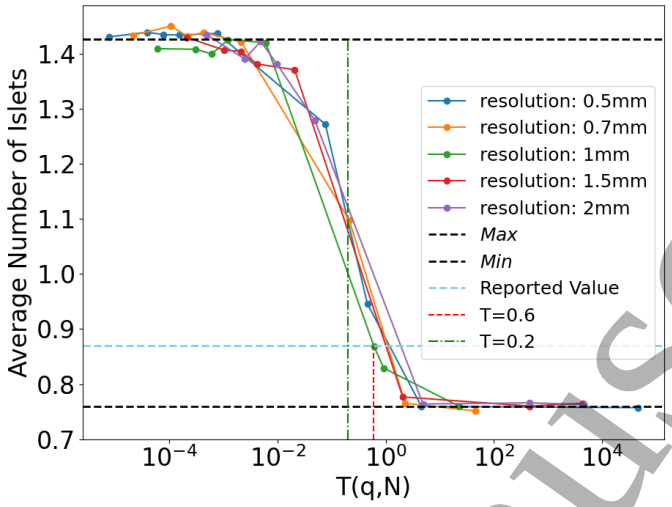


Figure 6: Average number of islets per patient ($\bar{N}_I$) for different grid resolutions and values of $T(q, N)$. The scatter plot, with different colors, represents results for various grid resolutions. The black dashed line at the top corresponds to the maximum theoretical $\bar{N}_I$ given by $\ln(\frac{1}{P_0})$, while the black dashed line at the bottom represents the minimum theretical value $1 - P_0$. The sky blue line indicates the $\bar{N}_I$ reported by Bajgira et al. [19]. The red dashed line marks the value of T that reproduces the reported $\bar{N}_I$ at a resolution of 1 mm. The green dot-dashed line represents the value of T that defines the midpoint of the transition between the two distinct tumor configuration states. The x-axis is displayed on a logarithmic scale.

the regime where voxel states are fully correlated ($T \gtrsim 10^1$), the islets predominantly cluster, filling the entire sampling space.

$\bar{V}_T$ exhibited a direct relationship with $T(q, N)$ and remained independent of grid resolution (Fig. 7d). When $T(q, N) \in [10^{-4}, 10^{-1}]$, the model produced $\bar{V}_T$ values within the expected range reported by Bajgira et al. [19]. Notably, at $T(q, N) = 0.02$, the simulations yielded an $\bar{V}_T$ of 583 mm^3, matching the literature estimate. Similar trends were observed for $\bar{V}_I$ (Appendix E).

Since $\bar{V}_T$ incorporates information from both $\bar{N}_I$ and $\bar{V}_I$, we examined tumor configurations at $T(q, N) = 0.02$. At this value, the marginal probability of MTP was $p = 0.028$ (Fig. 8a), and simulations indicated that MTP was absent in 23.8% of cases. Although the marginal probability was low, the simulations still produced cases with significant MTP volumes. Figure 8 illustrates a range of representative tumor configurations:

(b) A realization with a single islet of 429 mm^3, approximately close to the literature-reported mean V_T.

(c) A realization with a small islet of 1 mm^3, near the 5th percentile of the reported V_T distribution.

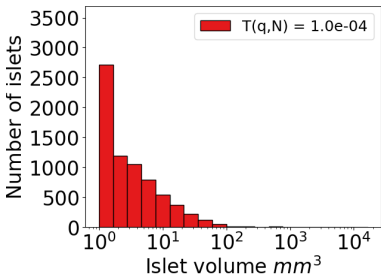
(d) A realization with an islet of 1849 mm^3, near the 95th percentile of the reported V_T distribution.

(e) An outlier case in which the tumor has infiltrated the entire prostate.

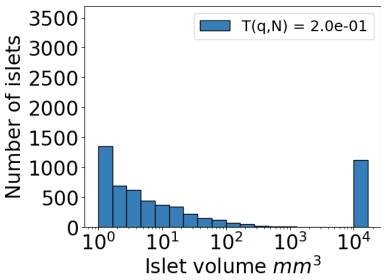
Furthermore, the simulated tumor islets did not exhibit any preferred orientation relative to the GTV boundaries (Fig. 8b–e).

3.2 Breast Simulations

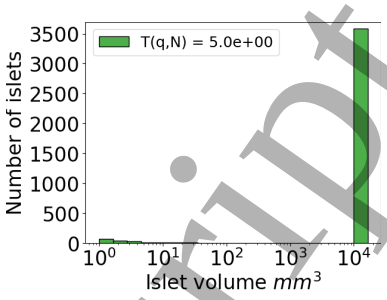
For different values of q, the optimization process determined the marginal probabilities of the layers, $\{\hat{p}_1, \ldots, \hat{p}_{12}\}$, that generated the $1 - \mu_i$ that best matched the w_i values reported in the literature (Fig. 9a).



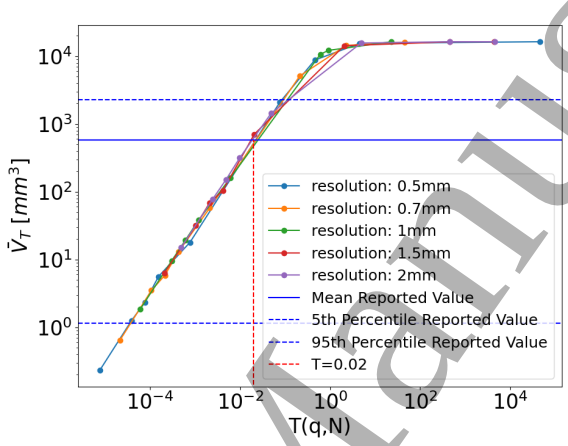
(a) Islet's volume distribution. Weak correlation



(b) Islet's volume distribution. Intermediate correlation



(c) Islet's volume distribution. Strong correlation



(d) Average total microscopic tumor volume

Figure 7: Dependence of MTP volume on $T(q, N)$. (a–c) Distributions of islet volumes for three representative T values at 1 mm resolution: (a) $T = 10^{-4}$ (weak correlation, red), (b) $T = 0.2$ (moderate correlation, blue), and (c) $T = 5$ (strong correlation, red). (d) Simulated $\bar{V}_T$ across varying grid resolutions and T values. The solid blue line denotes the $\bar{V}_T$ reported by Bajgira *et al.*, while the dashed blue lines indicate the 5th and 95th percentiles of the average total microscopic tumor volume per patient reported in [19]. All axes are shown on a logarithmic scale.

The results show that the marginal probabilities satisfy $\hat{p}_1 \geqslant \hat{p}_2 \geqslant \cdots \geqslant \hat{p}_{12}$ for all values of q except when q is approaching zero. Similar to the prostate case, the marginal probability decreases when $q \to 0$ (Fig 9b).

The simulations performed on the breast cancer patient showed that our model can not generate the $\bar{N}_I$ reported in Stroom et al. [25] for any value of q. The maximum $\bar{N}_I$ in our simulations was 1.3 and was achieved at q = 0.999. On the other hand, the estimated $\bar{V}_T$ obtained from Stroom et al. can be generated with the simulations when $q = 0.975$ (Appendix F) [25]. The relation between $\bar{V}_T$ and q follows a similar trend to the one observed in the prostate case, the microscopic tumor volume increases when the voxel states are more correlated. However, the microscopic tumor volume generated were one order of magnitude smaller when compared to the prostate case. Similar results were also observed for $\bar{V}_I$.

Since $\bar{V}_T$ encapsulates information about both $\bar{N}_I$ and $\bar{V}_I$, we analyzed the properties of the tumor configurations obtained from simulations with $q = 0.997$. These simulations revealed that the mean distance between the geometric center of the islets and the nearest GTV boundary followed a distribution similar to gaussian distribution with mean 0 mm and variance 7 mm^2 reported in Stroom et al. (Fig. 10a) [25]. The R^2 between our simulations and the function reported in the paper is 0.84 and the mean absolute error (MAE) is 0.01.

