Aus dem Lehrstuhl
für Funktionelle Genomik
Prof. Dr. Rainer Spang
der Fakultät für Medizin
der Universität Regensburg

Prospective and retrospective genetic tumour evolution simulation based on Mutual Hazard Networks

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Medizin

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Stefan Hansch

2025

Aus dem Lehrstuhl
für Funktionelle Genomik
Prof. Dr. Rainer Spang
der Fakultät für Medizin
der Universität Regensburg

Prospective and retrospective genetic tumour evolution simulation based on Mutual Hazard Networks

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Medizin

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Stefan Hansch

2025

Dekan:	Prof. Dr. med. Dipl.-Phys. Dirk Hellwig
1. Berichterstatter:	Prof. Dr. rer. nat. Rainer Spang
2. Berichterstatter:	Prof. Dr. med. Dipl.-Math. Bernd Salzberger
Tag der mündlichen Prüfung:	11.12.2025

1 Abstract

Contents

1.1 Abstract (English)

This work proposes a novel machine learning-based simulation concept that fuses a specific machine learning algorithm (Mutual Hazard Networks (MHN)) with a stochastic simulation algorithm (a modified Gillespie algorithm) to iteratively reconstruct and predict potential genetic mutation sequences and their interactions in tumours. A prototypical web interface has been developed to facilitate the intuitive utilisation of our simulation and prediction of individual tumours. The simulation model was evaluated and validated on several types of specific sets of genetic alterations in tumours from different studies and various cancer types. Additionally, a comparative analysis of learned models and their resulting simulations with real data from differing genetic evolutionary points was conducted, which enabled a first statistical significant prognostic prediction in certain types of lung cancer based on the gene *STK11 / LKB1*.

1.2 Abstract (German)

In dieser Arbeit wird das Konzept einer neuartigen, auf maschinellem Lernen basierenden Simulation vorgestellt, die einen spezifischen Algorithmus für maschinelles Lernen (Mutual Hazard Networks (MHN)) mit einem stochastischen Simulationsalgorithmus (einem modifizierten Gillespie-Algorithmus) verbindet, um mögliche genetische Mutationssequenzen und ihre Wechselwirkungen in Tumoren iterativ zu rekonstruieren und vorherzusagen. Um die intuitive Nutzung unserer Simulation und die Vorhersage individueller Tumore zu erleichtern, wurde ein Webseiten-Prototyp entwickelt. Das Simulationsmodell wurde anhand mehrerer Arten spezifischer genetischer Veränderungen in Tumoren aus verschiedenen Studien und unterschiedlichen Organsystemen evaluiert und validiert. Darüber hinaus wurde eine vergleichende Analyse der gelernten Modelle und der daraus resultierenden Simulationen mit realen Daten aus verschiedenen genetischen Evolutionspunkten durchgeführt. Diese ermöglichte eine erste statistisch signifikante prognostische Vorhersage auf der Basis des Gens *STK11 / LKB1* in bestimmten Lungentumoren.

2 Introduction

Although medical advances have improved the ability to prevent, detect and treat malignant tumours, cancer remains the leading cause of death in people under the age of 85 in the United States, as shown in the data presented by Siegel et al. [1].
The treatment of malignant tumours is a major challenge, partly due to the potential development of drug resistance. The development of drug resistance can occur not only with conventional therapies (such as cytostatics) but also with newer treatments that target specific genetic abnormalities. Examples of such methods include CAR-T cells, monoclonal antibodies, checkpoint inhibitors and other similar approaches [2]. In this context, the combination of different targeted therapies represents a potential strategy for overcoming such resistance [3].
A more comprehensive grasp of past and potential future individual genetic alterations may facilitate a deeper comprehension of tumourigenesis, thereby enhancing the efficacy of personalised cancer therapy. In order to address these questions, a new machine learning-based simulation is proposed. The approach combines a specific machine learning algorithm (Mutual Hazard Networks (MHN)) with a stochastic simulation algorithm (a modified Gillespie algorithm) in order to iteratively reconstruct and simulate possible genetic mutation sequences and their interactions in tumours.
The simulations are validated and evaluated as a concept on publicly available data sets. Initially, they are examined in general with respect to different types of cancer, including glioblastoma, renal cancer, and breast cancer. Subsequently, they are subjected to more detailed scrutiny with a particular focus on lung cancer.
Additionally, a comparative analysis of learned models and their resulting simulations with real data from differing genetic evolutionary points is conducted, which enables a first prognostic prediction. The prediction was conducted on lung cancer data utilising a specified tumour gene set to ascertain the probability of developing a poor prognostic mutation (*STK11 / LKB1*) through this methodology.
Moreover, to facilitate usability, a web interface has been developed. This allows the user to input a specific set of mutations in a tumour and the underlying learned possible genetic interactions, and then to study and simulate possible future mutations.
This work will commence with the present Introduction. The following Chapter 3 will present the fundamental methodologies and background knowledge, while Chapter 4 will delineate the methodologies and data employed in this study. The results obtained on simulated data will be examined in Chapter 5, and Chapter 6 will conclude with a discussion.

3 Background, related work and basic model restrictions

3.1 Evolution of cancer

Greaves et al. [4] and Nowell et al. [5] described a hypothesis for the evolution of tumour cells. This hypothesis states that tumours develop in a sequential evolutionary manner, which is in line with Darwin's theory of evolution. It is described as a process in which cell alterations (mainly due to genetics, epigenetics, or regulatory factors) occur with selective advantage, selective disadvantage, or no effect. Cells with a selective advantage over other competing cells are likely to predominate within a subpopulation. Tumours arise when this process occurs multiple times in cells, creating a cycle of clonal and sub-clonal evolution. A possible outcome of these cycles is defined as malignant if the cell composite leads to predetermined disadvantages for the organism.

In the contemporary context [4, 6], the field of cancer research is dominated by a predominant paradigm that perceives cancer progression as a multifaceted, individual process that occurs on an evolutionary basis. This paradigm is characterised by the presence of partially random mutational processes, which involve the emergence of genetic driver, passenger, and mutator alterations.

Stochastic processes in cancer evolution depend on complex underlying biological and biochemical actions and interactions, including, but not limited to

- Patient factors such as comorbidities, fitness, or immunodeficiency
- Tumour attributes (e.g. tumour age, localisation, or organ type)
- Medical therapy (e.g. surgery, medication, or radiation)
- Histological characteristics (e.g. cancer cell type, degree of malignancy, or cell receptor status)
- Underlying primarily genetic causes [7, 8] (e.g. mutations and clonal selection)
- Underlying non-directly genetic causes [7, 8, 9, 10] (e.g. the tumour microenvironment and non-genetic variability)

Cancer progression is a current research topic with very limited predictability for specific settings [8, 11, 12]. Recently, Diaz-Colunga et al. [8] compared different cancer progression models (CPMs) to prospectively predict short-term mutations in specified tumour evolution scenarios.

Short-term mutations of a given tumour genotype in this context were the next mutation n+1, leading to the new genotype. Therefore, MHN, among Conjunctive Bayesian Networks (CBN) [13, 14, 15], Oncogenetic Trees (OT) [16, 17], CAncer PRogression Inference (CAPRI) [18, 19] and CAncer PRogression Extraction with Single Edges (CAPRESE) [20] were investigated (citation, as seen in [8]). One shared conclusion was that CPMs could potentially provide a gain of knowledge for prospective tumour evolution. Still, more research is needed in specific settings for reliable and stable predictions for use in medical practice.

3.2 Mutual Hazard Networks

Mutual Hazard Networks (MHN), published by Schill et al. [21], is a machine learning algorithm that infers cyclic progression models from cross-sectional data using a continuous-time Markov process as a tumour progression model. Given events that can be represented as binary events (e.g. mutations, copy number alterations, methylation data or epigenetic data) in a collective of data sets, this algorithm can generate two types of connections between events. The first type is an interconnected network of events. It shows the frequency of occurrence and spontaneous fixation of binary events and their multiplicative effects on the rates of successive events. This rate can be facilitating or inhibiting. The second type of connection is a possible sequence of events calculated using maximum likelihood paths leading from a initial tumour progression state to the observed tumour state.
Unlike other algorithms in the field (e.g. Conjunctive Bayesian Networks (CBN) [22]), MHN allows cyclic graphs and therefore mutual exclusivity. Mutual exclusivity is the term given to the phenomenon in which if a specific event occurs, it is unlikely that another specific (mutually exclusive) genetic alteration will occur in the same tumour. This phenomenon is often observed in cancer genetic analysis [23]. It can occur, for example, when the events are in different pathways or potentially lethal.
Following the notation, variables and definitions of Schill et al. [21], the Markov process is specified by a system of n functions:

$$Q_{x_i,x} = f_i(x) = \exp(\theta_{ii} + \sum_{i=0}^{n} \theta_{ij} x_j) = \Theta_{ii} \prod_{x_j} \Theta_{ij}. \tag{1}$$

After an MHN-model is learned, the Θ-matrix is given by:

$$\Theta_{ij} := \mathrm{e}^{\theta ij} \in \mathbb{R}^{n \times n}. \tag{2}$$

The base risk of event *i*, given by Θ_{ii}, is defined as the rate in a neutral state, which means prior to other simulated events happening. When event *j* happens, the multiplicative effect of event *j* on the rate of event *i* is given by Θ_{ij}.
Furthermore, additional developments and extensions have been devised or are currently being developed for MHN [24, 25, 26], including an efficient computing library for Python [27], which is utilised in Section 5.4.

3.3 Gillespie algorithm

Gillespie et al. [28] described two main formalisms for simulating the time behaviour of spatially homogeneous chemical systems. The first one is the stochastic approach, which treats temporal evolution as a kind of random process driven by a single differential equation (the *"master equation"*). The other one is the traditional deterministic approach, which views temporal evolution as a continuous, fully predictable event-driven set of coupled ordinary differential equations (the *"reaction rate equations"*). The Gillespie algorithm was first published in 1977 [28] and was initially used to simulate (bio)chemical reaction systems with known reaction times and limited computing power. It uses a strictly derived Monte Carlo method. Monte Carlo methods are computational algorithms based on repeated random events.
The Gillespie algorithm uses the stochastic approach and therefore does not attempt to approximate infinitesimal time steps dt with finite time steps Δt.
Since its publication, (modified) Gillespie algorithms have been used in many modifications in various fields. See, for example, [29, 30, 31].

4 Methods and used data sets

4.1 Data set selection and restrictions

One type of publicly available data set employed in this study to obtain raw tumour data, as required in Section 5.1, is fully pre-processed and displayed in Section 4.2. These data were utilised in Schill et al. [21] and Gerstung et al. [13]. In this study, these data are employed to establish a comparison with a designated reference.
Another category of publically accessible data sets is utilised to obtain genetic profiles of tumour data that have undergone sequencing. However, the utilisation of these data in Chapter 5 necessitates preliminary processing for the purpose of conducting simulations and adapting to their underlying algorithms, as outlined in Section 4.3. In order to identify suitable data sets for the further validation of the data and to provide a foundation for real-world tumour data, the *cBioPortal* [32, 33] platform was employed.
It is important to note that all of the data sets utilised in this study, in conjunction with the used models, are reliant upon the analysis of bulk tissue mutational profiles. It is noteworthy that the precise correspondence of sequences to cellular or cell population levels remains to some extent ambiguous [34].

4.2 Reference data sets

The data set referred to as *Baudis* contains mutational profiles that were published with the *Progenetix database* (Baudis et al. [35]). For the purposes of this study, the preprocessed data that was published by Schill et al. [21] has been utilised. The data set is composed of 817 cases of breast cancer, 570 cases of colorectal cancer, and 251 cases of renal cell carcinoma, which were characterised by 10, 11, and 12 recurrent copy number alterations, respectively.

4.3 Preprocessing pipeline and basic model constraints

As data sets require preliminary processing for the purpose of our simulations, this is undertaken in accordance with a standardised procedure:
To uniformly convert ensemble transcript IDs to ensemble Hugo symbols *EnsDb.Hsapiens.v86* was used with the *R*-Package of Rainer J. [36]. The sequenced data set was filtered for binary mutation information without distinguishing between different mutation types (like frame-shift, missense or silent). For each data set, unless otherwise specified, and to simulate efficiently without deeper prior knowledge, the ten most mutated genes (samples with at least one mutation, according to the *cBioPortal* website [32, 33]) were utilised in the extraction process. When comparing two data sets, the most quantitatively mutated genes from each data set were used, and genes not examined by the other data set were discarded. This pipeline did not address comparability issues caused by different next-generation sequencing processing methods.
In addition to retrospective analyses of the evolution of de novo tumours, this study also employs a prospective approach to tumour simulations. The objective is to identify potential mutations that could occur in an existing tumour as it progresses further along its development trajectory. It was thus necessary to simulate events derived from data sets comprising genetic data from a minimum of two distinct temporal stages of development. The model was defined as requiring that the later data set must be a genetic descendant of the earlier data set. In the preparation process, it was essential to ensure data quality and to simulate only existing tumours. Therefore, data points without sequencing time or mutations in a set of the ten most frequently mutated genes were discarded.

