Biomarkers for the response to immunotherapy in patients with non-small cell lung cancer

Muller, M.

9 Modelling Diagnostic Strategies to manage toxic adverse-events following cancer immunotherapy

F. van Delft*, M. Muller*, R. Langerak, H. Koffijberg, V. Retèl,
D. van den Broek, M. IJzerman.

Med Decis Making. 2021;41(6):693-705.

*Contributed equally

Abstract

Background: Although immunotherapy (IMT) provides significant survival benefits in selected patients, approximately 10% of patients experience (serious) immune-related adverse events (irAEs). The early detection of adverse events will prevent irAEs from progressing to severe stages, and routine testing for irAEs has become common practice. Because a positive test outcome might indicate a clinically manifesting irAE that requires treatment to (temporarily) discontinue, the occurrence of false-positive test outcomes is expected to negatively affect treatment outcomes. This study explores how the UPPAAL modeling environment can be used to assess the impact of test accuracy (i.e., test sensitivity and specificity), on the probability of patients entering palliative care within 11 IMT cycles.

Methods:A timed automata-based model was constructed using real-world data and expert consultation. Model calibration was performed using data from 248 non–small-cell lung cancer patients treated with nivolumab. A scenario analysis was performed to evaluate the effect of changes in test accuracy on the probability of patients transitioning to palliative care.

Results: The constructed model was used to estimate the cumulative probabilities for the patients' transition to palliative care, which were found to match real-world clinical observations after model calibration. The scenario analysis showed that the specificity of laboratory tests for routine monitoring has a strong effect on the probability of patients transitioning to palliative care, whereas the effect of test sensitivity was limited.

Conclusion: We have obtained interesting insights by simulating a care pathway and disease progression using UPPAAL. The scenario analysis indicates that an increase in test specificity results in decreased discontinuation of treatment due to suspicion of irAEs, through a reduction of false-positive test outcomes.

Background

Non–small-cell lung cancer (NSCLC) is associated with significant mortality. The incidence of lung cancer is estimated to be 11.6% of all new cancer diagnoses worldwide and is considered the leading cause of cancer-related mortality [299]. Until September 2018, in the Netherlands, the first-line treatment in the metastatic disease setting (stage IV NSCLC) was chemotherapy with platinum doublets in patients without a targetable mutation, and patients presenting with a targetable mutation (e.g., epidermal growth factor receptor or anaplastic lymphoma kinase mutation) received a targeted therapy (e.g., erlotinib, crizotinib, or gefitinib).[13] Since then, the treatment landscape has fundamentally changed with the introduction of immunotherapy, which now has become a standard treatment. Initially, nivolumab and pembrolizumab were approved in the second-line setting, but, since 2018, immunotherapy based on PD-L1 expression with or without added doublet chemotherapy has been the standard first-line treatment. Clinical studies have shown that about 20% to 40% of patients respond to immunotherapy, with substantially prolonged survival benefit. Despite the clinical benefits of the use of immunotherapy treatments, immunotherapy is also known to be associated with immunogenic reactions that severely affect treatment schedules and outcomes [300]. To manage immunotherapy-related adverse events (irAEs), suspicion of irAEs is determined using a standardized set of blood tests. However, no clear guidance is available on aspects such as the frequency of tests, required test specifications, or interpretation of results. This might result in suboptimal diagnostics and outcomes, in terms of survival [301].

At the Netherlands Cancer Institute, a diagnostic panel was implemented and routinely used at 2-weekly intervals. Test outcomes are used to aid clinical decision making on treatment continuation, to confirm the suspicion of an irAE, and to grade the severity of an irAE. The diagnostic panel aimed to detect irAEs in an early stage to improve clinical management and outcomes. The diagnostic kit covers a broad spectrum of blood markers. Some of these tests are used in a different setting (e.g., screening for irAE without clinical complaints) as compared with their original routine use, which could result in suboptimal performance in detecting irAEs. In addition, there are no data on the optimal frequency and use of diagnostic tests for irAEs. Hence, there is an interest in optimizing the diagnostic workflow to, for example, reduce unnecessary testing. Optimizing the test sequence using a prospective or retrospective study design to test different diagnostic setups would be unfeasible given the number of possible diagnostic strategies, time, and financial constraints. In such cases, it is possible to construct simulation models based on real-world data to evaluate diagnostic strategies aimed at the detection of irAEs on disease management and patient outcomes.

This study explores the use of UPPAAL to model diagnostic strategies. UPPAAL was developed by computer scientists at Uppsala University (Sweden) and Aalborg University (Denmark) [302]. The UPPAAL software uses a distributed modeling paradigm that allows for modeling a system using networks' of timed automata (TA). The UPPAAL tool enables modelers to construct networks of TA. These networks consist of a finite set of automata with real-valued clocks and constraints. Automata can be seen as a state in which a predefined process will be executed automatically (e.g., changing an integer according to a predefined function). Within the network, clock values increase with equal speed, and clock values can be compared with integers to control transitions between automata[303]. Moreover, in these networks of TA, communication channels are used to allow for multiple types of synchronization signals that allow for communication between different automata in the network. Generally, TA models consist of multiple templates, with each template containing a network of automata used to model a specific function. The communication channels allow different templates to communicate with and influence each other. The ability to model substructures makes UPPAAL especially suited for modeling complex structures in which multiple agents have the ability to influence each other (e.g., a clinical pathway). In addition, UPPAAL provides extensive model-checking capabilities, which allow model developers to check the reachability of states or pathways. UPPAAL provides an environment that aids interdisciplinary communication (e.g., by using multiple templates to model subprocesses while using a graphical user interface). The heterogeneous treatment path and importance of event timing in the detection of irAEs requires a flexible modeling approach. In the field of health economics, patient-level Markov models, discrete event simulations (DES), or agent-based models are generally applied approaches when flexibility is required. A unique benefit of TA-based models over the currently preferred modeling approaches is the compositional model structure that allows modeled agents to interact with and influence each other through communication channels. This compositional nature of TA-based models makes them more flexible to adjust and also allows modeling of a continuous process that involves multiple decisions, as in the case of treatment of advanced cancers. A downside of TA-based models is the limitation in statistical distributions, which are limited to a uniform and exponential distribution in UPPAAL. However, in UPPAAL, there are workarounds that allow for the incorporation of other statistical distributions into the model[304].