The CTM demonstrated a radial dependence on distance from the GTV border, with probabilities de-

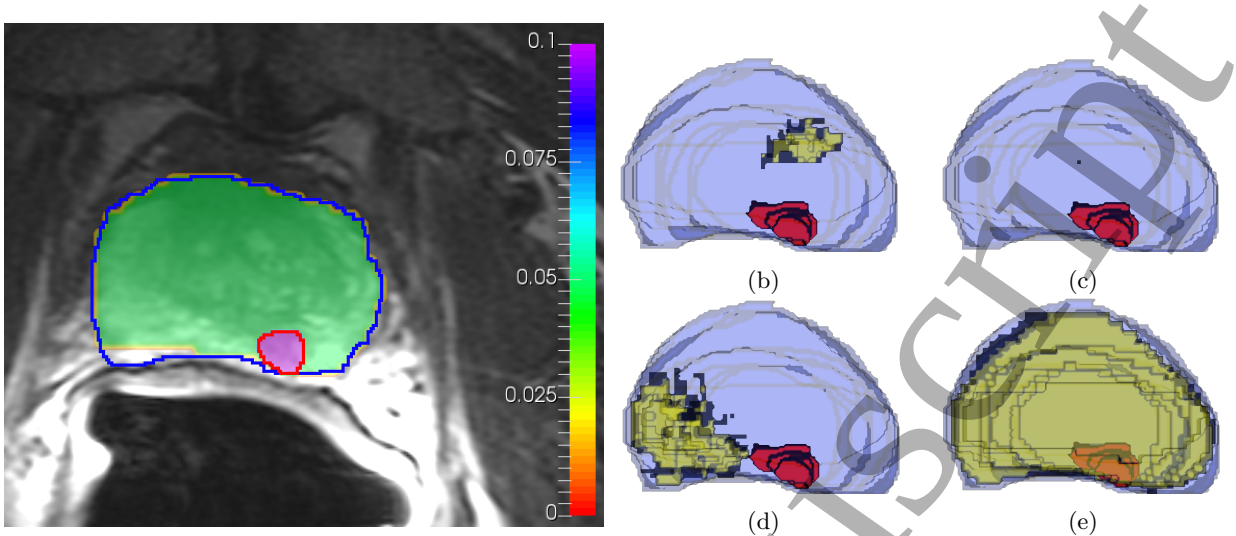
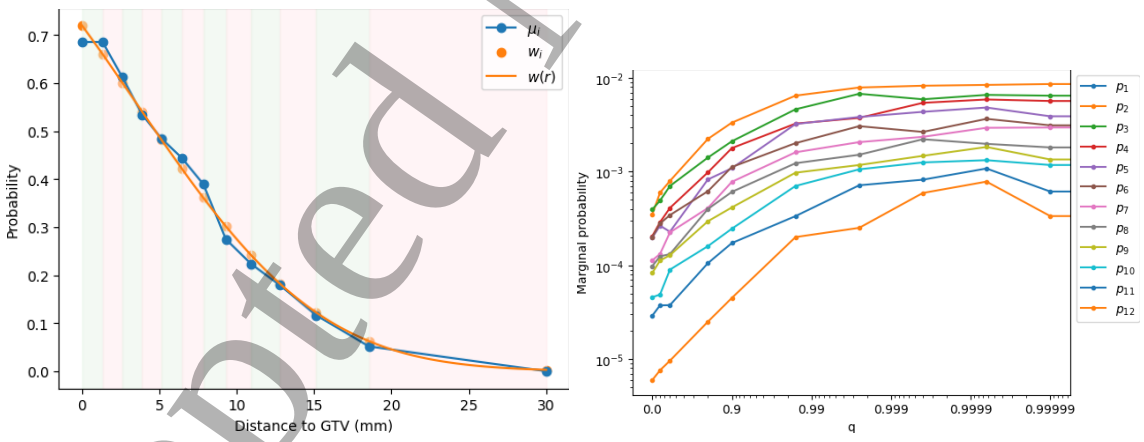


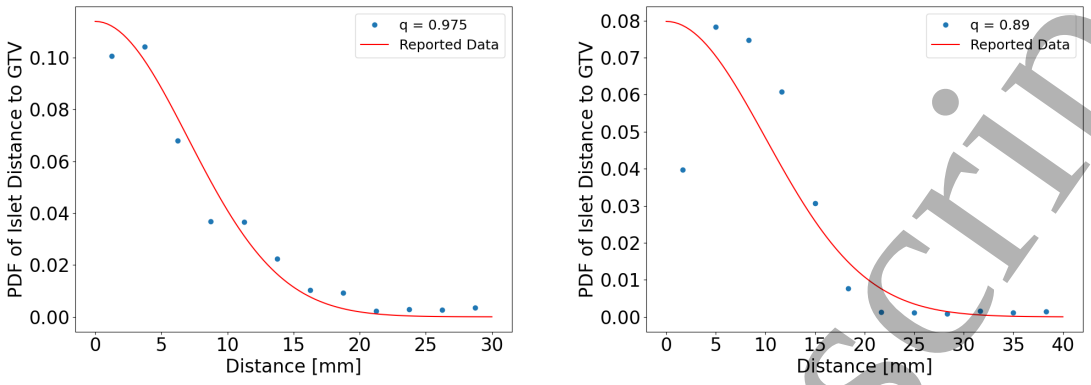
Figure 8: Microscopic tumor realizations generated for a prostate cancer patient. (a) The color map represents the CTM computed using the constant marginal probability model described in Section 2.4, overlaid on the diagnostic MRI. The blue contour delineates the anatomical barrier considered in the MTP simulation—the prostate—while the red contour outlines the GTV. (b–e) Different 3D realizations of MTP generated using a 1 mm resolution grid with $T(q, N) = 0.02$. The blue structure represents the prostate, the reddish/purple structure corresponds to the GTV, and the yellowish/greenish regions depict the microscopic tumor simulated with our model. Specifically, (b) a realization with a single islet of 429 mm^3, approximately close to the literature-reported mean V_T; (c) a realization with a small islet of 1 mm^3, near the 5th percentile of the reported V_T distribution; (d) a realization with an islet of 1849 mm^3, near the 95th percentile of the reported V_T distribution; and (e) an outlier case in which the tumor has infiltrated the entire prostate.



(a) Probability of finding MTP beyond a distance

(b) Marginal probability of layers vs q

Figure 9: Estimation of marginal probabilities by solving the optimization problem described in Equation 13. (a) A representative example ($q = 0.99999$) illustrating how well μ_i approximates w_i for the optimized marginal probabilities. Blue dots represent μ_i, while orange dots correspond to w_i. The background shading highlights the layer boundaries, with green and red indicating the extent of each layer, starting with layer 1 on the left. (b) Marginal probability estimated for each layer across different values of q.



(a) Breast cancer case (b) Lung cancer case

Figure 10: Probability density function (PDF) of the radial positions of microscopic tumor islets measured from the boundary of the GTV for breast and lung cancer cases. The scatter plots represent simulation results generated using our variable marginal probability model. The red line corresponds to the empirical function fitted by Stroom et al. based on their observations of the radial distribution of microscopic tumor islets [25].

creasing for voxels located farther from the target (Fig. 11a). The maximum marginal probability in the CTM was 0.007, occurring in the first layer. Using μ, we determined the distance beyond which the probability of finding any tumor voxel drops below 0.05, which was 20 mm. The isotropic expansion of the GTV with this distance represents the isoprobability shell 0.05. It can be noticed that this isoprobaiblity shell is smaller that than CTM.

Because the marginal probability decreases with increasing distance from the GTV, the simulations indicated that tumor islets preferentially clustered near the GTV boundary. Figure 11b–e presents representative tumor configurations:

(b) A realization with a volume approximately close to the mean of the V_T distribution, with $V_T = 162\,\text{mm}^3$.

(c) A realization near the 5th percentile of the V_T distribution, with $V_T = 8\,\text{mm}^3$.

(d) A realization near the 95th percentile of the V_T distribution, with $V_T = 552\,\text{mm}^3$.

(e) An outlier case with an exceptionally large tumor volume ($V_T = 1504\,\text{mm}^3$).

The simulations also showed that MTP was absent in 28.7% of the cases, and that the tumor islets frequently exhibited irregular, non-spherical shapes. In some realizations, streak-like MTP patterns were observed, similar to those illustrated in Figure 11d. Moreover, no tested value of q resulted in an MTP distribution that fully occupied the sampling domain.

3.3 Lung simulations

Similar to the breast cancer patient, the marginal probabilities across layers satisfy $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_{12}$ when voxel correlations are present, with probabilities decreasing as q decreases. It was also observed that the model was unable to replicate the $\bar{N}_I$ reported by Stroom et al. [25]. While their study reported an $\bar{N}_I$ of 2.47, the highest value obtained in our simulations was only 0.70.