4.4 *cBioPortal* data sets

The first data set of these data sets, referred to as *Metabric*, was published by Pereira et al., Rueada et al. and Curtis et al. [37, 38, 39]. This data set comprises targeted sequencing of 2509 primary breast tumours with 548 matched normal tissue samples. Of these tumour samples, 2369 samples remain after pre-processing.
The second data set, referred to as *Razavi* or *Razavi Breast*, published by Razavi et al. [40], includes sequencing (of the MSK panel) of tumour/normal sample pairs from 1918 breast cancers, of which after pre-processing, 1542 are available.
The third data set from Kan et al. [41] consists of 187 primary breast tumours from a Korean cohort (referred to as *SMC*). Of these, 185 records can be used after pre-processing.
Three data sets were employed for the progressive simulations:
One of these data sets was published by Jordan et al. [42], and referenced here as *Jordan Lung*. The cohort comprises 860 patients with metastatic lung adenocarcinoma who underwent genetic analysis and, if appropriate, targeted treatment according to the identified gene mutations. With the exception of patients with multiple samples or those also present in Jee et al. [43] (see description below), 764 samples with the 40 most common gene mutations were employed as a learning set in Section 5.4. The second data set was published by Jee et al. and will be referred to as *Jee Lung*. The data set features targeted sequencing from 1127 patients with metastatic non-small cell lung cancer (NSCLC) and ctDNA-guided therapy. In Section 5.3 of this data set 474 were used in the corresponding learned model for the individual tumour progression simulation (n=34 pairs). In Section 5.4, the data set was used for testing purposes, utilising the 40 most frequently mutated genes from the *Jordan Lung* data sets. From the initial set of 2621 genetic profiles, those comprising a single data point were primarily excluded (remaining n = 2233). Subsequently, data sets exhibiting identical mutations in a single patient (remaining n = 1604) were excluded, again resulting in the exclusion of patients with only one remaining data set (n = 1485). In consequence, 1054 sets resulted in a progressive simulation, and of these, 953 were deemed valid, as no conflicts were identified in the mutational order (as described in Section 5.3). Data sets lacking mutations in the 40 analysed genes have been excluded, leaving 562 sets (of 346 patients) with two evolutionary times. At last, the patients with lung problems that were seen in Jordan have been filtered out, leaving n = 532 (of 334 patients) data sets.
The last data set was published by Lengel et al. [44] and will be henceforth referred to as *Lengel Lung*. This data set contains 2532 lung adenocarcinomas and was collected with a focus on the study of metastatic organotropism. Of these lung adenocarcinomas, 1989 were used in the learning process and 25 pairs were used to test the individual progression simulation in Section 5.3.

Table 1
Synopsis of utilized data sets

Data set	Type	Section	n available	n used
Progenetic/Baudis	Breast cancer	5.2	-	817
Progenetic/Baudis	Colorectal cancer	5.2	-	570
Progenetic/Baudis	Renal cell Caricoma	5.2	-	251
Progenetic/Baudis	Glioblastoma	5.2	-	261
cBioPortal/Metabric	Breast cancer	5.2	2509	2369
cBioPortal/Razavi	Breast cancer	5.2	1918	1542
cBioPortal/SMC	Breast cancer	5.2	187	185
cBioPortal/Jee	Lung cancer	5.3	1127	474L, 34P*2S
cBioPortal/Lengel	Lung cancer	5.3	2532	1989L, 25P*2S
cBioPortal/Jee	Lung cancer	5.4	1127	334P, 532S
cBioPortal/Jordan	Lung cancer	5.4	860	764

* Note: The following is an overview of the used data sets, with the number of patients *n* in learning or testing indicated when there is no character. The following characters are further delineated in this table: The symbol *L* denotes the number of patients in the learned data, whilst *P* refers to the data points in the testing phase. The symbol *S* is used to refer to the available data points.

4.5 Modified Gillespie algorithm simulation using MHN

Definitions and algorithmic details of our modified Gillespie algorithm simulation using MHN are explained below.

The *impact* of gene J (corresponding index j) is defined as a constant number representing the sum of all multiplicative effects (as absolute values) that an event of gene J has on other simulated genes (indexed $0..n$).

$$\mathrm{Impact}(J) = \sum_{\substack{i=0 \\ i \neq j}}^{n} |\theta_{ij}|. \tag{3}$$

The probability of a given event X(corresponding index x) occurring next is defined as $P(X)$. Let I be the set of all possible events. We define two subsets:

- $N = \{y_1, \ldots, y_n\} \subset I$ contains events that have not yet occurred.
- $M = \{z_1, \ldots, z_m\} \subset I$ contains events that have already occurred.

$P(X)$ is calculated by:

$$P(X) = \frac{u}{s} \tag{4}$$

with

$$u = \mathrm{e}^{\theta_{xx} + \sum_{l \in M} \theta_{xl}} \tag{5}$$

and

$$s = \sum_{k \in N} \left(e^{\theta_{kk} + \sum_{l \in M} \theta_{kl}} \right) \tag{6}$$

The observation period T is defined as the (unknown) time from a hypothetical mutation-free individual or cell t_0 to the time of first diagnosis (here, the time of first sequencing). An event period t_{event} is defined as the time between the previous and the next event. In addition to mutations, the current event can be t_0, and the next event can be the end of the observation period T. When the actual time t and the actual event time t_{event} is greater than T, the end of the observation time is reached. According to Schill et al. [21], the simulated time is an abstract variable that cannot be translated into real time and was simulated using the exponential distribution.

Let a, b be random numbers with:

$$a, b \in \mathbb{R} : a, b \sim U(0, 1). \tag{7}$$

then

$$T = \frac{-\log(a)}{1} \tag{8}$$

and

$$t_{event} = \frac{-\log(b)}{s} \tag{9}$$

If an event occurs within the observation, a random number $c \in \mathbb{R} : c \sim U(0, 1)$ is used to determine which event $X \in N$ will be drawn. The event is selected such that the cumulative probability $\sum_{i \in N, i \leq X} P(i)$ first exceeds c. This process is referred to as *calculate random event*.

The probability $P_T(X)$ of an event X occurring in an observation period T is simulated using a modified Gillespie algorithm. A simplified iteration of this algorithm is described in the Algorithm 1. An iteration represents one observation period. After each iteration, a tally is made of each mutated gene. After 1000 simulations, the probability $P_T(X)$ is approximated from its observed appearances.

Algorithm 1 Iteration of modified Gillespie

Require: $T,\ \Theta$

$t \leftarrow 0, calculate\ all\ P(X)$
while $t < T \quad \& \quad N \neq \emptyset$ **do**
    $calculate\ s, t_{event}$
    if $t + t_{event} < T$ **then**
        $event = calculate\ random\ event$
        $update(N), update(M)$
        $update\ all\ remaining\ P(X)$
    end if
    $t \leftarrow t + t_{event}$
end while

4.6 Interactive tumour simulation

The development of the prototype software was conducted in *JavaScript*, utilising the *D3.js JavaScript* library. The prototype software comprises two distinct parts:
The first part is a web interface (also referred to as front-end), where a user, given a certain tumour genome and a corresponding learned Θ-matrix, can simulate automatically or manually further possible mutations with the statistics described in Section 4.5. Genes are arranged in accordance with the actual probability of mutation, in order to facilitate intuitive anticipation of subsequent potential mutations. By positioning the cursor over a specific gene, a preview of the calculated probability of that gene being affected by the event is provided.
The utilisation of such simulation software has the potential to assist medical professionals in enhancing their clinical decision-making capabilities in the future. As this software is intended primarily for scientific purposes, its design was inspired by the "*Publication Manual of the American Psychological Association*" [45] for scientific tables.
The second component is the back-end, the function of which is to provide and calculate statistical and simulation data. This data has previously been mentioned, and it is based on a Θ-matrix (see Section 4.5). Figure 1 presents a screenshot exemplifying the software interface. This part has been further developed for the purpose of validation and testing, a process which is described in Sections 5.1 and 5.2.

4.7 Statistical analysis software

The software *R* version 4.1.2 was utilised in Chapter 5 to preprocess data sets, and the *R* packages *survival* [46], *groupdata2* [47], *dplyr* [48] and *survminer* [49] were employed for further statistical analysis and the creation of statistical graphs. In the context of simulations, the software delineated in Section 4.6 will be utilised.

Interactive Table 1

Simulation of prospective tumor mutations using modified Gillespie algorithm

Gene	Impact on other genes	Mutation in period?	Mutation is next event?	Δ
-13q	2.61	3.44%	34.39%	NaN
-4q	0.34	3.81%	27.05%	29.98%
-9p	0.00	2.21%	10.36%	11.70%
-18q	0.03	1.06%	6.67%	-1.55%
+17p	3.33	0.90%	6.60%	-2.28%
-4p	2.93	0.71%	4.75%	-1.10%
-Xp	0.01	0.67%	4.47%	-1.04%
+Xp	0.00	0.58%	3.31%	-0.77%
-8p	1.52	0.37%	2.41%	-0.56%
-3p	0.73	0.00%	NaN	NaN
+17q	3.89	0.00%	NaN	NaN
-6q	1.63	0.00%	NaN	NaN

Note. Impact is calculated on how much the mutation of this gene influences the propability of mutation of other genes.

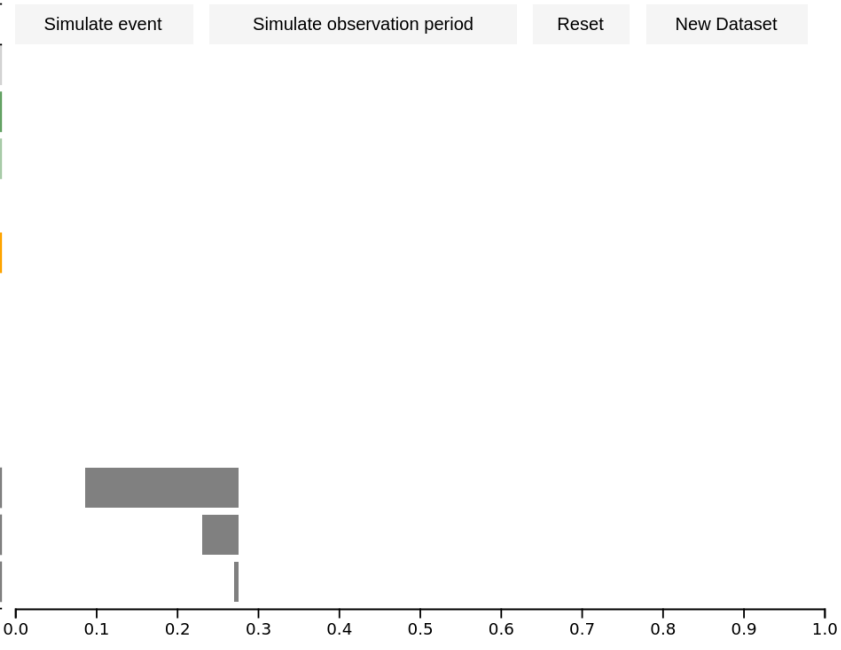


Figure 1

Screenshot of the interactive tumour simulation software. On the left, statistics with gene name, impact, $P_T(x)$, $P(x)$ and Δ. Δ represents the change in $P(x)$ due to the selected possible mutation (in light grey). On the right, time simulation with $T = 1$. The learned data set was a renal cancer data set from Baudis et al. with $\lambda = 0.01$. Three events (-3p, +17q and -6q) occurred at different times and are shown in grey. One event (-13q in light grey) is floating to show its effect on other genes by mouse hovering.

5 Results

In this Chapter, the developed simulation algorithm is first of all subjected to a validation for the correctness of its implementation in Section 5.1. This is followed by validation with different data sets in Section 5.2, and a highly specified progressive tumour simulation with comparison to cases with two temporal data points is performed in Section 5.3. The model will be extended and tested in a presumably realistic scenario with corresponding survival data in Section 5.4.

5.1 Internal Validation

In order to ascertain the correctness of the implemented algorithm in *JavaScript*, and due to the unavailability of a suitable modified Gillespie algorithm in *JavaScript* for comparison with the one implemented here, an internal validation was performed with an existing basic Gillespie algorithm in *Python* [50].
This validation was conducted after further development and modification of the original Gillespie algorithm. Utilising the *JavaScript* and *Python* implemented Gillespie simulation algorithm, three simulations with the reference data set of breast cancer, described in Section 4.2 and the corresponding Θ-matrix of $10.000 * 1.000$ tumours were performed with random exponential time using parameter $r = 1$ of the *Python* and the *JavaScript* algorithms.
Figure 2 shows a density plot comparing the simulated mutation count of the gene - *13q* between these two implementations. As the random exponential function was also implemented in *JavaScript*, this was validated with the pre-existing random number function used in MHN (visualised in Figure 3).

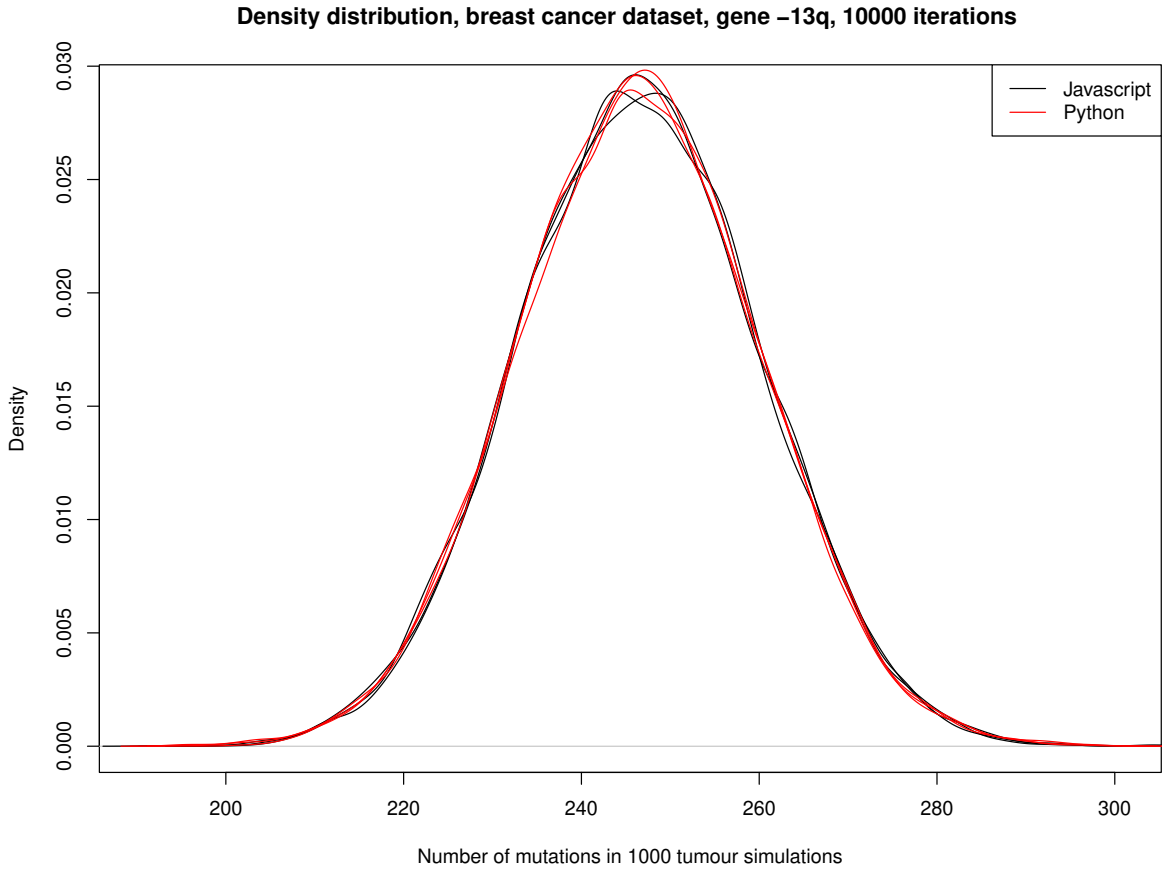


Figure 2

Density plot comparing the mutation count of the gene -13q with 10000 observations of 1000 tumour simulations from two implementations of the iterative Gillespie algorithm (Python: first data set: mean 246.3528, sd 13.57628, second data set: mean 246.527, sd 13.56794, third data set: mean 246.5293, sd 13.64655; JavaScript: first data set: mean 246.5714, sd 13.56432, second data set mean 246.5632, sd 13.60509, third data set: mean 246.5938, sd 13.4838). Breast cancer data set (Baudis) from Section 4.2 processed by MHN with $\lambda = 0.01$.