Our research aims to use a TA-based routine to model the clinical-diagnostic pathway of irAEs and to populate and calibrate the model using real-world survival data. This model will then be used to demonstrate the feasibility of UPPAAL by evaluating different test scenarios with increasing diagnostic performance of a broad spectrum of irAEs tests. We hypothesize that a TA-based model created in UPPAAL will be versatile enough to capture the complexity of the clinical path and decision making.

Methods

Study Cohort

The model was developed using a cohort of patients treated with nivolumab through the compassionate use program and regular care, containing 248 patients. A description of the study cohort, response assessment, and safety assessment was published in Lung Cancer in 2017[24]. Of these patients, 133 were recruited through the compassionate use program, whereas 115 patients started treatment in regular care. All patients received at least 1 line of previous treatment (chemotherapy) before nivolumab. In August 2015, nivolumab was available through the compassionate access program by Bristol-Myers-Squibb in 8 different hospitals in the Netherlands (NCT02475382). In this program, patients who had received at least 1 previous line of anticancer treatment were eligible to receive nivolumab if they had a good clinical performance with no or mild symptoms (World Health Organization performance status 0–1) and had adequate lab values for blood markers specific for organ (dys)function (e.g., aspartate aminotransferase, alanine aminotransferase, or creatinine)[247]. Data used in this study were limited to data acquired by the Netherlands Cancer Institute.

As part of routine care, patients were seen in the hospital every 2 weeks (wk), and laboratory tests were administered at baseline and every 2 wk thereafter. The laboratory assessment consisted of 30 blood tests including hematology, clinical chemistry, and hormonal measures, as depicted in Table 1.

Table 1 - An oversight of biomarkers included in the diagnostic panel.

Category	Measured biomarkers
Blood count	Hemoglobin, Hematocrit, Erythrocytes, MCV, Leukocytes, Neutrophil granulocytes, thrombocytes, cell differentiation
Liver function	Bilirubin, ALP, ASAT, ALAT, YGT, LDH
Clinical chemistry	CRP, Creatinine, GFR, Urea, sodium, potassium, phosphate, magnesium, glucose, total protein, albumin, calcium
Special chemistry	ACTH, Cortisol,

MCV, mean corpuscular volume; ALP, alkaline phosphatase: ASAT, aspartate aminotransferase; ALAT, alanine aminotransferase; YGT, gamma-glutamyl transferase; LDH, lactate dehydrogenase; CRP, C-reactive protein; GFR, glomerular filtration rate; ACTH, adrenocorticotropic hormone.

Disease progression was monitored through computed tomography imaging at 6 wk, 12 wk, 3 months (mo), 6 mo, 9 mo, 12 mo, and 15 mo after initiation of IMT. IMT was ceased in patients presenting with progressive disease. When any grade of irAE was clinically confirmed, patients were either withdrawn from IMT for the duration of a recovery period, which could take up to 5 wk, or IMT therapy was ceased definitively, and the patients proceeded to the next line of treatment. The NSCLC treatment landscape is heterogeneous. Therefore, in this model, the assumption is made that patients will transition to palliative chemotherapy after IMT is ceased definitively. During recovery, patients received appropriate treatment to recover from the incurred irAE. In practice, patients continued to the next line of therapy after discontinuation of IMT; in the model, we refer to this next line of treatment as "palliative care," since only the IMT phase was incorporated in the model. The data used from this cohort included the time on treatment in weeks, the frequency, and the incidence of toxicities and progressive disease. All relevant irAEs incorporated in the model are described in Table 2.

Table 2 - ***A*** **Description of immune related Adverse Events included in the model, based on data from the Nivo chohort (n=248).** The primary function of the test lies in the detection of irAEs; however, these tests results are also part of the diagnostic process.