Nevertheless, when setting $q = 0.89$, the simulated $\bar{V}_T$ matched the estimate from Stroom et al. (Appendix G) [25]. Under this condition, 10,000 simulations showed that the distribution of islets as a function of their

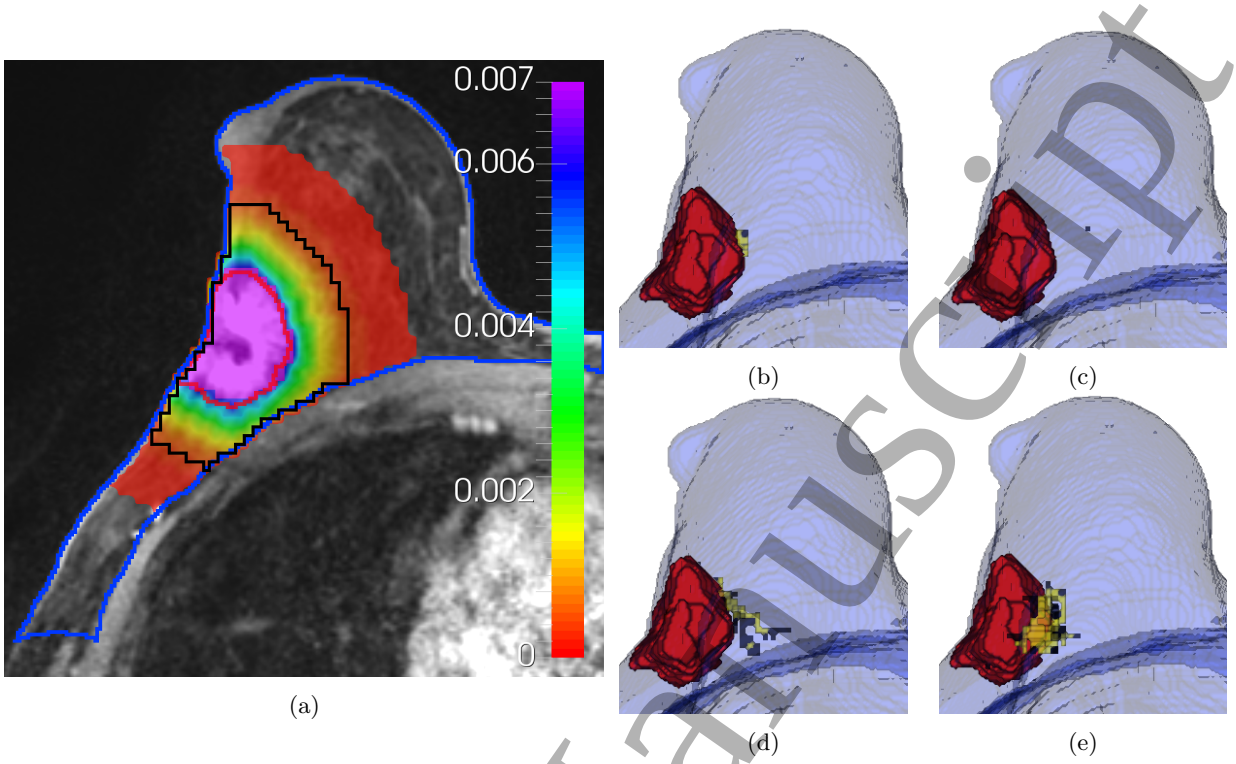


Figure 11: CTM and microscopic tumor realizations generated for a breast cancer patient. (a) The color map displays the MTP marginal probability calculated using the variable marginal probability model described in Section 2.5, overlaid on the diagnostic MRI. The blue contour represents the simulation space, accounting for anatomical barriers, the red contour outlines the GTV, and the black contour outlines the isoprobability shell 0.05. (b–e) Different 3D realizations of MTP generated using a 2 mm resolution grid and a setting of $q = 0.975$. The blue structure represents the sampling space, the reddish/purple structure corresponds to the GTV, and the yellowish/greenish regions depict the microscopic tumor realizations simulated with our model. Specifically, (b) a realization with a volume approximately close to the mean of the V_T distribution, with $V_T = 162\,\text{mm}^3$; (c) a single voxel corresponding to a realization near the 5th percentile of the V_T distribution, with $V_T = 8\,\text{mm}^3$; (d) a realization near the 95th percentile of the V_T distribution, with $V_T = 552\,\text{mm}^3$; and (e) an outlier case with an exceptionally large tumor volume ($V_T = 1504\,\text{mm}^3$).

distance from the GTV followed a trend similar to that reported in their study. However, the agreement was less strong than in the breast cancer case (Fig. 10b), as indicated by an R^2 of 0.63 and a MAE of 0.011 between our simulations and the function used in the paper.

The CTM reached a maximum marginal probability of 0.00014 and gradually decreased beyond the GTV boundary, with several layers exhibiting similar marginal probabilities (Fig. 12a). Despite the low marginal probabilities, the 0.05 isoprobability shell derived from the simulations extended 17 mm beyond the GTV boundary. The simulations also revealed the presence of small islets clustered near the GTV border. Figure 12b–e shows representative tumor configurations:

(b) A realization with a volume approximately close to the mean of the V_T distribution, with $V_T = 82\,\text{mm}^3$.

(c) A realization near the 5th percentile of the V_T distribution, with $V_T = 8\,\text{mm}^3$.

(d) A realization near the 95th percentile of the V_T distribution, with $V_T = 320\,\text{mm}^3$.

(e) An outlier case with an exceptionally large tumor volume ($V_T = 6616\,\text{mm}^3$).

The simulated islet volumes were consistently smaller than the GTV, and their shapes often displayed irregular, non-spherical shapes. In some realizations, streak-like MTP patterns were observed, similar to those illustrated in Figure 12e.