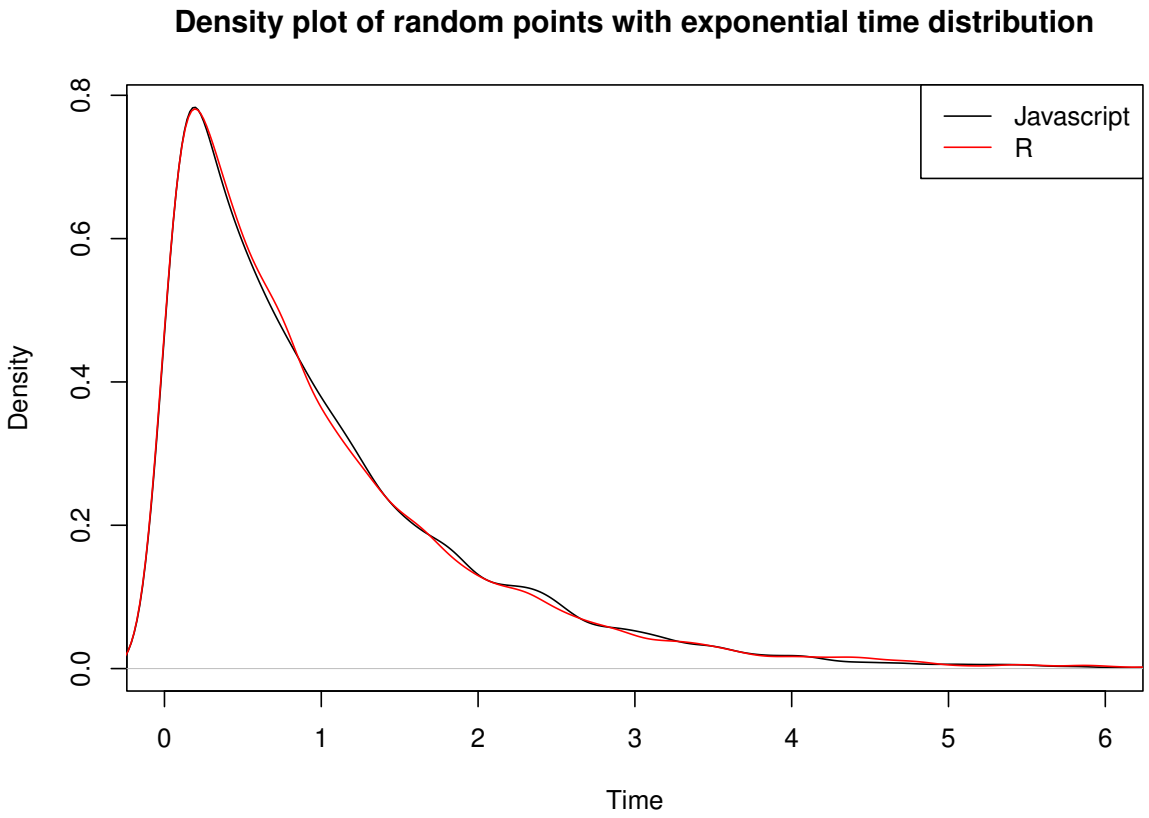


Figure 3

Density plot comparing two exponential distribution random functions with observation time 1 as a parameter, using R-function rexp(mean 0.9983856, SD 1.015696) as used by MHN and the developed JavaScript simulation algorithm (mean 0.9943825, SD 0.9809427). Time was simulated with 10000 data points and plotted in the range 0-6.

5.2 External Validation

In order to obtain validation of different data sets, a number of different data sets, as described in Chapter 4, are used. A semi-automated pipeline was developed, comprising the *JavaScript* implementation of the modified Gillespie algorithm for simulation and the unmodified MHN package in R for learning and training. The generation of simulated tumour data worked as follows:

1. Prepare a data set of n samples of tumours with g binary events, called raw tumour data
2. Generate a matrix Θ_{MHN} using MHN on the raw tumour data, Θ_{MHN} denoted as *MHN: data set name* (e.g. *MHN: Breastcancer*)
3. Generate n simulated tumour data using the modified Gillespie algorithm with Θ_{MHN}, called simulated tumour data
4. Generate a matrix Θ_{Gil} using MHN of n simulated tumour data, Θ_{Gil} referred as *SIM: data set name* (e.g. *SIM: Breast Cancer*)
5. Validate

Validation was performed *in-dataset*, meaning that an MHN data set (2) was checked against its raw tumour data (1) or a SIM data set (4) was checked against its simulated tumour data (3). The quality criteria employed were consistent with those used by Schill et al. and included the likelihood score *Score* implemented in MHN, 5-fold cross-validation (*5F-CV*), and the Akaike Information Criterion (*AIC*). Furthermore, a comparison was conducted between different data sets, designated as *Intra-Dataset*. The likelihood score, denoted as *M-Score*, was employed as a comparison parameter for SIM data sets with raw tumour data. The Kullback–Leibler divergence, represented as *KL*, was utilised to assess the similarity between processed SIM data and MHN data. Finally, referred to as *Cross-Dataset*, the Akaike Information Criterion, designated as *M-AIC*, was employed to evaluate the compatibility between SIM data and raw tumour data.
M-Score, *KL* and *M-AIC* were also used analogues between different data sets *MHN Dataset1 x SIM Dataset2*", meaning that a model was learnt on *Dataset1* and tested on *Dataset2*. *M-Score* compares a learned model of *Dataset2* and the raw data of *Dataset1*. For the *M-AIC*, the aforementioned *M-Score* was used with n of *Dataset2*. In order to validate this methodology, an initial evaluation of tumour data generated by the Gillespie simulation was conducted on the same samples that Schill et al. used for their validation. These were published by Baudis et al. [35] and are shown in Figure 2.

The second stage of the validation process entailed an evaluation of the tumour data generated by the Gillespie simulation on the samples described in Section 4.4 and depicted in Table 3 and Table 4.
In order to create a first visual impression of real-world data, Figure 4 presents a picture which was drawn like a screenshot of *cBioPortal* with *Razavi*-Data, alongside *SIM-Razavi*-Data from the tumour simulation.

Table 2
Validation of MHN against SIM Data using Baudis-Data

		In-Dataset[1]			Intra-Dataset[2]		
Data set	n	Score	5F-CV	AIC	M-Score	KL	M-AIC
MHN: Breast cancer	817	-5.62	-5.67	9191	-	-	-
SIM: Breast cancer	817	-5.49	-5.52	8978	-5.69	0.0223	9299
MHN: Colorectal cancer	570	-5.62	-5.69	6411	-	-	-
SIM: Colorectal cancer	570	-5.75	-5.80	6558	-5.71	0.0349	6515
MHN: Renal cell carcinoma	251	-4.87	-5.02	2445	-	-	-
SIM: Renal cell carcinoma	251	-5.20	-5.37	2608	-4.99	0.0589	2507
MHN: Glioblastoma	261	-7.70	-7.91	4020	-	-	-
SIM: Glioblastoma	261	-6.70	-6.88	3495	-8.05	0.1354	4200

* Note: $\lambda = 0.01$ and mutational gene data (*Baudis*) as described in Section 4.2 were used. [1, 2] : For more information, see Section 5.2.

Table 3
Validation of MHN against SIM Data using Metabric-Data, Razavi-Breast-Data

		In-Dataset[1]			Intra-/Cross-Dataset[2]		
Data set	n	Score	5F-CV	AIC	M-Score	KL	M-AIC
MHN: *Metabric**	2369	-4.70	-4.71	22261	-	-	-
SIM: *Metabric**	2369	-4.39	-4.40	20803	-4.79	0.0256	22673
MHN: *Razavi**	1542	-3.80	-3.81	11726	-	-	-
SIM: *Razavi**	1542	-3.58	-3.59	11055	-3.91	0.0354	12065
MHN: *Metabric***	2369	-3.65	-3.66	17311	-	-	-
SIM: *Metabric***	2369	-3.48	-3.49	16510	-3.73	0.0199	17658
MHN: *Razavi***	1542	-3.70	-3.71	11406	-	-	-
SIM: *Razavi***	1542	-3.56	-3.57	10982	-3.80	0.0307	11708
MHN *Metabric* x MHN *Razavi***	-	-	-	-	-3.72	0.0732	11459
MHN *Metabric* x SIM *Razavi***	-	-	-	-	-3.77	0.0633	11614
MHN *Razavi* x MHN *Metabric***	-	-	-	-	-3.80	0.0740	18008
MHN *Razavi* x SIM *Metabric***	-	-	-	-	-3.88	0.0995	18380

Note: $\lambda = 0.01$. [*] Ten most frequent genes of the single data sets according to *cbioportal* were used. [**] The ten most common measured genes in both data sets were analysed (*PIK3CA, TP53, PTEN, AKT1, ARID1A, KMT2C, GATA3, CDH1, MAP3K1, TBX3*) - but *ESR1, MUC16, AHNAK2, SYNE1, DNAH11* were filtered, as described in Section 4.4.
[1, 2] : For more information, see Section 5.2.

Table 4
Validation of MHN / SIM Data using Razavi and SMC Breast Cancer data

		In-Dataset[1]			Intra-/Cross-Dataset[2]		
Data set	n	Score	5F-CV	AIC	M-Score	KL	M-AIC
MHN: *Razavi**	1542	-3.70	-3.71	11406	-	-	-
SIM: *Razavi**	1542	-3.52	-3.53	10856	-3.80	0.0322	11717
MHN: *SMC**	185	-2.62	-**	970	-	-	-
SIM: *SMC**	185	-2.52	-**	935	-2.74	0.0648	1014
MHN *SMC* x MHN *Razavi**	-	-	-	-	-2.84	0.212	8770
MHN *SMC* x SIM *Razavi**	-	-	-	-	-2.91	0.231	8967
MHN *Razavi* x MHN *SMC**	-	-	-	-	-4.13	0.377	1528
MHN *Razavi* x SIM *SMC**	-	-	-	-	-4.47	0.675	1654

Note: $\lambda = 0.01$. * Ten most frequent genes of the *Razavi* breast cancer set according to *cbioportal* were used (*PIK3CA, TP53, CDH1, GATA3, TBX3, MAP3K1, KMT2C, PTEN, AKT1, ARID1A*). Gene *ESR1* was dismissed, as no mutations happened in the *SMC*-Data set. ** Insufficient number of mutated genes to perform 5F-CV.
[1, 2] : For more information, see Section 5.2.

Gene	Impact on other genes	Mutation in period?	Mutation is next event?
TP53	1.09	32.70%	29.87%
PIK3CA	1.26	33.20%	28.41%
GATA3	0.45	15.10%	10.45%
CDH1	0.51	15.70%	9.84%
ESR1	0.01	9.40%	4.34%
MAP3K1	0.00	6.80%	4.04%
PTEN	0.00	7.10%	3.69%
KMT2C	0.00	6.30%	3.51%
ARID1A	0.00	7.30%	3.18%
AKT1	0.00	5.40%	2.66%

Mutated Genes (1918 profiled samples)

Gene	#Mut	#	Freq
PIK3CA	826	725	37,8%
TP53	696	681	35,5%
CDH1	313	310	16,2%
GATA3	298	288	15,0%
ESR1	175	164	8,6%
MAP3K1	220	155	8,1%
KMT2C	173	154	8,0%
PTEN	155	137	7,1%
AKT1	109	107	5,6%
ARID1A	113	106	5,5%
TBX3	105	93	4,8%

Figure 4

On the left, a screenshot of the interactive tumour simulation software is shown; on the right, a picture as seen on BioPortal [51] is shown. The interactive tumour simulation visualises a learned model of SIM-Razavi data, as described in Section 4.4 and Section 5.2. The right shows the visualised raw (real) data of the Razavi data set as described in Section 4.4.

5.3 Progressive tumour simulation

In addition to retrospective analyses of the evolution of de novo tumours, this study employs a prospective approach to tumour simulations. The objective is to identify potential mutations that could occur in an existing tumour in the further course of its development. The data set was split into two parts: one containing only one evolutionary point per case, and one containing cases with multiple data points. Since suitable simulation data are rare, the single development point data set (with the number of data points n_L) was used to learn an MHN model, and the data set with multiple time points data points was used for testing and simulation. The temporal order of the test set was defined primarily by sequencing time and secondarily by location (as it can be assumed, for example, that a metastasis originates from a primary tumour). In the simulation process, performed by the simulation software described above on individual data points from the same patient, the number of mutated genes in the earlier (e_{bef}) and later sets were counted and compared, generating the number of mutations to simulate e_{sim} in one iteration. In addition, sets with mutations in the earlier data set that were absent in the subsequent data set were excluded, as they lack direct descendants within our model. Using the number of events, a learned model and the previous mutations of an individual tumour, 1000 simulations were run with these pairs of sets (n_T). Comparing these simulated events with real alterations gives us the number of right (R_{sim}) and wrong simulated events (W_{sim}). A descriptive statistical analysis followed this (results are shown in Figure 5). For comparability, a baseline simulation was performed without a priori knowledge, where all gene mutations have the same probability of occurring and don't affect the probability of alterations in other genes. This expected number of randomly right mutations R_{rand} (number of randomly wrong simulated W_{rand} analogue) was calculated by:

$$R_{rand}(x) = \sum_{x=0}^{n} (1000 * \frac{n-x}{10-h-x}) \tag{10}$$

As shown in Table 5, prospective analysis of this data set provides simulation results that are closer to reality than random events and is, as such, a legitimate approach for further research in progressive tumour simulation. In order to better understand this, it should be said that the quality and results of the trained and tested data depend not only on the algorithms used, but also on the quality, quantity and other factors (such as homogeneity) of the data sets. As also described in Diaz-Colunga et al. [8], more specific raw data sets of highly defined scenarios (as the scenario demonstrated above) are required. Furthermore, genetic or biomedical processes in this context are not deterministic but follow stochastic processes. This means that genetic alterations of an existing tumour can develop with different probabilities in different directions, without having only one possible genetic pathway.