Adverse Event	Probability of developing irAE during IMT therapy (%)	Number of reported events in 248 patients (n)	Time to development (Median (Range, days))	Time between development and Grade 3-4 AE (weeks)	Symptoms	First indication / Laboratory assessment	First symptom	Confirmation of diagnosis	Course of treatment	Complications when not treated
Pneumonitis*	9.6	10	60 (10-120)	2-4	Stuffiness	Patient	Patient: stuffiness	CT-scan or Bronchoscopy	~4-8 weeks improvement of symptoms	Pulmonary fibrosis
Colitis*		7	45 (15-180)	4	Diarrhea		Patient: Diarrhea	Coloscopy	~2 weeks until symptom relief	Bowel perforation
Dermatitis*		6	30 (10-120)	NA	Often: itch		Patient: itch		~2-4 weeks using the proper ointment	
Arthritis *		2	60 (40-90)	12	Joint pain		Patient: Thickened joints, pain	Physical assessment by a medical specilist or possibly rheumatoid factor	2-4 weeks with pain medication	
Pancreatitis	1.2	3	180 (100-200)	2-4	Stomach ache	Patient and lab: amylase, lipase, liver functon	Patient: pain in abdomen. Lab: Increased amylase/ lipase			
Hepatitis	2.8	7	90 (30-150)	4	Jaundice, feeling ill	Lab: ASAT, ALAT, YGT, ALP, Bilirubin	biomarker assessment		~ 4 weeks until improvement of lab values	
Hypofysitis	1.6	4	120 (90-300)	2	Feeling ill	Lab: Cortisol, ACTH, Na, K	Lab, or patient: feels ill		hospitalization: 2-4 until improvement	
Pancytopenia	0.4	1	60 (NA)	1-8	Bleedings, bruising easily	Lab: Hemoglobin, white blood cell count and differentiation, platelets	biomarker assessment	Shortage / absence Thrombocytes, leukocytes, erythrocytes		Infections, severe bleeding, anemia
Diabetes	0.4	1	10 (NA)	4-8	Thirst, urinating, blurry vision	Lab: Glucose	biomarker assessment, sometimes: thirsty, fatigue	Glucose: elevated	Insulin	Coma in case of severely elevated glucose

ASAT, aspartate aminotransferase; ALAT, alanine aminotransferase; YGT, gamma-glutamyl transferase; ALP, alkaline phosphatase; ACTH, adrenocorticoptropic hormone; Na, sodium; K, calcium

* Grouped as one immune related adverse event (irAE) based on the assumption that these irAEs will manifest with clear physical symptoms, and are generally discoved by patients themselves.

Expert Consultations

A multidisciplinary team involving experts in computer science, medical oncology, laboratory medicine, epidemiology, and decision science was involved in the development of the model. During the model development phase, 5 meetings were arranged with the multidisciplinary team, in which the research questions, model structure, model inputs, and results were discussed.

Model Construction

Although TA-based models have been established in other fields, in decision science, TA-based models have rarely been used or published. The care pathway described in this study consists of 2 distinct events that are monitored independently during the treatment process (i.e., the development and detection of irAEs and the development and detection of disease progression). Markov or DES models use “events” or “timing of events” to dictate the flow of patients through the model. Most of these models are built around 1 decision and process (i.e., a flow of subsequent actions). However, these models are less able to model asynchronous, parallel processes with multiple decision points and events causing an interruption of a process at arbitrary moments. A DES model does provide a more flexible approach, and depending on software-specific abilities, a DES model should be able to reflect the 2 independent subroutines. However, it would require a more complex model structure for which competing risks are defined for each combination of events. When using a Markov model, short cycle times could be used to allow for the evaluation of events at each cycle. However, a Markov model is less able to capture complex pathways with time-varying probabilities. UPPAAL provides the ability to model independent processes asynchronously, while synchronization channels can be used to interrupt processes in subroutines when necessary. Other merits of using UPPAAL are the ability to create substructures that represent a specific aspect of the simulated pathway and its model-checking engine. The ability to model substructures aids interdisciplinary communication, since each substructure is assigned its own template in a graphical user interface. In addition, the model-checking engine enables model developers to check the reachability of each state or pathway. A high-level overview of the clinical pathway is depicted in Figure 1. This high-level overview was translated into 6 templates used to capture different parts of the clinical pathway. In our model, the IMT is stopped if the patient has received 11 cycles of IMT, the patient develops progressive disease, or treatment is ceased because of irAEs. Therefore, our model adopts a time horizon of 66 wk, that is, 11 treatment cycles with a duration of 6 wk per cycle. A comprehensive overview of all transition probabilities, time constraints, and the underlying data source is provided in Table 3.

Table 3 - Model parameters and a description of the source on which the model parameter was based.

Description	Value	Unit	Source
irAE Occurrence: Grouped (Pneumonitis, Colitis, Dermatitis, Arthritis) - Hepatitis - Hypophysitis - Pancreatitis - Pancytopenia - Diabetes			
Time treatment start - occurrence of irAE, lower bound	1 - 4 - 3 - 14 - 8 - 1	Weeks	Patient data
Time treatment start - occurrence irAE, upper bound	26 - 21 - 43 - 29 - 9 - 2	Weeks	Patient data
Growth period (G1 - .G2 or G2 - .G3–4)	2 - 2 - 1 - 1 - 2 - 2	Weeks	Expert opinion
Probability of irAE occurrence, cycle 1	0.096 - 0.028 - 0.016 - 0.012 - 0.004 - 0.004	Probability	Expert opinion
Probability of irAE occurrence, cycle 2	0.192 - 0.056 - 0.032 - 0.024 - 0.008 - 0.008		
Probability of irAE occurrence, cycle 2	0.288 - 0.084 - 0.048 - 0.036 - 0.008 - 0.008		
Probability of irAE occurrence, cycle 2 – cycle 11	0.400 - 0.100 - 0.060 - 0.050 - 0.016 - 0.016		
Disease progression			
Probability of disease progression; cycle 1	0.38	Probability	Patiënt data
Probability of disease progression; cycle 2	0.29		
Probability of disease progression; cycle 3	0.22		
Probability of disease progression; cycle 4	0.168		
Probability of disease progression; cycle 5	0.128		
Probability of disease progression; cycle 6	0.097		
Probability of disease progression; cycle 7	0.074		
Probability of disease progression; cycle 8	0.056		
Probability of disease progression; cycle 9	0.04		
Probability of disease progression; cycle 10	0.033		
Probability of disease progression; cycle 11	0.025		