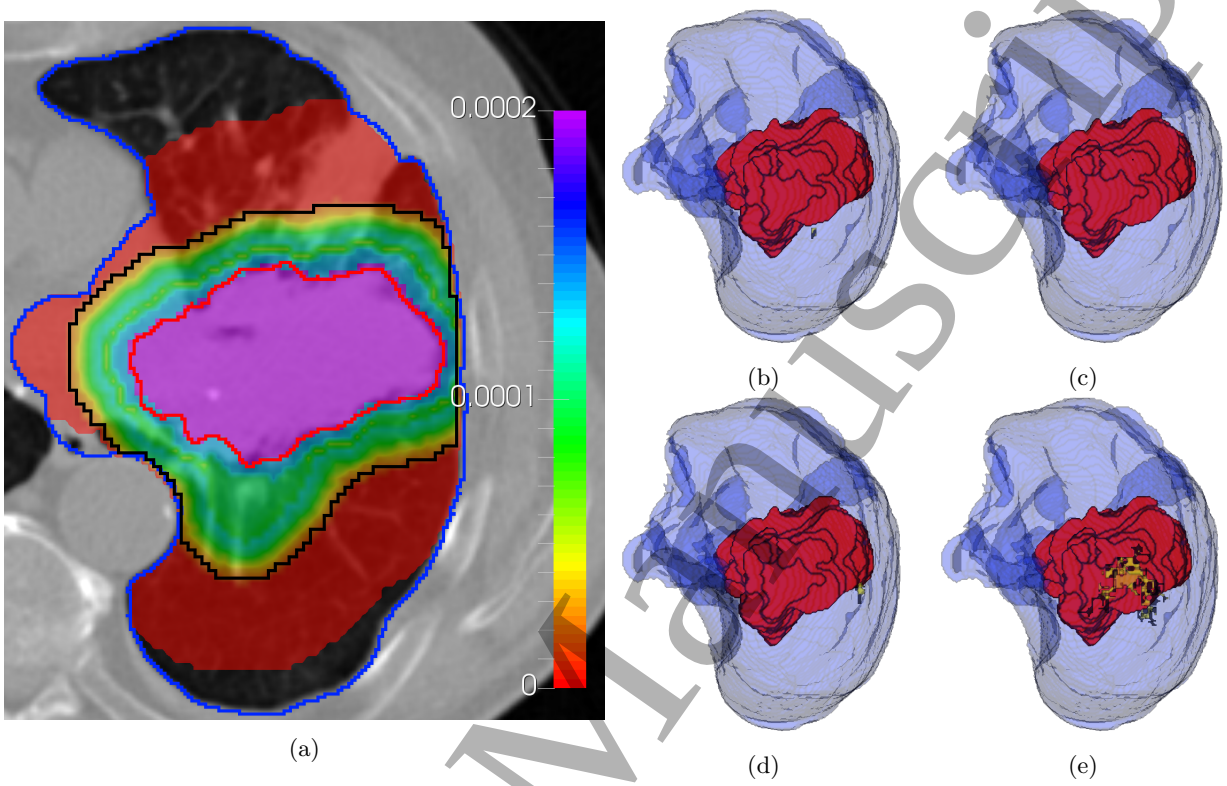


Figure 12: CTM and microscopic tumor realizations generated for a lung cancer patient. (a) The color map displays the MTP marginal probability calculated using the variable marginal probability model described in Section 2.5, overlaid on the diagnostic CT. The blue contour represents the simulation space, accounting for anatomical barriers, the red contour outlines the GTV, and the black contour outlines the isoprobability shell 0.05 generated from the simulations. (b–e) Different 3D realizations of MTP generated using a 2 mm resolution grid and a setting of $q = 0.89$. The blue structure represents the sampling space, the reddish/purple structure corresponds to the GTV, and the yellowish/greenish regions depict the microscopic tumor realizations simulated with our model. Specifically, (b) a realization with a volume approximately close to the mean of the V_T distribution ($V_T = 82\,\mathrm{mm}^3$); (c) a single voxel that corresponds to a realization near the 5th percentile of the V_T distribution ($V_T = 8\,\mathrm{mm}^3$); (d) a realization near the 95th percentile of the V_T distribution ($V_T = 320\,\mathrm{mm}^3$); and (e) an outlier case with an exceptionally large tumor volume ($V_T = 6616\,\mathrm{mm}^3$).

4 Discussion

The questions raised in the ICRU 83 report, as discussed in the introduction, highlight the necessity of clear statistical definitions and practical methods for incorporating probabilities into target definition and radiotherapy treatment planning [16]. In particular, the first question—how to define the "probability of occurrence" of tumor infiltration in a given tissue from a statistical perspective—can be directly addressed using the stochastic model presented in this work.

Our model provides a rigorous framework to estimate the probability of specific tumor configurations by computing the joint probability distribution or by employing Monte Carlo simulations. Through these simulations, various realizations of microscopic tumor presence (MTP) can be generated, and the probability of tumor presence in a given volume can be estimated as the fraction of simulations in which MTP occurs within that region.

The second question—on what clinical evidence these probabilities should be based—remains a challenge. In this study, available data consisted of population-based analyses that provide averages and standard deviations of microscopic tumor characteristics, such as islet radii, the percentage of patients with MTP, and the number of islets [28, 22, 10, 25]. While this data allow for the calibration of our stochastic model, more detailed information, including tumor staging and three-dimensional histopathological reconstructions, could enable more precise calibration. In particular, 3D histological reconstructions of MTP samples could be used to optimize the model parameters through maximum likelihood estimation, providing a more robust statistical foundation for our approach.

Finally, the third question—how to translate probabilities assigned to tissues into discretized grids for treatment planning—has been partially addressed in the literature. For instance, Bortfeld et al. demonstrated how microscopic tumor probabilities can be integrated with TCP in the treatment planning process [5]. Although their approach is computationally intensive, even for simplified models, it demonstrates the feasibility of incorporating probabilistic information into treatment optimization. However, to date, all models that incorporate probability during optimization rely solely on the clinical target map (CTM), neglecting voxel correlation. These models incorporated target uncertainty into treatment planning by using CTMs as weighting factors in the optimization loss function [3, 7, 24], this approach may result in underdosing regions where the actual risk of MTP exceeds 0.05. For instance, applying this method with the CTMs generated in our simulations would likely underdose the entire region outside the GTV, since the marginal voxel-level probabilities were all below 0.028 in the three cancer patients. However, the isoprobability surface defined at 0.05 corresponds to a much larger combined probability—where the true probability of microscopic tumor presence can reach values as high as 0.95. Consequently, relying only on marginal probabilities would lead to insufficient dose coverage. This underscores the importance of accounting for voxel correlations to ensure adequate target coverage during treatment optimization. It is also worth noting that, despite the relatively small marginal probabilities predicted by our models, the isoprobability shell at 0.05 still extends a considerable distance from the GTV border in the breast and lung cases. This occurs because the shell reflects the probability of detecting at least one tumor cell outside of the entire volume it encloses, rather than the probability at any single voxel.

Building on this motivation, a further way in which the probabilities generated by our stochastic model can support treatment planning is through the definition of non–isotropic CTVs in Euclidean space. If the CTV is defined as the volume that encompasses a given percentage of MTP realizations, our approach enables the exploration of all CTV shapes that satisfy this criterion by analyzing a large ensemble of simulations. Although these candidate CTVs correspond to isotropic expansions in the MTP probability space, their spatial manifestations in Euclidean space can become non–isotropic. This feature could be exploited in situations where conventional margin approaches, which typically rely on binary targets, are desired but the proximity of OARs to the GTV creates trade-offs between target coverage and OAR sparing. In such cases, one would prefer CTVs that carve out the OAR for preferential sparing, while extending farther in directions where compromise is less severe. We note that similar anisotropic solutions naturally arise in the dose space under probabilistic optimization, where the optimizer concentrates dose in regions with higher MTP probability while sacrificing other small, lower probability regions, in order to achieve the OAR constraints [6, 5].

The models developed in this work provide a framework for capturing voxel correlations while generating probabilistic results consistent with clinical observations reported in the literature. In the constant marginal probability model, this consistency is enforced by imposing a hard constraint that the parameter P_0 matches the observed probability of detecting MTP in a patient. In the variable marginal probability model, we ensure consistency by requiring that one minus the probability of all voxels outside the i-th shell being tumor-free, $1 - \mu_i$, corresponds to the probability of detecting MTP beyond the i-th shell, denoted as w_i. Moreover, our models explain how variations in voxel correlation influence the selection of the marginal probabilities assigned to individual voxels, thereby affecting the likelihood of different tumor configuration states. This behavior could have important implications for treatment planning. By enabling the calculation of joint probabilities, our models offer a principled way to better guide the target–OARs compromise through the assessment of the probability of finding MTP near an OAR. They can also support more accurate probabilistic robust evaluations by allowing the sampling of spatially realistic tumor configurations.

In addition to these practical applications, the models offer structural advantages that further enhance their usability and relevance. One of the most significant benefits is the limited number of tunable parameters, each grounded in sound statistical reasoning. This structure facilitates straightforward calibration using clinical data. The constant marginal probability model, for example, is defined by only two parameters, P_0 and T. The parameter P_0 is directly linked to the proportion of patients without MTP, allowing for immediate initialization based on available data. This, in turn, simplifies the tuning of T, as it remains the only free parameter in the model. A similar advantage is observed in the variable marginal probability model, where the probability of tumor presence beyond a given distance from the GTV, $\mathrm{w}(r)$, is used to determine $p_1, \ldots, p_{N'}$. The remaining parameter, q, is then calibrated to align with clinical observations, maintaining the model's adaptability while keeping it interpretable. The calibrated parameters that resulted from our study were $T = 0.02$, $q = 0.997$, and $q = 0.89$ for the prostate, breast, and lung cancer cases, respectively. Notably, these values lie well within the admissible parameter ranges, which explains why the resulting realizations display intermediate clustering behavior—neither isolated voxels nor entirely solid regions of MTP.

The parameter T reported in our study is not patient-specific, as it is defined independently of the total number of voxels in the prostate and calibrated to reproduce population-based MTP metrics. In contrast, the variable marginal probability model requires patient-specific calibration of q, since the probability of finding MTP in a given layer depends on the number of voxels within that layer. As detailed in the Methods section, T and q regulate the extent to which tumor voxels cluster together. From a biological perspective, these parameters can be interpreted as indicators of tumor aggressiveness, as they determine both the number and size of microscopic tumor islets (Figures 6 and 7b).

Our models naturally converge to previously established MTP models under specific conditions. The variable marginal probability model described in this work reduces to the model presented by Shusharina et al. as $q \to 0$ [23]. In the 1D case, it also converges to the model proposed by Buti et al. as $q \to 1$ [7], since Buti's model is fundamentally one-dimensional, assuming isotropy in the radial direction. Unlike these earlier models, our approach can also capture intermediate regimes, where voxels are neither fully independent nor fully correlated, i.e., a mixed correlation state. This was observed as the realizations in all the three cancer cases were clusters of many voxels of MTP rather than a single voxel. Beyond parameter calibration, the simplicity of our model's growing pattern also facilitates the computation of the joint probability of different tumor configurations. As outlined in the Methods section, the joint probability of a specific tumor configuration can be determined using Bayes' theorem, provided that the starting growing voxel is known.