Table 5
Simulation and Verification of progressive simulated tumour data

Data set	n_L	n_T	e_{bef}	e_{sim}	R_{Sim}	W_{Sim}	R_{rand}	W_{rand}	OR
Jee Lung*	474	34	29	43	13129	29871	6116.27	36883.73	2.65
Lengel Lung*	1989	25	34	32	7121	24879	5155.952	26844.05	1.49

* Note: $\lambda = 0.01$, *Ten most frequent genes of the data sets according to *cbioportal* were used. n_L: number of data points learned. n_T: number of data points tested. e_{bef}: events that happened before simulation. e_{sim}: number of simulated events. R_{Sim}: number of events simulated right. W_{Sim}: number of events simulated wrong. OR: Odds-Ratio.

5.4 Comparison of prospective simulated and real tumour data and possible prognostic impact in lung cancer

The following Section aims to provide a comparative analysis of simulated tumour data and real tumour data with regard to their possible prognostic impact. Therefore, a model Θ is derived from a study and its data set. The progressive tumour simulation is based on a different study, with the objective of utilising this data to ascertain its prognostic value.

The initial stage of the Section 5.4.1 involves the calculation and comparison of simulations of the aforementioned prospective simulated tumour data and real tumour data. Subsequent to this, the potential gene mutation in lung tumours of the gene *STK11 / LKB11* in tumours is examined for survival analysis (see Section 5.4.2).

Furthermore, the final Section 5.4.3 of the text presents a statistical analysis and survival graphs, which evaluate the potential impact of the gene mutation in question. This analysis is contextualised within the broader framework of the significance of other mutated genes and patient factors.

In order to achieve this objective, an extended model of the MHN was selected, as published in *Python* (as described in Section 3.2), with $\lambda = 0.0013$ and 40 analysed genes.

The first *Jordan Lung* data set (n=764 with 40 genes, as outlined in Section 4.4) with advanced (in this case metastasised) lung adenocarcinomas was employed for the purpose of training an MHN-Model (Θ-Matrix). Based on the aforementioned learned model Θ, a second data set is prepared for test simulations. This is the *Jee Lung* data set, which contains genetic sets of metastatic non-small cell lung cancer (NSCLC) but comprises multiple samples collected at different times or from other locations (n=532 simulations; see Section 4.4 for details). The resulting simulated set comprises 532 simulations between two evolutionary points.

5.4.1 Prospective simulated tumour and real tumour data

The initial analysis (see Table 6) was conducted in a manner analogous to that described in Section 5.3, with the exception that the temporal order of the test set was defined solely based on the number of mutated genes. A simulation was conducted with 1000 iterations using the aforementioned gene set of 532 real tumour data, incorporating both earlier and later gene sets. The number of mutations that occurred in reality, represented by e_{sim}, was used as the basis for the simulation.

Table 6
Prospective simulated tumour data and real tumour data

Data set	n_L	n_T	e_{bef}	e_{sim}	R_{Sim}	W_{Sim}	R_{rand}	W_{rand}	OR
Jee Lung	764	532	848	1203	204895	998105	80755.68	1122244	2.85

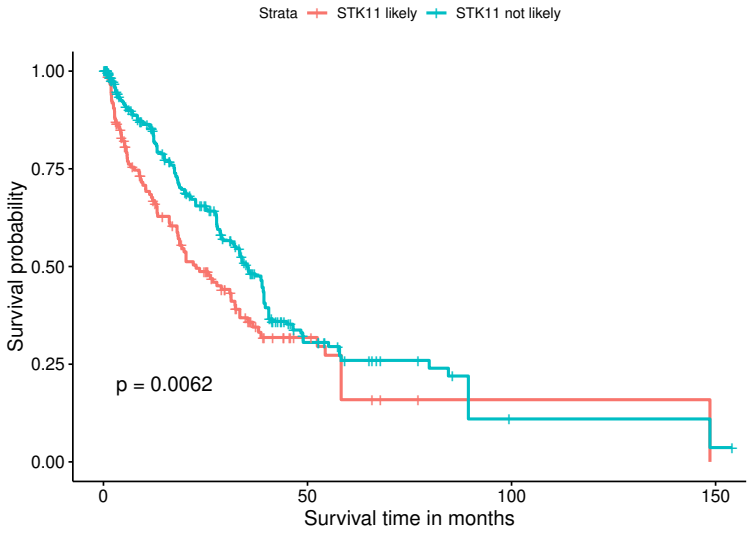
Note: $\lambda = 0.01$, 40 most mutated genes of *Jordan Lung* were used. n_L: number of data points learned by *Jordan Lung*. n_T: number of data points tested on *Jee Lung*. e_{bef}: events that happened before simulation. e_{sim}: number of simulated events. R_{Sim}: number of events simulated right. W_{Sim}: number of events simulated wrong. OR: Odds-Ratio. p-Value (Fisher's exact test/Chi-squared test) $< 2.2e-16$.

5.4.2 Survival analysis in tumours with a likely mutation in *STK11 / LKB1*

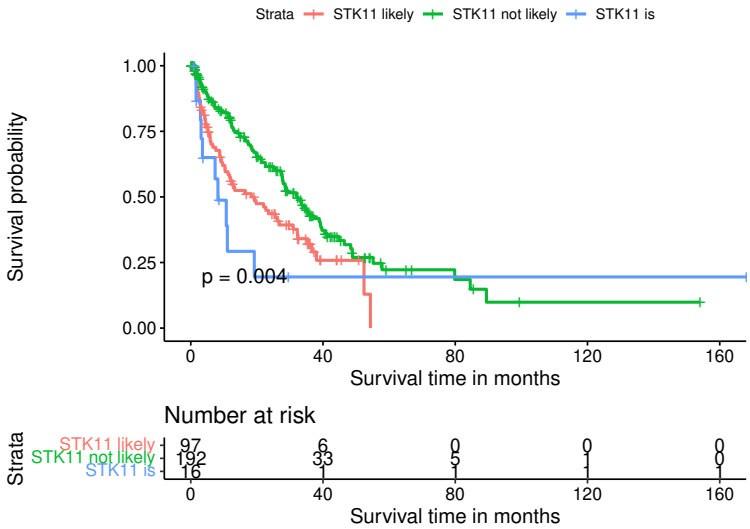
Subsequent to the aforementioned simulations and for the examination of the gene *STK11 / LKB1*, a further study was initiated. The gene *STK11 / LKB1* (referred to as *STK11*) is currently the subject of research in the field of non-small cell lung cancer (NSCLC) due to its association with a lack of response to immunotherapy in some patients [52, 53]. Consequently, a change in therapy regimen is being considered [54, 55]. The gene has been used to divide patients into two groups:
The first group encompasses instances where the occurrence of the event is deemed probable within the simulation models (with a probability greater than $> 20\%$, referred to as *STK11 probable* or *STK11 likely*). The second group comprises cases where the event is not probable (with a probability of $\leq 20\%$, referred to as *STK11 not probable* or *STK11 not likely*).
Within each group, different samples from the same patient have been included or excluded in the calculated Kaplan-Meier curves. At the time of sequencing, it is unclear whether additional samples will be forthcoming or if more samples will follow. Consequently, different approaches were simulated in which the same patient may be represented in the same or different groups multiple times, never, or once (see Figures 5 and 6). For more information, please see the supplementary data with Figures 15 and 16.
For Figure 5, the number of simulated events $e_{sim/tumour}$ was chosen as the number of events in real data. As the simulated number of events per tumour set, $e_{sim/tumour}$, that occurred in the real-world data set was not uniform and could distort the results, another simulation with real-world survival data was performed with $e_{sim/tumour} = 5$ (see Figure 6).
In Figures 5a and 6a, different samples of the same patients could be represented in different groups. Each group can have more than one sample from the same patient.
Figures 5b and 6b show data, where patients with data points in different groups are filtered. Within a group, a patient is represented once. Please note that tumours in which *STK11* is already mutated (referred to as *STK11 is*) are included (see below for further details).
As the Gene *STK11* has the capacity to mutate either before the initial genetic analysis or the first data point (referred to as *STK11 early* or *STK11 is*) or in the subsequent data point (referred to as *STK11 late*), Kaplan-Meier curves were constructed for the purpose of comparison. Figure 7 shows a not significant difference between *STK11 late* and *STK11 likely* data. Whereas Figure 8 provides an overview of the four groups (*STK11 early, late, unlikely and likely*).

Figure 5

The Kaplan-Meier curves illustrate forward simulations with a learned model Θ from Jordan Lung Data set with real tumour data of Jee-Lung, as described in Section 5.4. Number of events individually simulated per tumour $e_{sim/tumour}$. p-value calculated by log-rank test.



(a) $e_{sim/tumour}$ *as in real-world data. Patients with data points in different groups are not filtered. In each group, a patient can have multiple samples. As the Gene STK11 can mutate before the first genetic analysis or first data point (referred to as STK11 early or STK11 is) or in the later data point (referred to as STK11 late). Kaplan-Meier curves were constructed for comparison and are shown in Figure 7 and Figure 8. STK11-probable ($> 20\%$,n=149), STK11 not probable ($\leq 20\%$, n=330), **$p \leq 0.01$.*

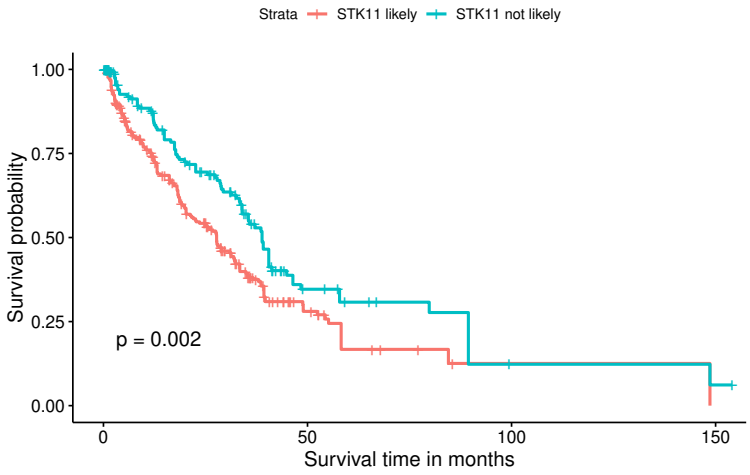


(b) $e_{sim/tumour}$ *as in real-world data. Patients with data points in different groups are filtered. In a group, each patient can have one sample only. STK11 probable ($> 20\%$,n=97), STK11 not probable ($\leq 20\%$, n=192), STK11 is mutated ($n = 16$), **$p \leq 0.01$.*

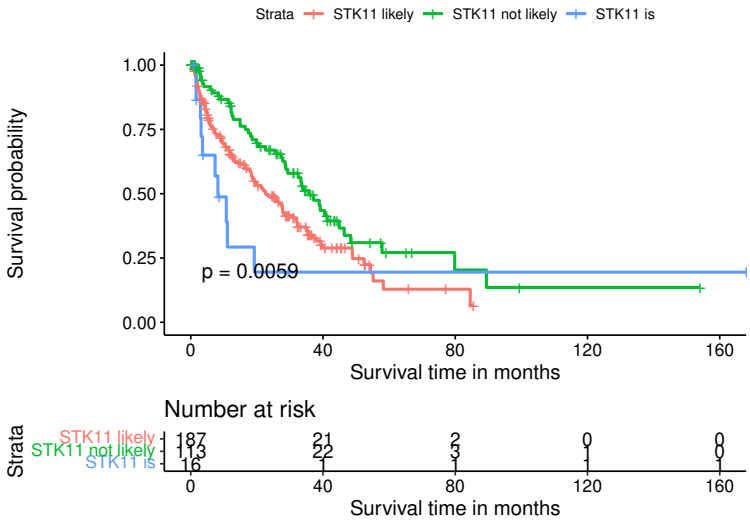
.

Figure 6

The Kaplan-Meier curves illustrate forward simulations with a learned model Θ from Jordan Lung Data set with real tumour data of Jee-Lung, as described in Section 5.4. Number of events individually simulated per tumour $e_{sim/tumour}$. p-value calculated by log-rank test.



(a) $e_{sim/tumour} = 5$. *Patients with data points in different groups are not filtered. In each group, a patient can have multiple samples. STK11 probable ($> 20\%$,n=282), STK11 not probable ($\leq 20\%$, n=197), **$p \leq 0.01$.*



(b) $e_{sim/tumour} = 5$. *Patients with data points in different groups are filtered. In a group, each patient can have one sample only. STK11 probable ($> 20\%$,n=187), STK11 not probable ($\leq 20\%$, n=113), STK11 is mutated ($n = 16$)* **$p \leq 0.01$.