Table 3 - Description of all model parameters and a description of the source on which the model parameter was based. *(Continued)*

Description	Value	Unit	Source
irAE Occurrence: Grouped (Pneumonitis, Colitis, Dermatitis, Arthritis) - Hepatitis - Hypophysitis - Pancreatitis - Pancytopenia - Diabetes			
Recovery			
Probability of recovery, grade 0 irAE (false-positive)	1	Probability	Expert opinion
Probability of recovery, grade 1 irAE, occurrence 1 - 2 - 3 - .3	1 - 0.9 - 0.8 - 0		
Probability of recovery, grade 2 irAE, occurrence 1 - 2 - 3 - .3	0.8 - 0.5 - 0 - 0		
Probability of recovery, grade 3–4 irAE, occurrence 1 - 2 - 3 - .3	0.5 - 0.2 - 0 - 0		
Probability of recovery, fast recovery	0.4		
Enter recovery after detection of a G2 or G3 irAE	0.5		
Duration recovery	5	Weeks	Expert opinion
Duration fast recovery	2	Weeks	Expert opinion
Maximum of irAEs allowed	6		Patient data
Diagnostic accuracy			
Sensitivity diagnostic path - applied to all 6 irAEs included	85	%	Model calibration
Specificity diagnostic path - applied to all 6 irAEs included	91		

irAE, immune-related adverse event.

The 6 templates in the model are referred to as “Protocol,”“Patient,”“Toxic,”“Test,”“Monitor,” and “Progression Check.” Each of these templates fulfills a specific role in the model. The Protocol template is built to indicate whether a patient should receive tests aimed at the detection of irAEs, move to a recovery state, or transition to palliative care. The Patient template keeps track of the physical state of a patient (e.g., the grade of irAE incurred). The Toxic template determines whether a patient will incur a certain irAE. For patients who would develop an irAE, the template is also used to determine the point in time at which the irAE will manifest itself, simulate the progression of the irAE to a more severe grade, and determine whether a patient will recover once the patient enters the recovery phase. The Test template simulates outcomes of tests aimed at the detection of irAEs. The Monitor template is used to log the time patients spend in palliative care, which was chosen as the primary model outcome. Disease progression is modeled using the Progression Check template, in which the probability of disease progression is derived from patient data and decreases over time (Table 3). An extensive model description was published in an online repository[305].

Modeling of irAEs and Recovery

The severity of irAEs is described in grades ranging from grade 0 to grade 5. The absence of an irAE is defined as grade 0, whereas grade 5 is used to represent death caused by an irAE[306]. Within our model, grade 3, 4, and 5 irAEs are aggregated in a "grade3_and_4" irAE, because all of these grades of irAEs manifest with severe physical ailments that requiring clinical management. IrAEs progress from stage 1 to stage 3 within prespecified time intervals. The described time intervals are based on the average time between irAE development and presentation of severe physical symptoms (i.e., grade 3 irAEs). Therefore, patients transition from grade 1 to grade 2 and from grade 2 to grade 3 in 0.5 times the time it takes from irAE development until symptomatic disease (i.e., grade 3 irAEs; Table 2, "Time between Development and Grade 3–4 AE" column).

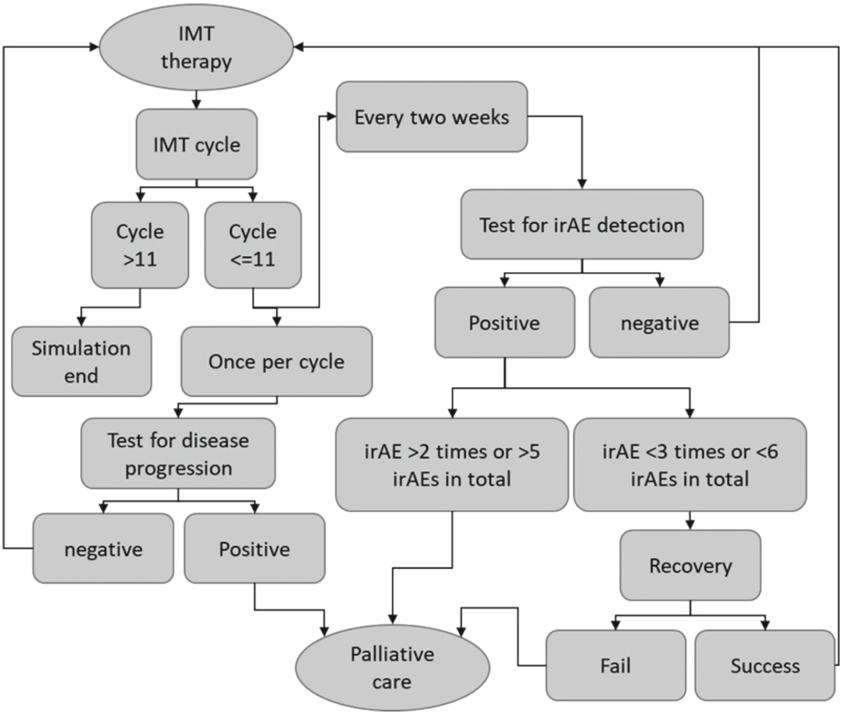

Figure 1 - High-level overview of the clinical pathway. One IMT cycle consists of 6 weeks of treatment with nivolumab. A test to detect progressive disease is performed once in every treatment cycle, and tests to detect irAEs are performed every two weeks. Patients diagnosed with progressive disease, incur a specific irAE a third time, or incur an irAE a sixth time transition to palliative care. Solid lines are used to depict standard transition options, the dashed line represents a conditional transition, i.e., the transition depends on the outcome of a separate process, in this case the detection of progressive disease. IMT-immunotherapy; irAE-immune related adverse event