While a model with few parameters offers certain advantages, it also has inherent limitations, particularly its reduced flexibility in simultaneously capturing multiple microscopic tumor properties. This is evident in the three patient cases: for prostate simulations, $T = 0.6$ generated realizations that matched the reported $\bar{N}_I$ but not $\bar{V}_T$, whereas $T = 0.02$ matched $\bar{V}_T$ but not the $\bar{N}_I$. Similar discrepancies were observed in the breast and lung simulations. These examples illustrate that, although the model provides a reasonable first-order approximation, it cannot perfectly reproduce all tumor metrics simultaneously. Introducing additional parameters could potentially improve the simultaneous fitting of different microscopic tumor features; however, the limited availability of reliable MTP data from histopathology studies poses significant challenges

for calibrating more complex, multi-parameter models.

Another limitation of our approach lies in its dependence on the specific sampling algorithm illustrated in Figure 2. The algorithm tends to favor growth along the principal axes (x, y, z) originating from the randomly selected seed voxel. This can lead to the appearance of "streaks" structures in some realizations (as observed, for example, in Figs. 11d & 12e), resulting from stronger voxel-to-voxel Pearson correlations along the grid axes compared to off-axis directions. Consequently, alternative sampling strategies, such as incorporating second-order neighborhoods, adopting non-axial growth geometries, or varying the correlation parameter across voxels, could produce realizations with different spatial patterns or morphologies.

It is important to note that our model was primarily designed to reproduce population-level characteristics of microscopic tumor presence reported in the literature, such as the total microscopic tumor volume and the number of tumor islets, both of which are independent of the specific tumor shape. Generating more anatomically realistic spatial patterns would require voxel-level tumor infiltration data from histopathology to more accurately calibrate spatial correlations. However, such detailed data are currently very difficult to obtain and were not available for this study.

In comparison with the work of Stroom et al., who proposed a CTV margin calculation formula based on the standard deviation of the islet distribution and a correction factor derived from radiotherapy simulations, our approach provides an alternative, data-driven perspective. Although we do not introduce an explicit analytical expression for margin estimation, the CTV margin naturally emerges from our model as the isotropic expansion of the GTV that encloses 95% of the simulated MTP realizations. Applying this approach yields CTV margins of 20 mm for breast cancer and 17 mm for lung cancer, which are in the order of magnitude to the margins predicted by the Stroom formula (15.4 mm and 22.6 mm, respectively). The remaining differences of approximately 5 mm can reasonably be attributed to the distinct underlying assumptions of the two methods.

Future work will focus on integrating both the CTM and voxel correlation information into the treatment planning process to enable probabilistically robust treatment strategies. Additionally, we aim to investigate more physically grounded models, such as the Ising model and its variants, to further refine the representation of microscopic tumor spread and improve the accuracy of treatment optimization [26, 2].

5 Conclusion

In this study, we developed two first-principles stochastic models to simulate microscopic tumor presence (MTP) under different assumptions, incorporating a correlation factor to account for spatial dependencies between neighboring voxels. This approach enables a more realistic representation of tumor infiltration patterns compared to existing methods. By leveraging these models, we can estimate both the marginal probability of MTP and the joint probability of different tumor configurations, providing a complete probabilistic description of MTP. Our simulations reproduced key features of microscopic tumor extension reported for prostate, breast, and lung cancers, effectively capturing their distinct spatial patterns.

Importantly, our findings highlight the limitations of using marginal probabilities alone as weights in treatment optimization and emphasize the necessity of incorporating voxel correlations to properly account for spatial uncertainty. This work lays the foundation for integrating probabilistic modeling into the description of MTP, with potential implications for improving radiotherapy treatment planning.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Actions under Grant Agreement No. 955956. The work was also supported by the National Cancer Institute of the United States under grant number R01CA266275, and by the Therapy Imaging Program (TIP) funded by the Federal Share of program income earned by Massachusetts General Hospital on C06CA059267, Proton Therapy Research and Treatment Center.

References

[1] Hugo J. W. L. Aerts et al. *Data From NSCLC-Radiomics*. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI. Version 4. 2014.

[2] L. Amoudruz, G. Buti, L. Rivetti, A. Ajdari, G. Sharp, P. Koumoutsakos, S. Spohn, A. L. Grosu, and T. Bortfeld. "Ising energy model for the stochastic prediction of tumor islets". In: *arXiv preprint arXiv:2508.20804v2* (Sept. 2025). PMID: 41001577; PMCID: PMC12458594.

[3] Ivar Bengtsson, Anders Forsgren, and Albin Fredriksson. "Implications of using the clinical target distribution as voxel-weights in radiation therapy optimization". In: *Physics in Medicine & Biology* 68.9 (Apr. 2023), p. 095005. DOI: 10.1088/1361-6560/acc77b. URL: https://dx.doi.org/10.1088/1361-6560/acc77b.

[4] Kamalika Bhattacharjee, Nazma Naskar, Souvik Roy, and Sukanta Das. "A survey of cellular automata: types, dynamics, non-uniformity and applications". In: *Natural Computing* 19.2 (2020), pp. 433–461. ISSN: 1572-9796. DOI: 10.1007/s11047-018-9696-8. URL: https://doi.org/10.1007/s11047-018-9696-8.

[5] Thomas Bortfeld, Nadya Shusharina, and David Craft. "Probabilistic definition of the clinical target volume—implications for tumor control probability modeling and optimization". In: *Physics in Medicine & Biology* 66.1 (Jan. 2021), 01NT01. DOI: 10.1088/1361-6560/abcad8. URL: https://dx.doi.org/10.1088/1361-6560/abcad8.

[6] Gregory Buti, Nadya Shusharina, Ali Ajdari, Edmond Sterpin, and Thomas Bortfeld. "Exploring trade-offs in treatment planning for brain tumor cases with a probabilistic definition of the clinical target volume". In: *Medical Physics* 50.1 (2023), pp. 410–423. DOI: https://doi.org/10.1002/mp.16097. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16097. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16097.

[7] Gregory Buti, Kevin Souris, Ana Maria Barragán Montero, John Aldo Lee, and Edmond Sterpin. "Introducing a probabilistic definition of the target in a robust treatment planning framework". In: *Physics in Medicine & Biology* 66.15 (July 2021), p. 155008. DOI: 10.1088/1361-6560/ac1265. URL: https://dx.doi.org/10.1088/1361-6560/ac1265.

[8] Sorcha Campbell et al. "Evaluation of Microscopic Disease in Oral Tongue Cancer Using Whole-Mount Histopathologic Techniques: Implications for the Management of Head-and-Neck Cancers". In: *International Journal of Radiation Oncology*Biology*Physics* 82.2 (2012), pp. 574–581. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2010.09.038. URL: https://www.sciencedirect.com/science/article/pii/S036030161003347X.

[9] Danna Daniels, Dorit Last, Katya Cohen, Yael Mardor, and Michal Sklair-Levy. *Standard and Delayed Contrast-Enhanced MRI of Malignant and Benign Breast Lesions with Histological and Clinical Supporting Data (Advanced-MRI-Breast-Lesions)*. https://doi.org/10.7937/C7X1-YN57. Version 2. 2024.

[10] Andriy Fedorov, Clare Tempany, Robert Mulkern, and Fiona Fennessy. *Data From QIN PROSTATE*. https://doi.org/10.7937/K9/TCIA.2016.fADs26kG. 2016. DOI: 10.7937/K9/TCIA.2016.fADs26kG.

[11] Claudio Fiorino, Robert Jeraj, Catharine H. Clark, Cristina Garibaldi, Dietmar Georg, Ludvig Muren, Wouter van Elmpt, Thomas Bortfeld, and Nuria Jornet. "Grand challenges for medical physics in radiation oncology". In: *Radiotherapy and Oncology* 153 (Dec. 2020), pp. 7–14. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2020.10.001. URL: https://doi.org/10.1016/j.radonc.2020.10.001.

[12] B. Fleury, J. Thariat, R. Barnoud, G. Buiret, F. Lebreton, B. Bancel, M. Poupart, and M. Devouassoux-Shisheboran. "Approche anatomopathologique de l'extension microscopique des carcinomes épidermoïdes ORL : implications pour la définition du volume cible anatomoclinique". In: *Cancer/Radiothérapie* 18.7 (2014), pp. 666–671. ISSN: 1278-3218. DOI: https://doi.org/10.1016/j.canrad.2014.04.006. URL: https://www.sciencedirect.com/science/article/pii/S1278321814000742.