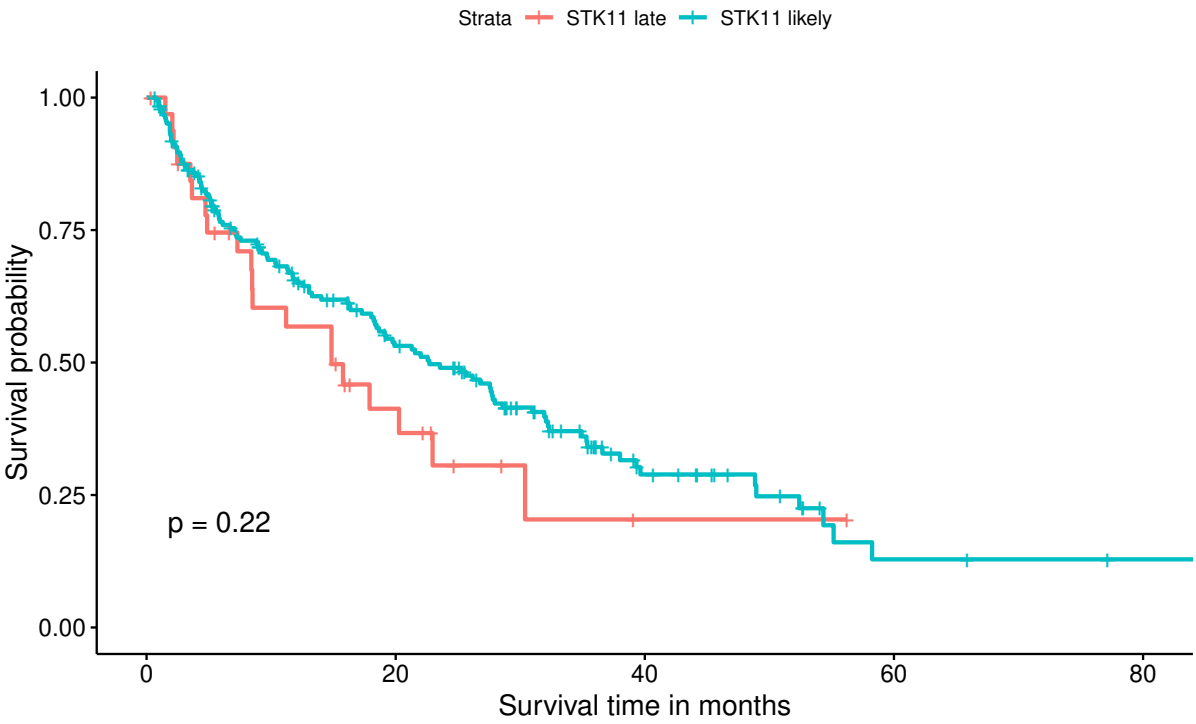


Figure 7

The following comparison is made between simulated data, in which STK11 is likely ($n = 187$) and the case in real data, in which STK11 mutates in a late data point ($n = 34$). The p-value ($p = 0.22$) shows no significant difference.

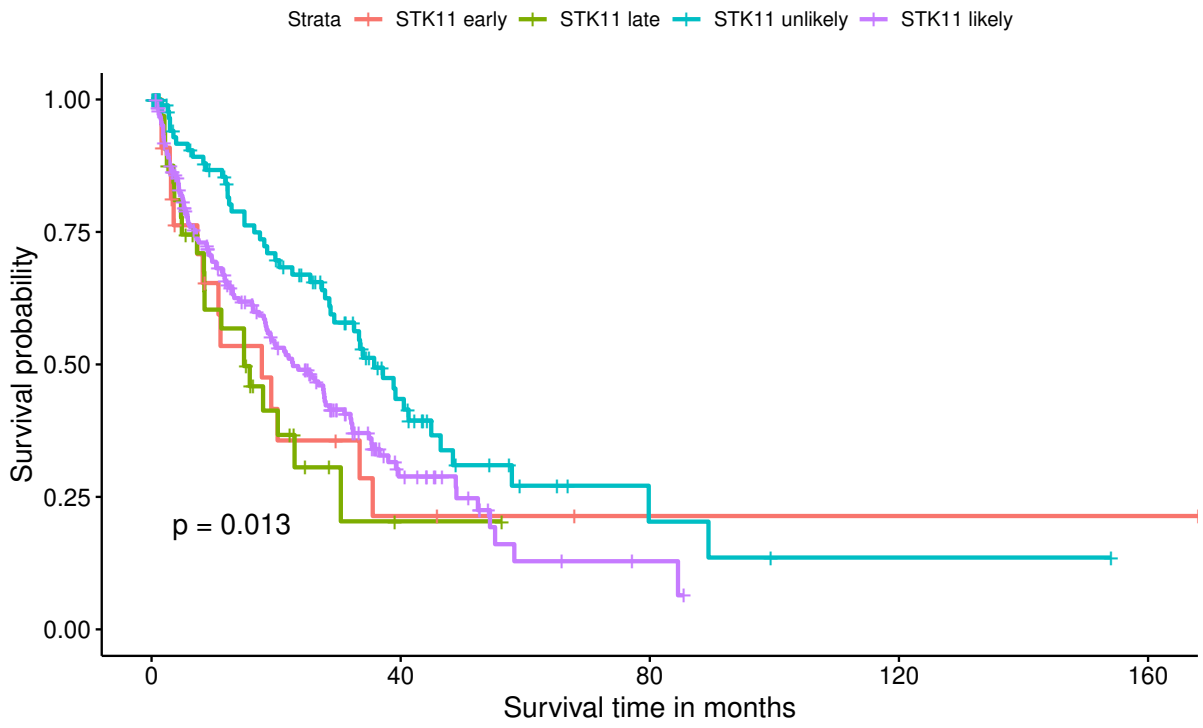


Figure 8
*The following comparison was made between the early ($n = 24$) and late ($n = 34$) STK11-mutation groups, as well as the simulated probable ($n = 187$) and not probable ($n = 113$) groups. *$p \leq 0.05$.*

As demonstrated by the data presented above, which shows statistical significance in accordance with the hypothesis that a likely *STK11* mutation in our model is a prognostic factor, further studies are to be conducted with the following constraints:
The progression model with $e_{sim/tumour} = 5$ appears to be a more appropriate approach, particularly in light of the uncertainty surrounding the number of mutations within a bulk tumour that are responsible for its further development. For further details on this topic, readers are referred to Section 4.1. Consequently, the progression model with five mutations will be utilised.
The grouping model, in which patients are permitted to have only one sample within a single group, appears to be the most statistically appropriate model. For this reason, it will be used for further investigations.

5.4.3 Possible *STK11* / *LKB1* mutations in the context of statistical survival with other genes and patient factors

The following investigation will firstly establish the context of *STK11* mutations in relation to other genes and survival rates. Subsequently, it will be contextualised within a statistical framework, with patient factors being taken into consideration.
In the initial phase of the investigation, the genes that exhibited the most significant impact on *STK11* were selected, and a univariate Cox regression analysis was conducted, as illustrated in Table 7. For the genes that demonstrated significance in the initial analysis, a multivariate Cox regression was conducted, and the results are presented in Table 8.

Table 7
Univariate Cox regression in specified genes and probable STK11-mutations.

Gene	Beta	HR (95% CI for HR)	Wald-test	p-value
STK11-likely	0.47	1.6 (1.1-2.3)	7.2	*0.0073* **
KRAS	0.45	1.6 (1.1-2.2)	7.1	*0.0077* **
KEAP1	1.2	3.3 (1.5-7.6)	8.2	*0.0043* **
EGFR	-0.46	0.63 (0.44-0.91)	6.1	*0.014* *
TP53	0.39	1.5 (1.1-2)	5.8	*0.016* *
PTPRD	-15	3e-07 (0-Inf)	0	0.99
MET	-0.09	0.91 (0.51-1.6)	0.09	0.77
SMARCA4	0.62	1.9 (0.26-13)	0.38	0.54
RBM10	0.62	1.9 (0.26-13)	0.38	0.54

Note: Significance codes: 0.001 '***', 0.01 '**', 0.05 '*'. Genes with the most impact on *STK11* derived the Θ-model from *Jordan Lung* were selected. The gene *EHB1* was dismissed, as no occurrence was found in *Jee Lung*.
Abbreviations: *Beta*: Beta coefficient, *95% CI for HR*: Hazard ratio (95 % confidence interval), *Wald-Test*: Wald Test.

Table 8
Multivariate Cox regression analysis in specified genes and probable STK11-mutations, for more details, see Table 11.

Gene	Coef	HR	SE(Coef)	Wald	p-value
STK11-likely	0.40429	1.49823	0.42696	0.947	0.34370
KRAS	0.44573	1.56163	0.20149	2.212	*0.02695* *
EGFR	0.09762	1.10254	0.42504	0.230	0.81835
KEAP1	1.34981	3.85668	0.48631	2.776	*0.00551* **
TP53	0.43001	1.53727	0.18301	2.350	*0.01879* *

Note: Significance codes: 0.001 '***', 0.01 '**', 0.05 '*'.
Abbreviations: *coef*: Coefficient (B), *HR*: Hazard ratio (exponential of the coefficient (B)), *SE(Coef)*: Standard error of the coefficient, *Wald*: Wald statistic value.

Given the finding that multiple genes demonstrated significance in a multivariate Cox regression, it is proposed that these genes should be explored individually with Kaplan-Meier curves in the context of possible *STK11*-mutations. As demonstrated in the study by Boeschen et al. [56], mutations in *KRAS*, *STK11*, and *KEAP1* exhibit significant co-occurrence in NSCLC, leading to a consequent reduction in the number of cases available for the subsequent analyses. It should be noted that some of the figures present data on a comparatively limited number of cases. However, these figures are incorporated for the sake of comprehensiveness.

- *KRAS* is shown in Figure 9 and Figure 10
- *KEAP1* is shown in Figure 11. It is notable that no cases were observed in which *KEAP1* was mutated and *STK11* was likely, and thus the case in which *KEAP1* was mutated and *STK11* was unlikely was included in Figure 11.
- *TP53* is shown in Figure 12 and 13
- *EGFR* In contrast to the other genes analysed, the presence of mutations in the epidermal growth factor receptor (*EGFR*) does not show any cases where *STK11* is likely and *EGFR* is mutated. Furthermore, a mutation in *EGFR* shows a worse survival curve than in the case where a *STK11* mutation is likely. It can be hypothesised that these genes are mutually exclusive (see Figure 14).

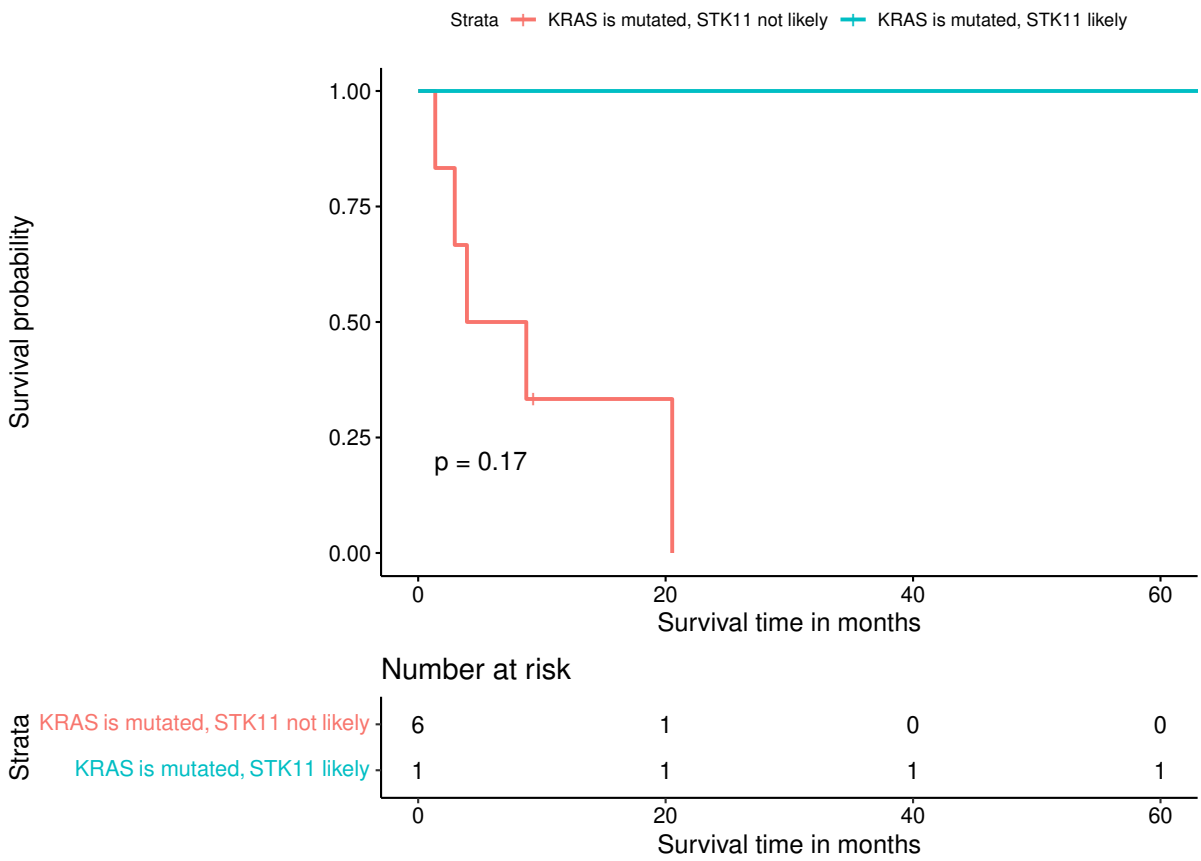


Figure 9

In this particular instance, the analysis indicates that KRAS is mutated. The data presented in the graphs demonstrates a high likelihood of a STK11-mutation ($n = 1$) and a low likelihood of a STK11-mutation ($n = 6$), as described in the preceding Section 5.4.