As shown in Figure 1, patients in whom 1 of the tests results in a positive outcome can either enter a recovery phase or transition to palliative care. During the recovery period, patients are withheld from IMT for a duration of 5 wk. Patients who are diagnosed with an irAE within the first IMT cycle transition to a “fast recovery” state, in which the recovery period is reduced to 2 wk. The fast recovery state is introduced to resemble clinical decision making during the first IMT cycle. During the first cycle, physicians strive to optimize the chances of IMT to have a beneficial effect. During recovery, the test for disease progression continues according to the 6-wk schedule, and patients diagnosed with progressive disease during recovery transition to palliative care directly. Recovery from irAEs in the recovery phase is dependent on a prespecified probability. This recovery probability is based on the IMT cycle number, the grade of irAE, and the number of previous irAEs (Table 3). A transition to palliative care is made when recovery from the irAE fails. IMT treatment is ceased indefinitely after entering palliative care. Moreover, the model allows for recovery of the same type of irAE twice, and patients are allowed to recover 5 times from any combination of irAEs included in the model. In case a specific irAE occurs for the third time or a patient develops an irAE for the sixth time, a transition to palliative care is made directly without entering the recovery phase.

Model Calibration

Because the accuracy of the diagnostic pathway (i.e., the combined accuracy of the tests and interpretation of test results by a physician) aimed at the detection of irAEs is unknown because of a paucity of information regarding the accuracy of tests in this specific application (i.e., the detection of irAEs and the influence of a physician interpreting these test results), model calibration is performed to improve the accuracy of the diagnostic path and ensure the internal validity of the model by comparing the model output to real-world patient data. The model was calibrated by changing the input values for the sensitivity and specificity of the diagnostic pathway and comparing the probability of patients entering palliative care over time. The cumulative probability of patients entering palliative care over time is calculated based on real-world data using R statistical software version 3.6.1 and the ecdf function included in the stats package[307, 308]. Within UPPAAL, a query is run to simulate 11 treatment cycles (i.e., 66 wk) and 100,000 patients. The probability of entering palliative care over time is retrieved directly from the query output in UPPAAL. Model outputs are compared with real-world data through data visualization using the ggplot2 package (version 3.2.1) in R[309]. The model calibration is considered successful when the model outputs are within the confidence bounds surrounding the real-world patient data. Confidence bounds are generated according to the Dvoretzky–Kiefer–Wolfowitz inequality[310].

One of the shortcomings of using only the test outcomes is the incapability of expressing the overall physical state of the patient, which might be of great influence concerning the decision on treatment continuation. In practice, test results are interpreted by a physician; this interpretation step is likely to result in an increase in the accuracy of the diagnostic process. Herein we define the accuracy of the diagnostic process as the accuracy of the test after interpretation of the test results by a physician. The accuracy of the diagnostic process is expressed in terms of test sensitivity and specificity. Conversely, the test accuracy is still used to refer to the sensitivity and specificity of each test when outcomes are solely compared with the threshold values for disease detection. Since the accuracy of the diagnostic process is unknown, we adjusted this accuracy until the described model outcomes closely match the observed patient data.

Scenario Analysis

A scenario analysis is performed to assess the influence of changes in diagnostic accuracy, that is, the sensitivity and specificity of the diagnostic process on the probability of patients entering palliative care within 66 wk of IMT. The scenario analysis makes use of a query, which provides the probability of patients entering palliative care within 66 wk. For this scenario analysis, 14 scenarios with different input values for the test sensitivity and specificity are drafted, including 2 sensitivity values to represent a high and low test sensitivity. The scenario analysis includes 7 specificity values chosen after empirical tests during model calibration show that a specificity lower than 88% results in a probability of 1 that patients would enter palliative care before week 66.

Results

Model Construction

As described in the Methods section, the constructed model consists of 6 templates, with each template fulfilling a specific function in modeling the clinical pathway. Here, we describe 2 templates in more detail to provide insight into the inner workings of the model.