[13] Vincent Grégoire et al. "Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines". In: *Radiotherapy and Oncology* 126.1 (Jan. 2018), pp. 3–24. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2017.10.016. URL: http://dx.doi.org/10.1016/j.radonc.2017.10.016.

[14] Tim Head, Manoj Kumar, Holger Nahrstaedt, Gilles Louppe, and Iaroslav Shcherbatyi. *scikit-optimize/scikit-optimize: v0.8.1*. Version v0.8.1. Sept. 2020. DOI: 10.5281/zenodo.4014775. URL: https://doi.org/10.5281/zenodo.4014775.

[15] Petra J. van Houdt et al. "Histopathological Features of MRI-Invisible Regions of Prostate Cancer Lesions". In: *Journal of Magnetic Resonance Imaging* 51.4 (2020), pp. 1235–1246. DOI: https://doi.org/10.1002/jmri.26933. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jmri.26933. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.26933.

[16] International Commission on Radiation Units and Measurements. *Prescribing, Recording, and Reporting Intensity-Modulated Photon-Beam Therapy (IMRT)*. Tech. rep. Report 83. ICRU, 2010. URL: https://www.icru.org/report/prescribing-recording-and-reporting-intensity-modulated-photon-beam-therapy-imrticru-report-83/.

[17] Xue Meng, Xindong Sun, Dianbin Mu, Ligang Xing, Li Ma, Baijiang Zhang, Shuqiang Zhao, Guoren Yang, Feng-Ming (Spring) Kong, and Jinming Yu. "Noninvasive Evaluation of Microscopic Tumor Extensions Using Standardized Uptake Value and Metabolic Tumor Volume in Non-Small-Cell Lung Cancer". In: *International Journal of Radiation Oncology*Biology*Physics* 82.2 (Feb. 2012), pp. 960–966. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2010.10.064. URL: http://dx.doi.org/10.1016/j.ijrobp.2010.10.064.

[18] Leyla Moghaddasi, Eva Bezak, and Loredana G. Marcu. "Current challenges in clinical target volume definition: Tumour margins and microscopic extensions". In: *Acta Oncologica* 51.8 (Sept. 2012), pp. 984–995. ISSN: 1651-226X. DOI: 10.3109/0284186x.2012.720381. URL: http://dx.doi.org/10.3109/0284186X.2012.720381.

[19] Ali Mohammadian Bajgiran, Sara Afshari Mirak, Shirin Shakeri, et al. "Characteristics of missed prostate cancer lesions on 3T multiparametric-MRI in 518 patients: based on PI-RADSv2 and using whole-mount histopathology reference". In: *Abdominal Radiology* 44.3 (2019), pp. 1052–1061. DOI: 10.1007/s00261-018-1823-6.

[20] Christopher F. Njeh. "Tumor delineation: The weakest link in the search for accuracy in radiotherapy". In: *Journal of Medical Physics / Association of Medical Physicists of India* 33 (2008), pp. 136–140. URL: https://api.semanticscholar.org/CorpusID:5262423.

[21] Nina N. Sanford et al. "Individualization of Clinical Target Volume Delineation Based on Stepwise Spread of Nasopharyngeal Carcinoma: Outcome of More Than a Decade of Clinical Experience". In: *International Journal of Radiation Oncology, Biology, Physics* 103.3 (Mar. 2019), pp. 654–668. ISSN: 0360-3016. DOI: 10.1016/j.ijrobp.2018.10.006. URL: https://doi.org/10.1016/j.ijrobp.2018.10.006.

[22] Annemarie C. Schmitz, Maurice A.A.J. van den Bosch, Claudette E. Loo, Willem P.Th.M. Mali, Harry Bartelink, Maria Gertenbach, Roland Holland, Johannes L. Peterse, Emiel J.Th. Rutgers, and Kenneth G. Gilhuijs. "Precise correlation between MRI and histopathology – Exploring treatment margins for MRI-guided localized breast cancer therapy". In: *Radiotherapy and Oncology* 97.2 (2010), pp. 225–232. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2010.07.025. URL: https://www.sciencedirect.com/science/article/pii/S0167814010004780.

[23] Nadya Shusharina, David Craft, Yen-Lin Chen, Helen Shih, and Thomas Bortfeld. "The clinical target distribution: a probabilistic alternative to the clinical target volume". In: 63.15 (July 2018), p. 155001. DOI: 10.1088/1361-6560/aacfb4. URL: https://dx.doi.org/10.1088/1361-6560/aacfb4.

[24] A Smolders, I Bengtsson, A Forsgren, A Lomax, D C Weber, A Fredriksson, and F Albertini. "Robust optimization strategies for contour uncertainties in online adaptive radiation therapy". In: *Physics in Medicine & Biology* 69.16 (July 2024), p. 165001. DOI: 10.1088/1361-6560/ad6526. URL: https://dx.doi.org/10.1088/1361-6560/ad6526.

[25] Joep Stroom, Kenneth Gilhuijs, Sandra Vieira, Wei Chen, Javier Salguero, Elizabeth Moser, and Jan-Jakob Sonke. "Combined Recipe for Clinical Target Volume and Planning Target Volume Margins". In: *International Journal of Radiation Oncology*Biology*Physics* 88.3 (2014), pp. 708–714. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2013.08.028. URL: https://www.sciencedirect.com/science/article/pii/S0360301613030344.

[26] Salvatore Torquato. "Toward an Ising model of cancer and beyond". In: *Physical Biology* 8.1 (Feb. 2011). Epub 2011 Feb 7, p. 015017. DOI: 10.1088/1478-3975/8/1/015017.

[27] Jan Unkelbach et al. "The role of computational methods for automating and improving clinical target volume definition". In: *Radiotherapy and Oncology* 153 (Dec. 2020), pp. 15–25. ISSN: 0167-8140. DOI: 10.1016/j.radonc.2020.10.002. URL: https://doi.org/10.1016/j.radonc.2020.10.002.

[28] Judith van Loon et al. "Microscopic Disease Extension in Three Dimensions for Non–Small-Cell Lung Cancer: Development of a Prediction Model Using Pathology-Validated Positron Emission Tomography and Computed Tomography Features". In: *International Journal of Radiation Oncology*Biology*Physics* 82.1 (2012), pp. 448–456. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2010.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0360301610032384.

[29] Shalini K. Vinod, Michael G. Jameson, Myo Min, and Lois C. Holloway. "Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies". In: *Radiotherapy and Oncology* 121.2 (2016), pp. 169–179. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2016.09.009. URL: https://www.sciencedirect.com/science/article/pii/S0167814016343316.

[30] E. Weiss and C. Hess. "The Impact of Gross Tumor Volume (GTV) and Clinical Target Volume (CTV) Definition on the Total Accuracy in Radiotherapy". In: *Strahlentherapie und Onkologie* 179 (2003), pp. 21–30. DOI: 10.1007/s00066-003-0976-5. URL: https://doi.org/10.1007/s00066-003-0976-5.

[31] Constantinos Zamboglou et al. "Uncovering the invisible—prevalence, characteristics, and radiomics feature–based detection of visually undetectable intraprostatic tumor lesions in 68GaPSMA-11 PET images of patients with primary prostate cancer". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48 (2020), pp. 1987–1997. URL: https://api.semanticscholar.org/CorpusID:227034057.

A Appendix: Relation of Pearson Correlation Coefficient with Parameter q

Given a 2D grid with 9 elements, where each voxel has the same marginal probability p (Fig 13), the Pearson correlation coefficient between two neighbor voxels, C_8 and C_9, is defined as:

$$\rho_{C_8,C_9} = \frac{\mathbb{E}[C_8C_9] - \mathbb{E}[C_8]\mathbb{E}[C_9]}{\sqrt{\mathbb{E}[C_8^2] - (\mathbb{E}[C_8])^2}\sqrt{\mathbb{E}[C_9^2] - (\mathbb{E}[C_9])^2}}. \tag{14}$$

Because the state of each voxel is Bernoulli distributed, $\mathbb{E}[C_8] = \mathbb{E}[C_8^2] = \mathbb{E}[C_9] = \mathbb{E}[C_9^2] = p$. Therefore,

$$\rho_{C_8,C_9} = \frac{\mathbb{E}[C_8C_9] - p^2}{p - p^2}. \tag{15}$$

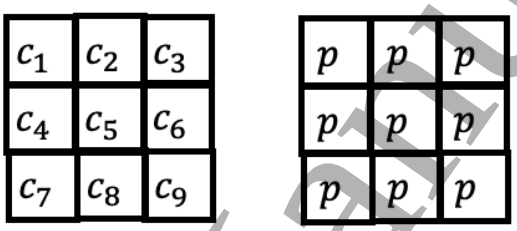


Figure 13: 2D 9-voxel grid with a uniform marginal probability distribution. The left image labels each voxel with its corresponding random variable, while the right image illustrates the marginal probability assigned to each voxel.