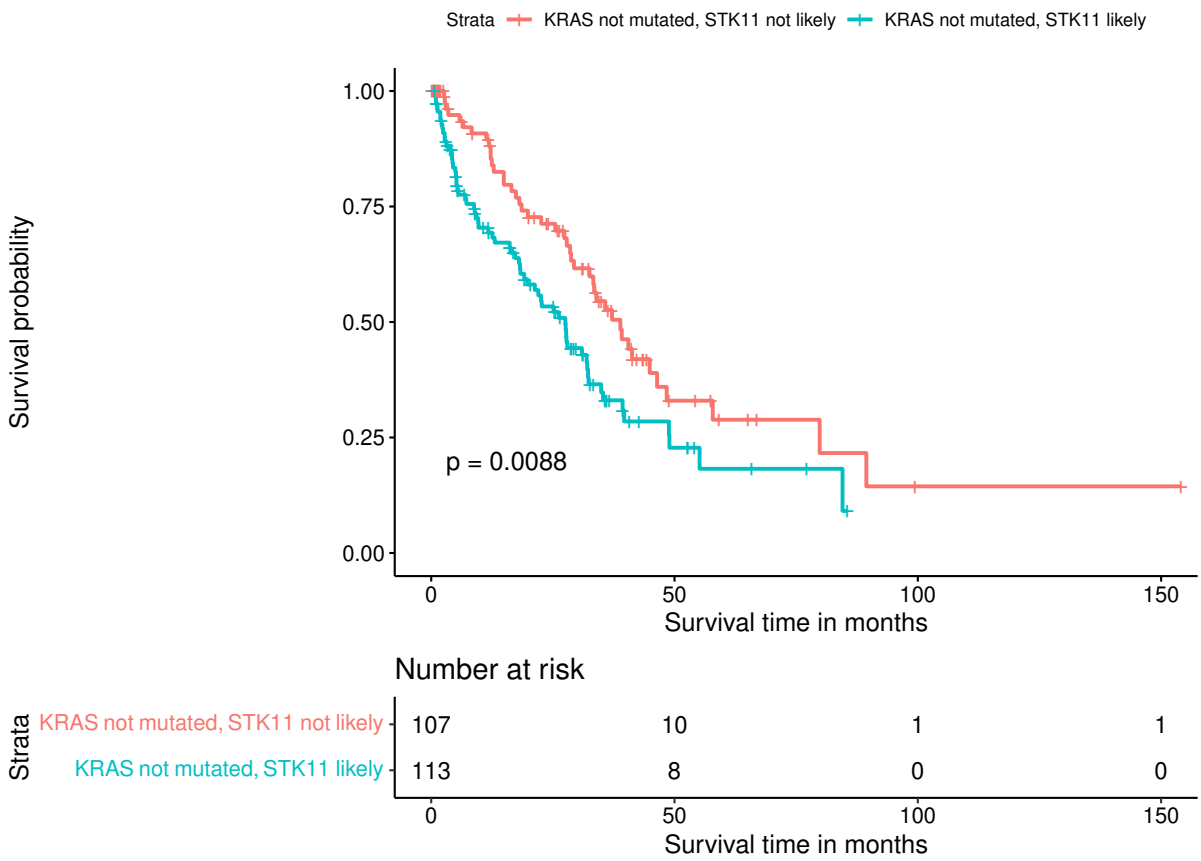


Figure 10

*In this particular instance, the analysis indicates that KRAS is not mutated. The data presented in the graphs demonstrates a high likelihood of a STK11-mutation ($n = 113$) and a low likelihood of a STK11-mutation ($n = 107$), as described in the preceding Section 5.4. **$p \leq 0.01$.*

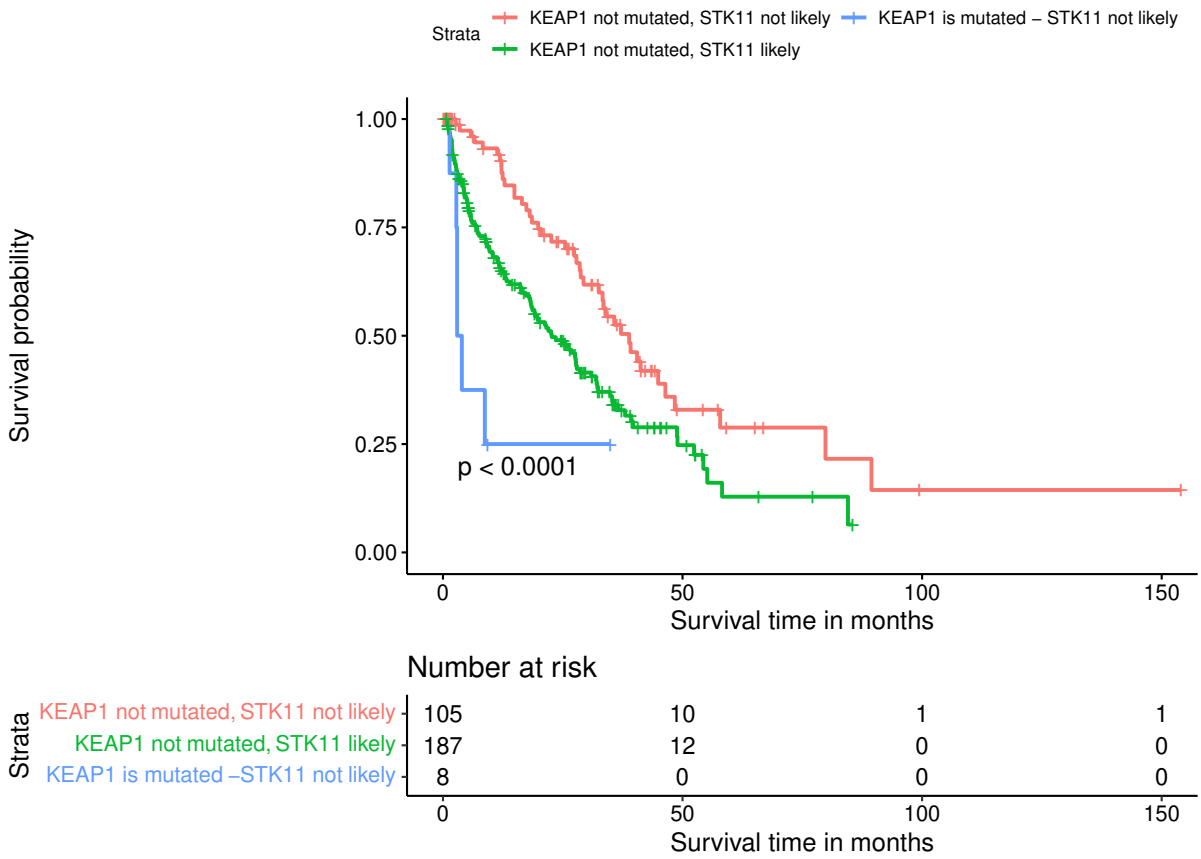


Figure 11

*In this particular instance, the analysis shows the gene KRAS. The data demonstrates that KEAP1 is not mutated, with a high likelihood of a STK11-mutation ($n = 187$) and a low likelihood of a STK11-mutation ($n = 105$). It also includes the case where KEAP1 is mutated and STK11 is likely ($n = 8$). ***$p \leq 0.001$.*

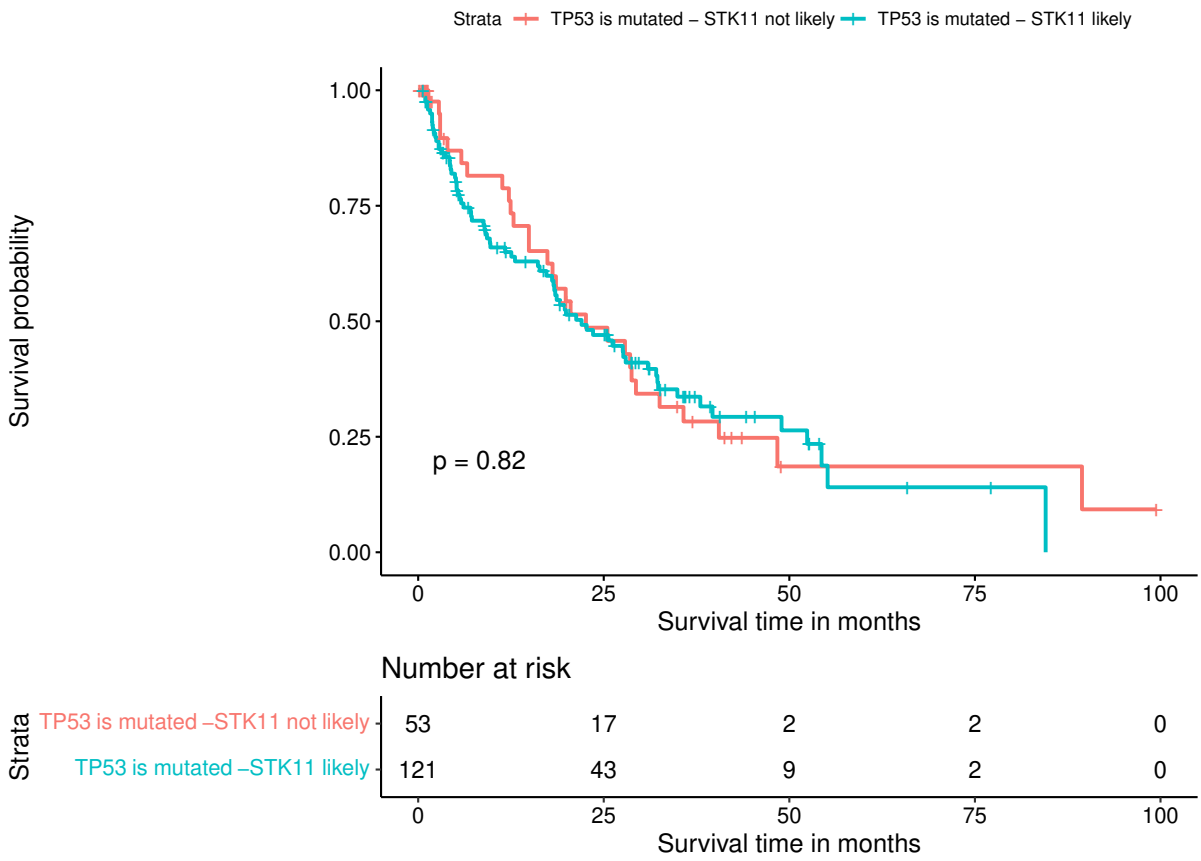


Figure 12

In this particular instance, the analysis shows the gene TP53. The data demonstrates, that TP53 is mutated, with a high likelihood of a STK11-mutation ($n = 121$) and a low likelihood of a STK11-mutation ($n = 53$).

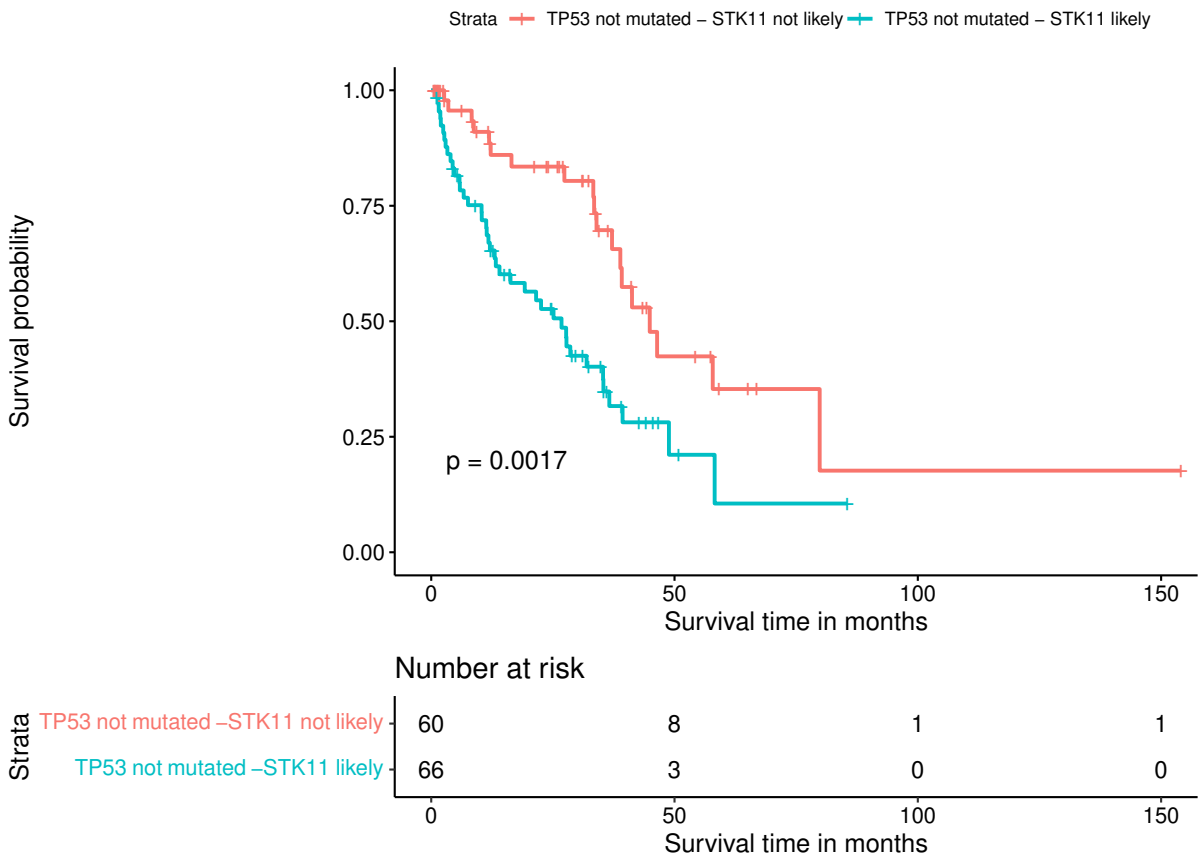


Figure 13

*In this particular instance, the analysis shows the gene TP53. The data demonstrates that TP53 is not mutated, with a high likelihood of a STK11-mutation ($n = 66$) and a low likelihood of a STK11-mutation ($n = 60$). **$p \leq 0.01$.*