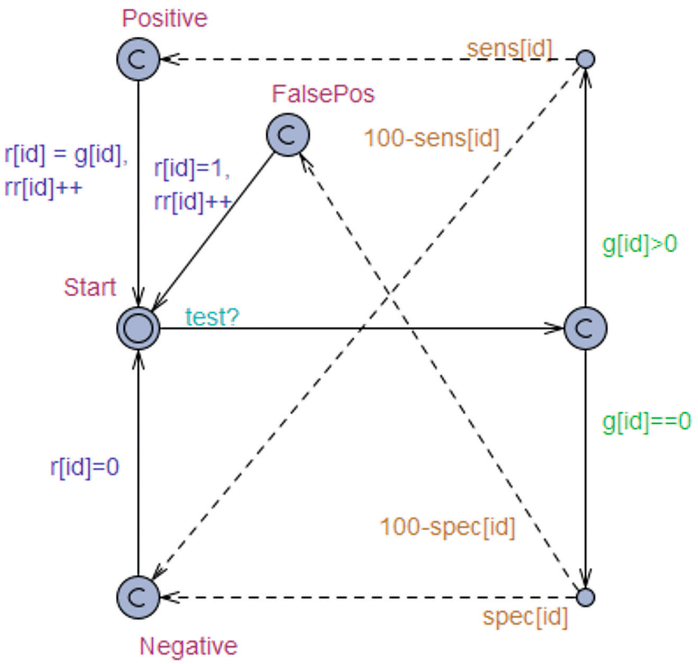

Figure 2 - Template “Test”, simulation of tests aimed to detect irAEs based on the test sensitivity, specificity, and presence of an irAEs. Solid line: transition path, dashed line: transition based on a probability of that line being executed. Green text: Guards, a requirement that must be met to allow for the transition to occur. Orange text: Probability, the probability with which a transition will occur. Blue text: update, once the transition occurs the defined parameters will receive an update. Light blue text: synchronization, the transition name followed by a “?” is a receiving channel and the transition will take place once the synchronization signal is received. If the name is followed by a “!” the channel will be used as a broadcasting channel and a synchronization signal will be send once the transition occurs.

Figure 2 depicts the template used to model the 6 different diagnostic processes that correspond to the irAEs included in the model. The template is replicated 6 times during a simulation, and each copy of the template is assigned an irAE through an identifier ([id]). Each test template is initiated in the “Start” location, and all tests are performed simultaneously when a signal is received through the synchronization channel indicated by “test?” When this synchronization signal is received, a transition is made from the start location to the location indicated by “irAE present?” The test outcomes depend on the presence of an irAE; in the model, “g[id]” is used to indicate the grade of irAE for each of the irAEs included in the model. In case a patient presents with the irAE corresponding to the respective test, the patient will follow the path “g[id]>0.” Patients free of the irAE continue through the path indicated by “g[id]==0.” For patients who present with an irAE, the test outcomes depend on the test sensitivity defined by “sens[id].” Patients presenting with an irAE transition to the location “Positive” in case of a positive test result. The probability of this true-positive test result is equal to the test sensitivity, whereas the probability of a false-negative test result (i.e., a transition to the location “Negative”) is equal to 1 minus the test sensitivity. Patients free of an irAE follow the path downward from “irAE present?” and can receive either a true-negative (transition to the location “Negative”) or false-positive test result (i.e., transition to the location “FalsePos”). The probability of receiving a true-negative test result equals the specificity of a test (spec[id]). The probability of receiving a false-positive test result is equal to 1 minus the test specificity. From the location “Positive,”“FalsePos,” or “Negative,” the patient returns to start. This transition automatically updates the values of “r,” which represents the grade of irAE incurred, and “rr,” which represents the number of times a specific irAE occurred.

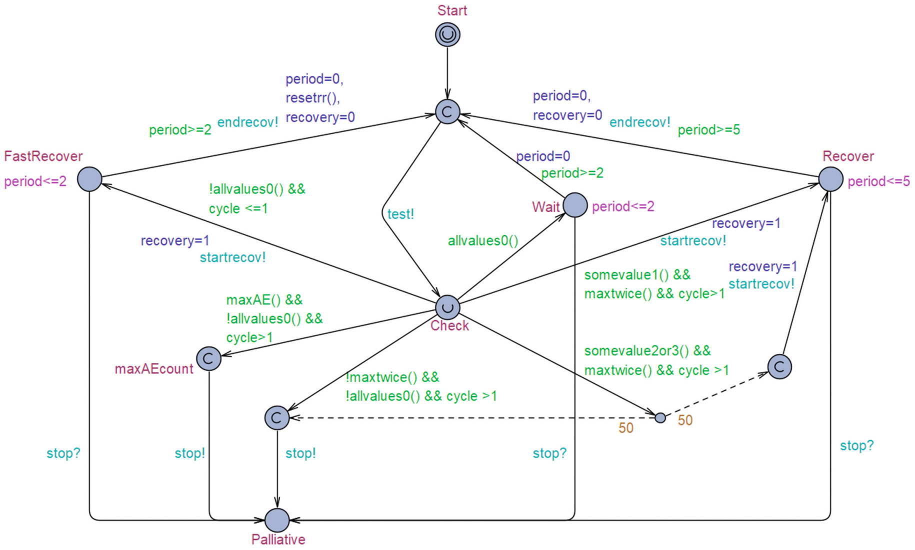

Figure 3 - Template “Protocol”, simulates the test protocol and clinical decision making. Solid line: transition path, dashed line: transition based on a probability of that line being executed. Green text: Guards, a requirement that must be met to allow for the transition to occur. Orange text: Probability, the probability with which a transition will occur. Blue text: update, once the transition occurs the defined parameters will receive an update. Light blue text: synchronization, the transition name followed by a “?” is a receiving channel and the transition will take place once the synchronization signal is received. If the name is followed by a “!” the channel will be used as a broadcasting channel and a synchronization signal will be send once the transition occurs. Pink text: invariant, an upper limit for the maximum time until the next transition has to occur from this location.