The key challenge is determining $\mathbb{E}[C_8C_9]$. Since our method involves randomly selecting a voxel in the grid as the initial growth point, the expectation is computed by averaging over all possible starting voxels. Let $\mathbb{E}_i[C_8C_9]$ denote the expected value given that the growth starts at voxel i. Then, the overall expectation is:

$$\mathbb{E}[C_8C_9] = \frac{1}{9}\sum_{i=1}^{9}\mathbb{E}_i[C_8C_9] = \frac{1}{9}\sum_{i=1}^{9}\sum_{\text{all states}} c_8c_9\mathbb{P}_i(\text{state}), \tag{16}$$

where $P_i(\text{state})$ represents the joint probability of a particular state given that voxel i was the starting point. By introducing equation (16) in equation (15), we obtain:

$$\rho_{C_8,C_9} = \frac{1}{9}\sum_{i=1}^{9}(\rho_{C_8,C_9})_i, \tag{17}$$

where $(\rho_{C_8,C_9})_i$ is the Pearson correlation coefficient between C_8 and C_9, considering voxel i as the starting point. Each $(\rho_{C_8,C_9})_i$ can be expressed as a weighted sum of different powers of the correlation factor q. For instance:

$$
\begin{aligned}
(\rho_{C_8,C_9})_7 &= q, \\
(\rho_{C_8,C_9})_5 &= \frac{1}{2}(q + q^3), \\
(\rho_{C_8,C_9})_4 &= \frac{1}{2}\left(q + \frac{1}{2}q^3 + \frac{1}{2}q^5\right), \\
(\rho_{C_8,C_9})_1 &= \frac{1}{2}\left(q + \frac{1}{3}q^3 + \frac{1}{3}q^5 + \frac{1}{3}q^7\right).
\end{aligned}
$$

A key observation is that the closer the starting voxel is to the voxels whose correlation is being assessed, the smaller the exponent of q. This behavior arises from the design of the visiting algorithm. Additionally, in each case, the coefficients multiplying q sum to 1, ensuring that when $q = 1$, the Pearson correlation coefficient equals 1, as expected.

Although this example is presented in 2D, the same visiting rules apply in 3D, making these results generalizable. Then, we can express the correlation coefficient between two neighboring voxels as:

$$\rho = \sum_{n=1}^{\infty} a_n q^{2n-1}, \tag{18}$$

where a_n are coefficients that satisfy:

$$\sum_{n=1}^{\infty} a_n = 1. \tag{19}$$

Note that the magnitude of a_n will depend on the size of the grid and the pair of voxels that we want to assess the correlation.

B Appendix: Taylor Series Approximation of Equation 6

To approximate Equation 6 using a Taylor series, we first rewrite it as:

$$\left(\frac{P_0}{1-p}\right)^{\frac{1}{N-1}} = 1 - p + pq. \tag{20}$$

Assuming that the number of voxels N is large and thus $x = \frac{1}{N-1}$ is small, we expand the term $\left(\frac{P_0}{1-p}\right)^x$ in a Taylor series of x around $x = 0$:

$$\sum_{n=0}^{\infty}\left[\left(\ln\frac{P_0}{1-p}\right)^n \frac{1}{n!}x^n\right] = 1 - p + pq. \tag{21}$$

By neglecting terms of the series with orders $n = 2$ and higher, we obtain:

$$1 + \left(\ln\frac{P_0}{1-p}\right)x \approx 1 - p + pq. \tag{22}$$

After re-inserting $x = \frac{1}{N-1}$, this leads to

$$\frac{P_0}{1-p} \approx e^{-(N-1)(1-q)p}. \tag{23}$$

Finally, since N is assumed to be large, we approximate $N - 1$ as N in the exponential term, resulting in:

$$\frac{P_0}{1-p} \approx e^{-N(1-q)p}. \tag{24}$$

C Appendix: Conditional probability variable marginal probability model

Consider a 1D grid (Fig. 4a) where each voxel follows the *non-tunneling assumption*. For two neighboring voxels, C_i and C_k, with marginal probabilities p_i and p_k such that $p_i \geqslant p_k$, the conditional probabilities are defined as:

$$\begin{aligned}
&\mathbb{P}(C_k = 1 \mid C_i = 1) = aq + (1-q)p_k, \\
&\mathbb{P}(C_k = 1 \mid C_i = 0) = (1-q)p_k, \\
&\mathbb{P}(C_i = 1 \mid C_k = 1) = q + (1-q)p_i, \\
&\mathbb{P}(C_i = 1 \mid C_k = 0) = bq + (1-q)p_i.
\end{aligned} \tag{25}$$

Here, a and b are scalar factors that modulate the probability of copying or flipping the state of the neighboring voxel.

To determine a and b, we use the expectation constraints:

$$\mathbb{E}[C_i] = p_i, \quad \mathbb{E}[C_k] = p_k. \tag{26}$$

Expanding these expectations:

$$p_k = \frac{1}{2}\Big(p_k + p_i\big(aq + (1-q)p_k\big) + (1-p_i)(1-q)p_k\Big), \tag{27}$$

$$p_i = \frac{1}{2}\Big(p_i + p_k\big(q + (1-q)p_i\big) + (1-p_k)\big(bq + (1-q)p_i\big)\Big). \tag{28}$$

Solving for a and b, we obtain:

$$a = \frac{p_k}{p_i}, \quad b = \frac{p_i - p_k}{1 - p_k}. \tag{29}$$

These values ensure that the marginal probabilities remain consistent across the system.

D Appendix: Data Calculated from Bajgira et al.

In this section, we describe how the prostate cancer population data from Bajgira et al. [19] was obtained and processed for our analysis.

Patient and Tumor Classification

Table 1 of their study reports histologic findings for 518 patients who underwent multiparametric MRI followed by prostatectomy. Tumors in their paper can be classified into:

1. Index tumors (i): the main lesion in each patient, defined as the tumor with the highest Gleason Score. For the purpose of our analysis, these were identified as the gross tumor volumes (GTVs).
2. Islet tumors (j): additional lesions with lower Gleason Score, considered multifocal satellites of the index tumor.
3. Total tumors (T): the full set of index and islet tumors.

The study also distinguishes between lesions detected by MRI and those that were missed but later identified by histopathology.

Notation and Data Summary

For clarity, we define below the notation and numerical values extracted from Bajgira et al. [19], which are used throughout this appendix.

Symbol	Definition	Value
N_{patients}	Total number of patients	518
N_T	Total number of missed tumors	563
N_i	Number of missed index tumors	111
N_j	Number of missed islet tumors ($N_T - N_i$)	452
μ_T	Mean radius of missed tumors	9.4 mm
μ_i	Mean radius of missed index tumors	15.8 mm
μ_j	Mean radius of missed islet tumors	to be calculated
σ_T	Std. dev. of missed tumors	0.72 mm
σ_i	Std. dev. of missed index tumors	8.8 mm
σ_j	Std. dev. of missed islet tumors	to be calculated

Table 3: Summary of variables and values used in the appendix.

Proportion of Patients Without Microscopic Disease

Among the 518 patients, 396 presented multifocal disease (i.e., one or more islet tumors in addition to the index tumor). If we assume that every multifocal case includes at least one MRI-invisible lesion, the probability of a patient having no microscopic multifocal disease is estimated as:

$$P_0 = 1 - \frac{396}{518} = 0.28.$$

This assumption is supported by the observation that of the

$$567 = \underbrace{1085}_{\text{Total tumors}} - \underbrace{518}_{\text{Index tumors}}$$

islet tumors, only

$$115 = \underbrace{522}_{\text{Detected tumors}} - \underbrace{407}_{\text{Detected index tumors}}$$

were detected by MRI, meaning the majority were invisible.

Average Number of Missed Islet Tumors

The average number of missed islet tumors per patient is

$$\frac{N_T - N_i}{N_{\text{patients}}}.$$

Substituting the reported values:

$$\frac{563 - 111}{518} = 0.87.$$

Average Radius of Missed Islet Tumors

The mean radius of missed islet tumors is obtained from the weighted difference:

$$\mu_j = \frac{\mu_T \cdot N_T - \mu_i \cdot N_i}{N_j}.$$

Substituting the reported values:

$$\mu_j = \frac{9.4\,\text{mm} \cdot 563 - 15.8\,\text{mm} \cdot 111}{452} = 7.82\,\text{mm}.$$

Standard deviation of missed islet tumors

We derive σ_j from first principles, starting with sums of squares.