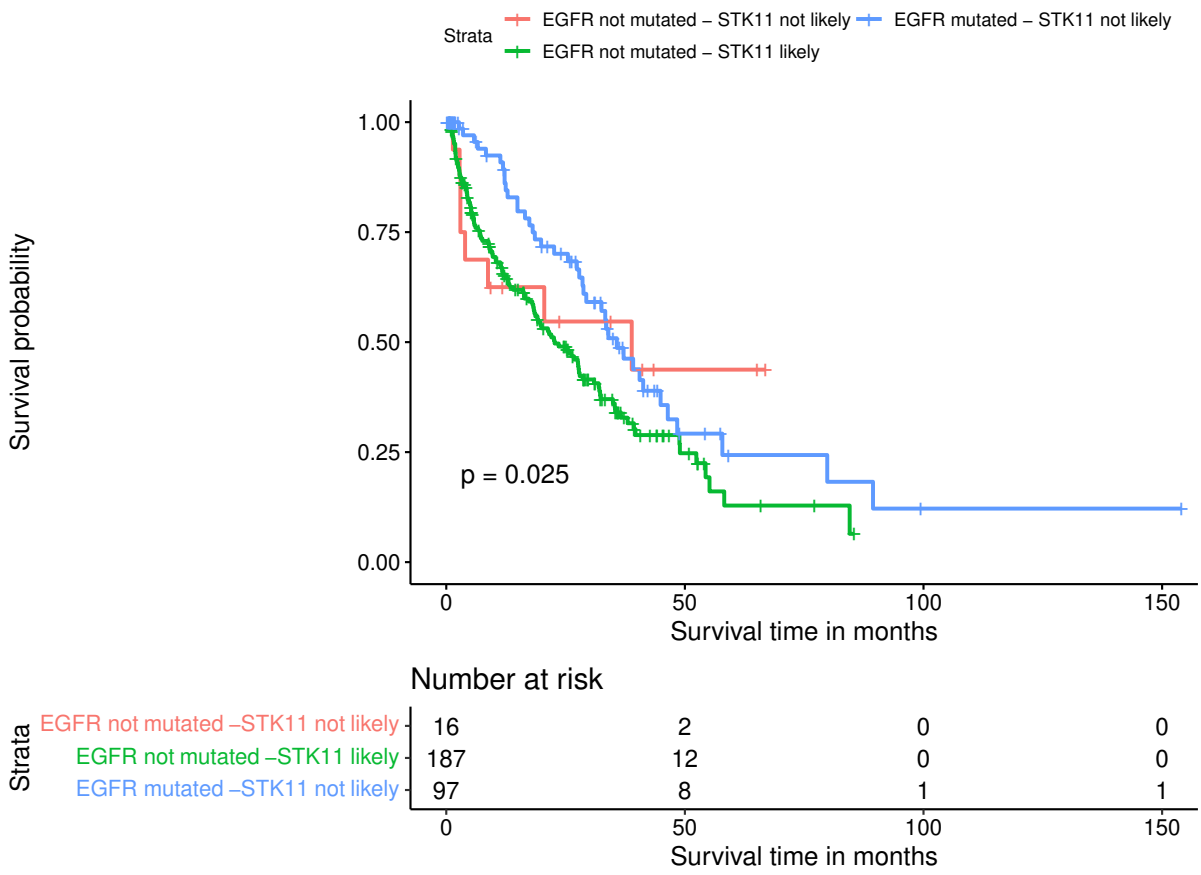


Figure 14

*In this particular instance, the analysis shows the gene EGFR. The data demonstrates that EGFR is not mutated, with a high likelihood of a STK11-mutation ($n = 187$) and a low likelihood of a STK11-mutation ($n = 16$). It also includes the case where EGFR is mutated and STK11 is not likely ($n = 97$). *$p \leq 0.05$, p-Value without "EGFR not mutated - STK not likely": **$p = 0.0072$.*

The last investigation will employ univariate Cox regression in probable *STK11*-mutations and patient data. Initially, the base factors of patients and likely *STK11*-mutations were selected, as illustrated in Table 9. For the factors that demonstrated significance in the first analysis, a multivariate Cox regression was conducted and is presented in Table 10.

Table 9
Univariate Cox regression in probable STK11-mutations and patient factors.

Gene	Beta	HR (95% CI for HR)	Wald-test	p-value
STK11	0.46	1.6 (1.1-2.2)	7.1	*0.0079* **
Sex	-0.23	0.8 (0.58-1.1)	2.1	0.15
Prior treatment	0.9	2.5 (1.7-3.6)	22	*3.1e-06* ***
Smoking status	-0.3	0.74 (0.53-1)	3.2	0.073
Age current	0.0076	1 (0.99-1)	1.2	0.27

Note: Significance codes: 0.001 '***', 0.01 '**', 0.05 '*'.
Abbreviations: *Beta*: Beta coefficient, *95% CI for HR*: Hazard ratio (95 % confidence interval), *Wald-Test*: Wald Test.

Table 10
Multivariate Cox regression analysis in probable STK11-mutations and patient factors, for more information see Table 12.

Gene	Coef	HR	SE(Coef)	Wald	p-value
STK11	0.4032	1.4967	0.1752	2.302	*0.0213* *
Prior treatment	0.8670	2.3799	0.1931	4.490	*7.12e-06* ***

Note: Significance codes: 0.001 '***', 0.01 '**', 0.05 '*'.
Abbreviations: *coef*: Coefficient (B), *HR*: Hazard ratio (exponential of the coefficient (B)), *SE(Coef)*: Standard error of the coefficient, *Wald*: Wald statistic value.

6 Conclusion and Discussion

In conclusion, the feasibility and validation of a machine learning-based simulation utilising MHN and a modified Gillespie algorithm have been demonstrated. Furthermore, a scientific prototype simulation software interface for individual tumour evolution simulation has been developed, and a statistically significant prognostic prediction of tumour evolution has been performed.
In order to evaluate and validate the simulation software, three different types of validation and one kind of prognostic prediction were conducted:

- First, genetic sets of tumours randomly generated on a learned model were relearned and then compared with the original learned data set. This was performed on the four different cancer types, which were used by Schill et al. [21] (originally published by Baudis et al [35]).

- Second, tumours of the same type (breast cancer) were prepared from three different data sets (*Razavi* [40], *Metabric* [37, 38, 39], *SMC* [41]). Each data set was learned, random tumours were generated from this learned model and relearned, and each data set from a different study was compared at least once with data from another study.

- Third, a model was defined in which tumours with tumour data with at least two different evolutionary developmental points were prepared. *Lengel* [44] and *Jee* [43], both lung cancer data sets, were used as baselines. Here, tumours with less than two evolutionary points were used to learn a model; this model was then used with the earlier data sets with more than one evolutionary point. Progressive simulations were run on these tumours and then compared with the data sets of their biological genetic descendants.

- The prognostic prediction entailed applying a learned model Θ derived from a lung cancer data set (*Jordan Lung* [42]) to the progressive tumour simulation of another lung cancer data set (*Jee Lung* [43]). A statistically significant prognostic value was demonstrated for the prediction of the gene *STK11 / LKB1*. This gene is currently subject of research in the field of non-small cell lung cancer (NSCLC) due to its association with a lack of response to immunotherapy in some patients, where a change in therapy regimen is being considered.

The above-described results are promising, but there are limitations:

- A reduced and simplified gene set was used
- The data sets originate from (now) retrospective different studies of the same or different cancer (sub)types, including different patient cohorts and, in some cases, small numbers of suitable subjects. In addition, different pipelines were used for the genetic sequencing of the data sets, with different sets of genes being analysed.
- Although the prediction of *STK11* shows promising results, it's possible that the actual results will not only show a single possible gene mutation, but rather a network of genes, which could also result in a medical prediction score rather than a gene mutation prediction.

It is further noted that this type of genetic analysis and simulation can be applied to other areas of emerging resistance, such as viral diseases (e.g. HIV, or the development of more mutation-stable vaccines) and multidrug-resistant outbreaks.

6.1 Supplementary information

Source code is available on: `https://github.com/spang-lab/MHN-Gill-Sim`

Table 11
Further information to Table 8: Multivariate Cox regression analysis with probable STK11-mutations and specified genetic mutations.

Statistical test	p-value
Likelihood ratio test= 24.18 on 5 df	$p = 2e-04$
Wald test = 25.01 on 5 df	$p = 1e-04$
Score (logrank) test = 26.68 on 5 df	$p = 7e-05$

Note: Concordance= 0.631 (se = 0.022).

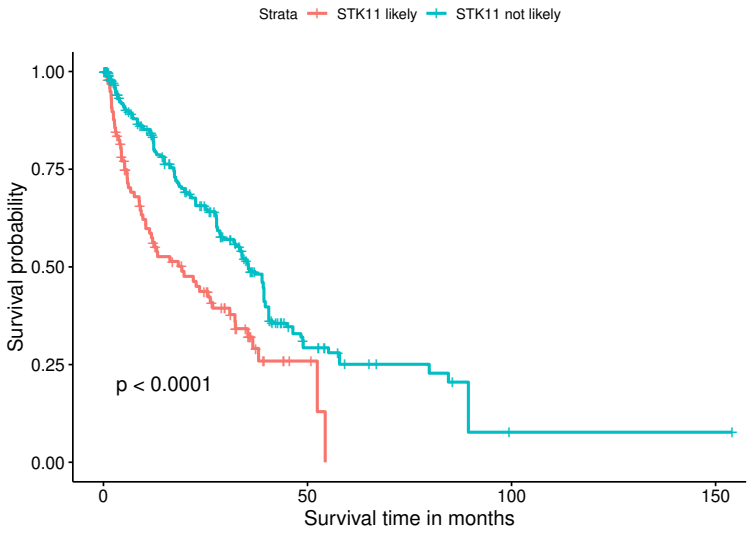
Table 12
Further information to Table 10: Multivariate Cox regression analysis with probable STK11-mutations and patient factors.

Statistical test	p-value
Likelihood ratio test= 30.28 on 2 df	$p = 3e-07$
Wald test = 27.08 on 2 df	$p = 1e-06$
Score (logrank) test = 28.46 on 2 df	$p = 7e-07$

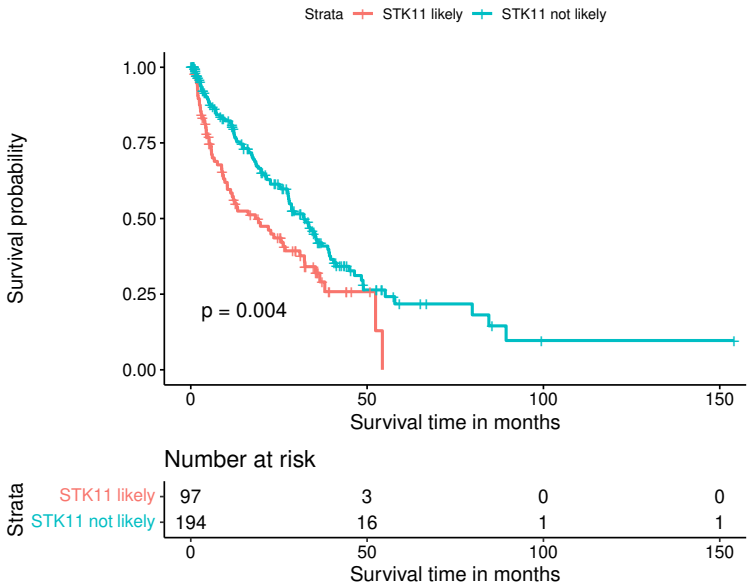
Note: Concordance= 0.644 (se = 0.02).

Figure 15

The Kaplan-Meier curves illustrate forward simulations with a learned model Θ from Jordan Lung Data set with real tumour data of Jee-Lung, as described in Section 5.4. Number of events individually simulated per tumour $e_{sim/tumour}$. p-value calculated by log-rank test.



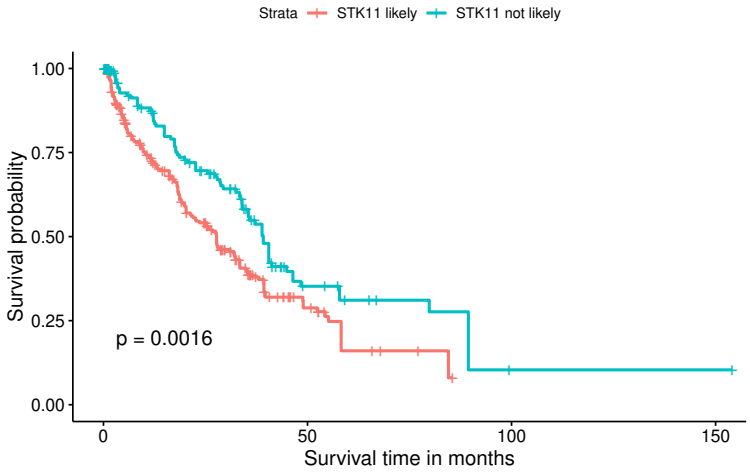
(a) $e_{sim/tumour}$ *as in real-world data. Patients with data points in both groups are filtered. In each group, a patient can have multiple samples. STK11 probable* ($> 20\%$*,n=98), STK11 not probable* ($\leq 20\%$*, n=286),* ***$p \leq 0.001$.



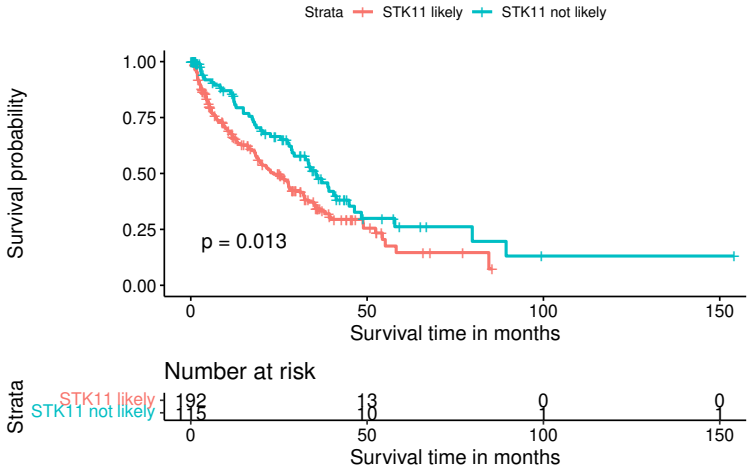
(b) $e_{sim/tumour}$ *as in real-world data. Patients with data points in both groups are filtered. In a group, each patient can have one sample only. STK11 probable* ($> 20\%$*, n=97), STK11 not probable* ($\leq 20\%$*, n=194),* ****$p \leq 0.01$.

Figure 16

The Kaplan-Meier curves illustrate forward simulations with a learned model Θ from Jordan Lung Data set with real tumour data of Jee-Lung, as described in Section 5.4. Number of events individually simulated per tumour $e_{sim/tumour}$. p-value calculated by log-rank test.