The test protocol template as depicted in Figure 3 is initiated in the “Start” location. This template is used to simulate the IMT treatment cycles and interpretation of test results. The actions taken in the test protocol depend on a time in weeks indicated by “period.” During IMT, patients receive tests every 2 wks aimed at the detection of irAEs. The tests are performed when the patient transitions from the location “Neutral” to the location “Check.” During this transition, the communication channel “test!” is activated. Test results are evaluated in the location “Check.” Depending on the test result, patients can either continue the standard test sequence when no irAEs are found, that is, the patient transitions to the location “Wait” through the path indicated by “allvalues0()”. Go into a fast recovery phase “FastRecover” if the patient is diagnosed with an irAE during the first treatment cycle (path: !allvalues0 && cycle<=1). Enter the normal recovery phase “Recover” if the patient is diagnosed with a grade 1 irAE, is diagnosed with an irAE fewer than 3 times, and has completed at least 1 IMT cycle (path: somevalue1() && maxtwice() && cycle >1). Patients diagnosed with a grade 2 or 3 irAE, who have been diagnosed with an irAE fewer than 3 times, and who have completed at least 1 IMT cycle have an equal probability of 0.5 of either entering the recovery phase (location: Recover) or transitioning to palliative care (location: Palliative). This probability of 0.5 is indicated by the number 50 near the dashed arrows. Patients who receive a third positive test result for one of the included irAEs or who are diagnosed with an irAE for the sixth time transition to palliative care (location: Palliative) directly. The fast recovery period is defined to last 2 wk, as defined by the guard “period>=2” and the invariant “period<=2”, meaning the transition has to occur when the value of period equals 2. The standard recovery period is defined to last 5 wk (guard: period>=5, invariant: period<=5). The probability of recovery depends on the grade of irAE and the number of times the patient is diagnosed with the irAE. This probability is looked up in a table using the notation c[g[id]][rec[id]], in which “c” indicates the probability of recovery based on “g,” which represents the grade of irAE, and “rec” represents the number of real detected irAEs (i.e., the number of previously incurred true-positive test results).

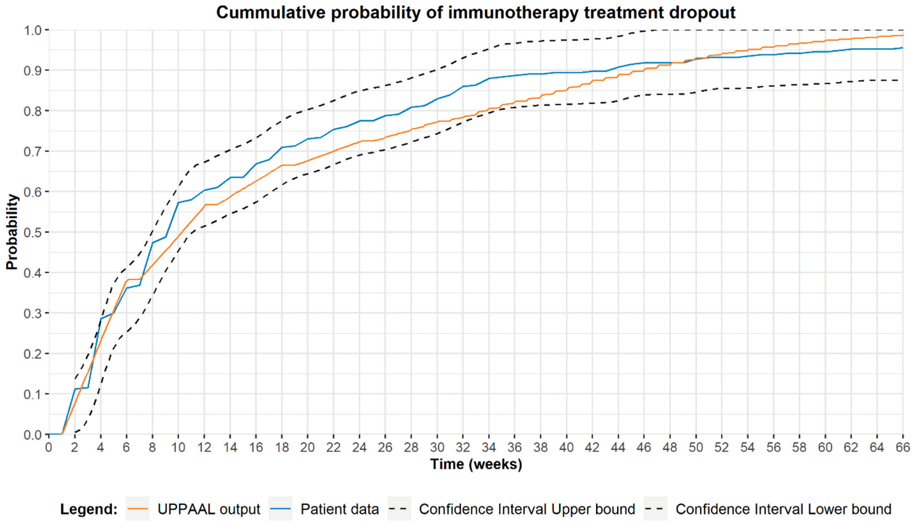

Figure 4 - Model calibration, the probability of patients transitioning to palliative care over time. The blue line represents patient data, the orange line depicts model output, and the black lines represent the confidence bounds surrounding the patient data based on the Dvoretzky–Kiefer–Wolfowitz inequality. The model calibration was performed using the accuracy of the diagnostic process, satisfactory results (i.e. model outputs are located within confidence bounds surrounding the patient data over the full 66 week period) were provided using a sensitivity and specificity of 85% and 91%, respectively. Model outcomes were derived using the query: E[<=66;100 000](max:paltime)

Model Calibration

The model was calibrated by comparing the cumulative probability distribution of patients entering palliative care over time to observed patient data. Ultimately, a sensitivity and specificity of 85% and 91% provided a satisfactory fit, respectively. The choice was based on the visual fit of the model outcome as compared with patient data and its corresponding confidence bounds. However, after calibration, the model still underestimates the probability of patients entering palliative care slightly between week 8 and 48 of IMT and overestimates this probability from week 50 until week 66. Figure 4 depicts the cumulative probability distribution of patients entering palliative care over time, derived from patient data and model outcomes.

Scenario Analysis

A scenario analysis was performed to evaluate the effect of the accuracy of the diagnostic process on the probability of patients entering palliative care. Table 4 depicts the probability of patients entering palliative care within 11 IMT cycles given a combination of sensitivity and specificity values for the diagnostic process (i.e., test results, including interpretation of test results by a physician). Our results show that changes in test specificity can have a significant effect on the probability of patients entering palliative care within 11 IMT cycles, with a difference of 15% between a test specificity of 88% and 99% (Table 4). Moreover, there was no significant difference between the scenario with the high sensitivity and low sensitivity in patients entering palliative care before week 66 of IMT, ceteris paribus.

Table 4 - Outcomes of the scenario analysis. The probability of patients transitioning to palliative care within 11 IMT cycles given a pre-specified combination of sensitivity and specificity of the diagnostic path. The specified diagnostic accuracy of the diagnostic path was applied to all six tests corresponding to the six immune related adverse events included in the model. Model outcomes were derived using the query Pr[\=66](\.Protocol.Palliative). The top row represents the test specificity, the left most column represents the two scenarios including a high and low sensitivity.