Let the observed radii of the N_T missed tumors be $x_1, \ldots, x_{N_T}$. Partition the indices so that the first N_i are the missed index tumors and the remaining N_j are the missed islet tumors (with $N_T = N_i + N_j$). Then the total sum of squares splits as

$$\sum_{n=1}^{N_T} x_n^2 = \sum_{n\in i} x_n^2 + \sum_{n\in j} x_n^2.$$

Divide both sides by N_T to obtain the total second moment:

$$\frac{1}{N_T}\sum_{n=1}^{N_T} x_n^2 = \frac{1}{N_T}\sum_{n\in i} x_n^2 + \frac{1}{N_T}\sum_{n\in j} x_n^2.$$

Recognizing subgroup averages,

$$\mathbb{E}[X^2]_T = \frac{N_i}{N_T}\left(\frac{1}{N_i}\sum_{n\in i} x_n^2\right) + \frac{N_j}{N_T}\left(\frac{1}{N_j}\sum_{n\in j} x_n^2\right),$$

so

$$\boxed{\mathbb{E}[X^2]_T = \frac{N_i\,\mathbb{E}[X^2]_i + N_j\,\mathbb{E}[X^2]_j}{N_T}}.$$

Using $\mathbb{E}[X^2]_k = \sigma_k^2 + \mu_k^2$ for $k \in \{T, i, j\}$ gives

$$\sigma_T^2 + \mu_T^2 = \frac{N_i(\sigma_i^2 + \mu_i^2) + N_j(\sigma_j^2 + \mu_j^2)}{N_T}.$$

Solving for σ_j yields the population-standard deviation expression

$$\boxed{\sigma_j = \sqrt{\frac{N_T(\sigma_T^2 + \mu_T^2) - N_i(\sigma_i^2 + \mu_i^2)}{N_j} - \mu_j^2}\,.}$$

by replacing the values from table 3 we obtained $\sigma_j = 5.76$ mm

Average islet volume calculation

To estimate the average volume of tumor islets ($\bar{V}_I$), we modeled their radius as a random variable

$$r \sim \mathcal{N}(\mu_j, \sigma_j),$$

where μ_j and σ_j are the mean and standard deviation of the islet radius obtained in the previous subsection. Since tumor radii cannot be negative, we applied a truncated normal distribution restricted to $r > 0$. Assuming spherical geometry, the volume of an islet is

$$V = \frac{4}{3}\pi r^3.$$

We generated a large sample of radii from the truncated normal distribution, computed the corresponding spherical volumes, and summarized the resulting distribution. The estimated population mean volume was

$$\mathbb{E}[V] \approx 513\,\mathrm{mm}^3,$$

with the 5th and 95th percentiles at $1\,\mathrm{mm}^3$ and $2024\,\mathrm{mm}^3$, respectively, providing an empirical 90% interval for the islet volumes.

Average Total Microscopic Tumor Volume

The Average Total Microscopic Tumor Volume ($\bar{V}_T$) per multifocal patient was calculated by multiplying $\bar{V}_I$ by the average number of islets per patient:

$$\bar{V}_T = \bar{V}_I \times \frac{N_j}{N_{\text{multifocal patients}}} = 513\,\mathrm{mm}^3 \times \frac{452}{396} \approx 584\,\mathrm{mm}^3.$$

Similarly, the 5th and 95th percentiles of the $\bar{V}_T$ distribution were obtained by scaling the corresponding percentiles of $\bar{V}_I$ distribution by the same factor:

$$V_{5\%} \approx 1\,\mathrm{mm}^3 \times 1.14 \approx 1.14\,\mathrm{mm}^3, \quad V_{95\%} \approx 2024\,\mathrm{mm}^3 \times 1.14 \approx 2307\,\mathrm{mm}^3.$$

Thus, the $\bar{V}_T$ provides an estimate of the total microscopic tumor burden from islets in multifocal patients, with an empirical 90% interval of approximately 1–2307 mm^3.

E Prostate Simulation Results

In this section we show the simulations of the average islet volume ($\bar{V}_I$) for the prostate patient (Fig 14).

F Breast Simulation Results

In this section we show the simulations of the average total microscopic tumor volume ($\bar{V}_T$) for the breast patient (Fig 15).

G Lung Simulations Results

In this section we show the simulations of the average total microscopic tumor volume ($\bar{V}_T$) for the lung patient (Fig 16).

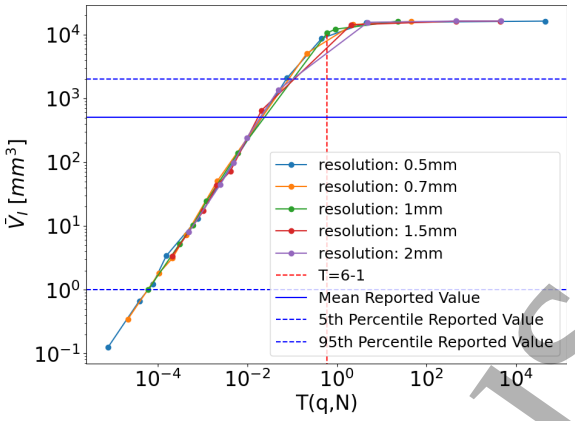


Figure 14: Simulated $\bar{V}_I$ for the prostate case across different grid resolutions and values of $T(q, N)$. Each color represents a different grid resolution. The solid blue line indicates the $\bar{V}_I$ reported by Bajgira et al., while the dashed black lines correspond to the 5th and 95th quantiles estimated from [19]. The red dashed line marks the value of T that produces the $\bar{V}_I$ reported in Bajgira et al. [19]. Both axes are in logarithmic scale.

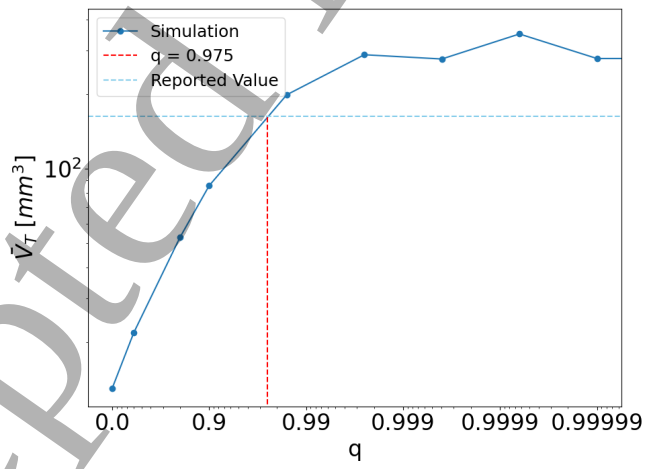


Figure 15: $\bar{V}_T$ simulated using a grid resolution of 2 mm for the breast cancer case. The skyblue dash line shows the $\bar{V}_T = 162\ mm^3$ estimated from Stroom et al. and the red dashed line shows the value of q needed to replicate that result by doing simulations [25]

.

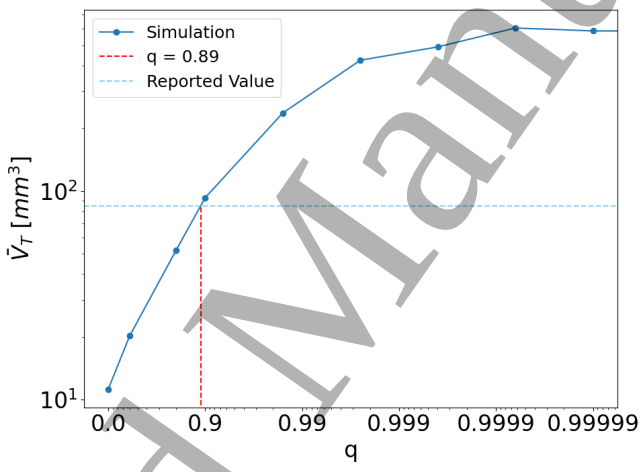


Figure 16: $\bar{V}_T$ simulated in the lung case using a grid resolution of 2 mm. The skyblue dash line shows the $\bar{V}_T = 82\ mm^3$ estimated from Stroom et al. and the red dashed line shows the value of q needed to replicate that result by doing simulations [25].