(a) $e_{sim/tumour} = 5$. *Patients with data points in both groups are filtered. In each group, a patient can have multiple samples. STK11 probable ($> 20\%$, n=252), STK11 not probable ($\leq 20\%$, n=184), *$p \leq 0.01$.*



(b) $e_{sim/tumour} = 5$. *Patients with data points in both groups are filtered. In a group, each patient can have one sample only. STK11 probable ($> 20\%$, n=192), STK11 not probable ($\leq 20\%$, n=115), *$p \leq 0.05$.*

References

1. Siegel RL, Giaquinto AN, and Jemal A. Cancer statistics, 2024. CA: a cancer journal for clinicians 2024; 74:12–49. DOI: 10.3322/caac.21820
2. Vasan N, Baselga J, and Hyman DM. A view on drug resistance in cancer. Nature 2019; 575:299–309. DOI: 110.1038/s41586-019-1730-1
3. Jin H, Wang L, and Bernards R. Rational combinations of targeted cancer therapies: background, advances and challenges. Nature Reviews Drug Discovery 2023; 22:213–34. DOI: 10.1038/s41573-022-00615-z
4. Greaves M and Maley CC. Clonal evolution in cancer. Nature 2012; 481:306. DOI: 10.1038/nature10762
5. Nowell PC. The clonal evolution of tumor cell populations. Science 1976; 194:23–8. DOI: 10.1126/science.959840
6. Orr HA. The genetic theory of adaptation: a brief history. Nature Reviews Genetics 2005; 6:119–27. DOI: 10.1038/nrg1523
7. Greaves M. Evolutionary determinants of cancer. Cancer discovery 2015; 5:806–20. DOI: 10.1158/2159-8290.CD-15-0439
8. Diaz-Colunga J and Diaz-Uriarte R. Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next? PLOS Computational Biology 2021 Dec; 17:1–23. DOI: 10.1371/journal.pcbi.1009055
9. Albini A and Sporn MB. The tumour microenvironment as a target for chemoprevention. Nature Reviews Cancer 2007; 7:139–47. DOI: https://doi.org/10.1038/nrc2067
10. Brock A, Chang H, and Huang S. Non-genetic heterogeneity—a mutation- independent driving force for the somatic evolution of tumours. Nature Reviews Genetics 2009; 10:336–42. DOI: 10.1038/nrg2556
11. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, and Gerlinger M. Cancer evolution and the limits of predictability in precision cancer medicine. Trends in cancer 2016; 2:49–63. DOI: 10.1016/j.trecan.2015.11.003
12. Ramazzotti D, Angaroni F, Maspero D, Ascolani G, Castiglioni I, Piazza R, Antoniotti M, and Graudenzi A. Longitudinal cancer evolution from single cells. bioRxiv 2020 :2020–1. DOI: 10.1101/2020.01.14.906453
13. Gerstung M, Baudis M, Moch H, and Beerenwinkel N. Quantifying cancer progression with conjunctive Bayesian networks. Bioinformatics 2009; 25:2809–15. DOI: 10.1093/bioinformatics/btp505

14. Gerstung M, Eriksson N, Lin J, Vogelstein B, and Beerenwinkel N. The temporal order of genetic and pathway alterations in tumorigenesis. PloS one 2011; 6:e27136. DOI: 10.1371/journal.pone.0027136

15. Montazeri H, Kuipers J, Kouyos R, Böni J, Yerly S, Klimkait T, Aubert V, Günthard HF, Beerenwinkel N, and Study SHC. Large-scale inference of conjunctive Bayesian networks. Bioinformatics 2016; 32:i727–i735. DOI: 10.1371/journal.pcbi.1008363

16. Szabo A and Boucher KM. Oncogenetic trees. *Handbook of cancer models with applications*. USA: World Scientific, 2008 :1–24. DOI: doi.org/10.1142/6677

17. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, and Schäffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of computational biology 1999; 6:37–51. DOI: 10.1089/cmb.1999.6.37

18. Ramazzotti D, Caravagna G, Olde Loohuis L, Graudenzi A, Korsunsky I, Mauri G, Antoniotti M, and Mishra B. CAPRI: efficient inference of cancer progression models from cross-sectional data. Bioinformatics 2015; 31:3016–26. DOI: 10.1093/bioinformatics/btv296

19. Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R, De Sano L, Mauri G, Moreno V, Antoniotti M, and Mishra B. Algorithmic methods to infer the evolutionary trajectories in cancer progression. Proceedings of the National Academy of Sciences 2016; 113:E4025–E4034. DOI: 10.1073/pnas.1520213113

20. Loohuis LO, Caravagna G, Graudenzi A, Ramazzotti D, Mauri G, Antoniotti M, and Mishra B. Inferring tree causal models of cancer progression with probability raising. PloS one 2014; 9:e108358. DOI: 10.1371/journal.pone.0108358

21. Schill R, Solbrig S, Wettig T, and Spang R. Modelling cancer progression using mutual hazard networks. Bioinformatics 2020; 36:241–9. DOI: 10.1093/bioinformatics/btz513

22. Beerenwinkel N, Eriksson N, and Sturmfels B. Conjunctive bayesian networks. en. Bernoulli 2007; 13:893–909. DOI: 10.3150/07-BEJ6133

23. Deng Y, Luo S, Deng C, Luo T, Yin W, Zhang H, Zhang Y, Zhang X, Lan Y, Ping Y, et al. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. Briefings in Bioinformatics 2019; 20:254–66. DOI: 10.1093/bib/bbx109

24. Schill R, Klever M, Lösch A, Hu YL, Vocht S, Rupp K, Grasedyck L, Spang R, and Beerenwinkel N. Overcoming Observation Bias for Cancer Progression Modeling. bioRxiv 2023 :2023–12. DOI: 10.1101/2023.12.03.569824

25. Schill R, Klever M, Rupp K, Hu YL, Lösch A, Georg P, Pfahler S, Vocht S, Hansch S, Wettig T, et al. Reconstructing Disease Histories in Huge Discrete State Spaces. KI-Künstliche Intelligenz 2024 :1–11. DOI: 10.1007/s13218-023-00822-9

26. Rupp K, Lösch A, Hu YL, Nie C, Schill R, Klever M, Pfahler S, Grasedyck L, Wettig T, Beerenwinkel N, et al. Modeling metastatic progression from cross-sectional cancer genomics data. bioRxiv 2024 :2024–1. DOI: 10.1101/2024.01.30.577989

27. Vocht S et al. mhn: A Python Package for Analyzing Cancer Progression with Mutual Hazard Networks. Available from: https://learnmhn.readthedocs.io/en/latest/index.html [Accessed on: 2025 Mar 21]

28. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry 1977; 81:2340–61. DOI: 10.1021/j100540a008

29. Lu T, Volfson D, Tsimring L, and Hasty J. Cellular growth and division in the Gillespie algorithm. IEE Proceedings-Systems Biology 2004; 1:121–8. DOI: https://doi.org/10.1049/sb:20045016

30. Rao CV and Arkin AP. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. The Journal of chemical physics 2003; 118:4999–5010. DOI: https://doi.org/10.1063/1.1545446

31. Bernstein D. Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics 2005; 71:041103. DOI: https://doi.org/10.1103/PhysRevE.71.041103

32. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 2013; 6:pl1–pl1. DOI: 10.1126/scisignal.2004088

33. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. 2012. DOI: 10.1158/2159-8290.CD-12-0095

34. McDonald TO, Chakrabarti S, and Michor F. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. Nature genetics 2018; 50:1620–3. DOI: 10.1038/s41588-018-0217-6

35. Baudis M and Cleary ML. Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 2001; 17:1228–9. DOI: https://10.1093/bioinformatics/17.12.1228

36. Rainer J. EnsDb. Hsapiens. v86. Bioconductor 2017. DOI: 10.18129/B9.bioc.EnsDb.Hsapiens.v86

37. Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature communications 2016; 7:11479. DOI: 10.1038/ncomms11479

38. Rueda OM, Sammut SJ, Seoane JA, Chin SF, Caswell-Jin JL, Callari M, Batra R, Pereira B, Bruna A, Ali HR, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature 2019; 567:399–404. DOI: 10.1038/s41586-019-1007-8

39. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012; 486:346–52. DOI: 10.1038/nature10983

40. Razavi P, Chang MT, Xu G, Bandlamudi C, Ross DS, Vasan N, Cai Y, Bielski CM, Donoghue MT, Jonsson P, et al. The genomic landscape of endocrine-resistant advanced breast cancers. Cancer cell 2018; 34:427–38. DOI: 10.1016/j.ccell.2018.08.008

41. Kan Z, Ding Y, Kim J, Jung HH, Chung W, Lal S, Cho S, Fernandez-Banet J, Lee SK, Kim SW, et al. Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. Nature communications 2018; 9:1725. DOI: 10.1038/s41467-018-04129-4

42. Jordan EJ, Kim HR, Arcila ME, Barron D, Chakravarty D, Gao J, Chang MT, Ni A, Kundra R, Jonsson P, et al. Prospective comprehensive molecular characterization of lung adenocarcinomas for efficient patient matching to approved and emerging therapies. Cancer discovery 2017; 7:596–609. DOI: 10.1158/2159-8290.CD-16-1337

43. Jee J, Lebow ES, Yeh R, Das JP, Namakydoust A, Paik PK, Chaft JE, Jayakumaran G, Rose Brannon A, Benayed R, et al. Overall survival with circulating tumor DNA-guided therapy in advanced non-small-cell lung cancer. Nature Medicine 2022; 28:2353–63. DOI: 10.1038/s41591-022-02047-z

44. Lengel HB, Mastrogiacomo B, Connolly JG, Tan KS, Liu Y, Fick CN, Dunne EG, He D, Lankadasari MB, Satravada BA, et al. Genomic mapping of metastatic organotropism in lung adenocarcinoma. Cancer Cell 2023; 41:970–85. DOI: 10.1016/j.ccell.2023.03.018

45. Association AP et al. Publication manual of the american psychological association,(2020). 2019 :428. DOI: 10.1037/0000165-000

46. Therneau TM. A Package for Survival Analysis in R. R package version 3.8-3. 2024. DOI: 10.32614/CRAN.package.survival. Available from: https://CRAN.R-project.org/package=survival

47. Olsen LR. Creating Groups from Data. 2024. DOI: 10.32614/CRAN.package.groupdata2. Available from: https://cran.r-project.org/web/packages/groupdata2/index.html

48. Hadley W, Romain F, Lionel H, and Kirill Müllerand Davis V. A Grammar of Data Manipulation. 2023. DOI: 10.32614/CRAN.package.dplyr. Available from: https://cran.r-project.org/web/packages/dplyr/index.html

49. Alboukadel K, Kosinski M, Przemyslaw B, and Fabian S. Drawing Survival Curves using 'ggplot2'. 2024. DOI: 10.32614/CRAN.package.survminer. Available from: https://cran.r-project.org/web/packages/survminer/index.html

50. Sasaki K. Build your own Gillespie algorithm. Available from: https://github.com/karinsasaki/gillespie-algorithm-python/blob/master/build_your_own_gillespie_solutions.ipynb [Accessed on: 2023 Oct 14]

51. (MSK) MSKCC. cBioPortal FOR CANCER GENOMICS. 2023. Available from: https://www.cbioportal.org/study/summary?id=breast_msk_2018 [Accessed on: 2023 Dec 18]

52. Xu K, Lu W, Yu A, Wu H, and He J. Effect of the STK11 mutation on therapeutic efficacy and prognosis in patients with non-small cell lung cancer: a comprehensive study based on meta-analyses and bioinformatics analyses. BMC cancer 2024; 24:491. DOI: https://doi.org/10.1186/s12885-024-12130-y

53. Chen Y, Lee K, Woo J, Kim Dw, Keum C, Babbi G, Casadio R, Martelli PL, Savojardo C, Manfredi M, et al. Evaluating predictors of kinase activity of STK11 variants identified in primary human non-small cell lung cancers. Human Genetics 2025 :1–16. DOI: https://doi.org/10.1007/s00439-025-02726-0

54. Sumbly V and Landry I. Unraveling the role of STK11/LKB1 in non-small cell lung cancer. Cureus 2022; 14. DOI: 10.7759/cureus.21078

55. Mograbi B, Heeke S, and Hofman P. The importance of STK11/LKB1 assessment in non-small cell lung carcinomas. Diagnostics 2021; 11:196. DOI: 10.3390/diagnostics11020196

56. Boeschen M, Kuhn CK, Wirtz H, Seyfarth HJ, Frille A, Lordick F, Hacker UT, Obeck U, Stiller M, Bläker H, et al. Comparative bioinformatic analysis of KRAS, STK11 and KEAP1 (co-) mutations in non-small cell lung cancer with a special focus on KRAS G12C. Lung Cancer 2023; 184:107361. DOI: https://doi.org/10.1016/j.lungcan.2023.107361

7 Danksagung und Acknowledgements

Allen, die zum Gelingen der Dissertation beigetragen haben, möchte ich meinen herzlichen Dank aussprechen.
Mein besonderer Dank gilt Herrn Prof. Dr. rer. nat. Rainer Spang für die exzellente Betreuung während der Anfertigung meiner Dissertation.
Herrn Prof. Dr. med. Bernd Salzberger möchte sehr ich für die Unterstützung und Zweitkorrektur danken.
Des Weiteren möchte ich dem gesamten Team der Spang-Group für die interdiszplinäre Kooperation und die freundliche Zusammenarbeit während meiner Doktorarbeit und darüber hinaus meinen Dank aussprechen, insbesondere Rudolf Schill, Andreas Lösch und Linda Hu.
Es sei an dieser Stelle der Deutschen Forschungsgemeinschaft (DFG) für die Gewährung eines Promotionsstudiums gedankt.
Florian Volker Schlieckau danke ich für die Unterstützung und die Anregungen zu dieser Arbeit.
Nicht zuletzt bedanke möchte ich meiner Familie herzlich danken:
Meinen Eltern, meinem Bruder Gerhard, meiner Schwester Isabella, meiner Partnerin Linda und meiner Tochter Mila - möchte ich meinen herzlichsten Dank aussprechen. Ohne ihre Unterstützung und ihr Verständnis wäre diese Arbeit nicht möglich gewesen. Für ihre Unterstützung widme ich diese Arbeit meiner Tochter Mila.

Hinweis zur Nutzung von TCGA-Daten:
The results shown here are in part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.