		Specificity						
Problability of patients transitioning to palliative care with 11 treatment cycles.		**.88**	**.90**	**.92**	**.94**	**.96**	**.98**	**.99**
Sensitivity	**.60**	0.99	0.99	0.98	0.95	0.89	0.84	0.83
	.90	0.99	0.99	0.98	0.95	0.89	0.84	0.83

Discussion

Managing IMT-induced irAEs is one of the great challenges in cancer management today. Several attempts have been made to predict patients' susceptibility to irAEs in the treatment of solid tumors using immune checkpoint inhibitors. A variety of biomarkers have been studied in this context, including pretreatment serum antibody levels in melanoma patients[311] and baseline thyroid peroxidase, thyroglobulin, follistatin, and human interferon-inducible protein-10 levels in NSCLC patients [312, 313]. Moreover, a review by von Itzstein et al [314] demonstrated the large variety of biomarkers studied in relation to the diagnosis and prediction of irAEs. Although predictive biomarkers might aid the identification of patients with a high susceptibility for irAEs, monitoring is still needed to identify any occurring irAEs. To our knowledge, there are no previous studies that have aimed to construct a model to analyze the development and detection of IMT-induced adverse events in lung cancer patients. In this article, we present a model that can eventually be used to evaluate the influence of test accuracy, timing, and composition of the diagnostic test panel on treatment continuation. Optimization of the diagnostic test panel might ultimately lead to a cost reduction through less frequent testing or a reduction in biomarkers included in the test panel. In addition, as depicted in Table 1, the current diagnostic panel consists of 30 biomarkers assigned to 1 of 4 analysis categories (i.e., blood count, liver function, clinical chemistry, and special chemistry). Although the current model aggregates all relevant biomarkers and their interpretation into a single test per irAE, an extension of the proposed model could be used to evaluate the added value of each individual biomarker in the diagnostic panel. Currently, a physician needs to analyze results from all biomarkers included in the diagnostic panel, resulting in a complex decision scheme. However, optimization of the diagnostic panel results in removal of tests with little added value from the test panel, resulting in a potentially less complex decision scheme. Although optimization of the test sequence would be unfeasible in prospective or retrospective studies because of the number of possible diagnostic strategies, a modeling approach allows for the evaluation of a large number of test sequences with less financial and time constraints. Our results indicate the feasibility of developing and calibrating a TA-based model developed in UPPAAL to simulate IMT in lung cancer patients, including disease progression and development of irAEs.

During model development, the probability of completing all 11 IMT treatment cycles was chosen as the outcome parameter to calibrate the model and compare different scenarios. This parameter was chosen in combination with the time horizon, which was limited to the IMT treatment period. It is known that the care pathway in NSCLC is very heterogeneous, and various treatment options are available after IMT. This heterogeneity in therapeutic pathways makes it unfeasible to include all relevant pathways in the model and to incorporate the effect of IMT on quality of life and survival during subsequent treatment lines. In the model, it is assumed that a 5-wk treatment cessation during a recovery period does not influence treatment outcomes. However, ceasing treatment too early might result in withholding a potentially beneficial treatment from patients.

With regard to model calibration, the model still slightly underestimates the probability of patients transitioning to palliative care during the period of week 8 to week 48 of treatment. Conversely, the model slightly overestimates this probability for the remaining 18 wk. These differences might be explained by 3 modeling challenges.

First, the diagnostic process of irAE detection is difficult, and there is a lack of strict guidelines on restarting IMT after a patient recovers from an irAE. Second, irAEs occur with a relatively low incidence, and patients are more likely to stop IMT because of disease progression. This might result in an underestimation of the probability of developing irAEs or the probability of recurrence after recovery. In the data set used during model construction, the percentage of incurred irAEs ranged from 0.4% up to 6.5% for individual irAEs, whereas in total, 18.1% of patients incurred an irAE. This low prevalence directly affects the uncertainty regarding the timing of events and the probability of occurrence. However, this does not affect the viability of UPPAAL in modeling the clinical pathway or the ability to calibrate the model. Third, little is known about the actual sensitivity and specificity of the diagnostic process in this specific setting, mostly because thresholds for disease detection are derived from other patient groups or a healthy population. In addition, because test results contribute to the decision-making process, the sensitivity and specificity of the diagnostic process can never be 100%, as many other factors influence the decision on treatment continuation. Moreover, within this process, it is unclear how much the true diagnostic accuracy of the tests is influenced by the physician, since the data used during model construction provide information only on the actual clinical decision. Although the true accuracy of the test is unknown, the model could be used to assess the influence of removing a test or changing the test frequency on the probability of completing all IMT cycles on cohort level.

We performed the scenario analysis to evaluate the influence of the accuracy of the diagnostic process on the probability of patients transitioning to palliative care. This scenario analysis shows a strong influence of test specificity, although the influence of test sensitivity did not appear to affect model outcomes. These results match our expectations given the low incidence of irAEs. The absence of an effect of test sensitivity might partially be explained by the low incidence of irAEs. However, the relatively high-test frequency is also likely to lower the effect of changes in test sensitivity on model outcomes, as this could limit the impact of false-negative test results on model outcomes. Moreover, the direct effect of test sensitivity is not fully captured by the outcome measure, because a low-test sensitivity might only delay the time until the irAE is detected in the model. This delayed detection could occur through a positive test later on or due to a symptomatic presentation of the irAE. Hence, the influence of the sensitivity of the diagnostic path on the probability of patients entering palliative care before week 66 is limited. However, there is a difference in the recovery probability depending on the grade of irAE at the time of diagnosis. Therefore, a high-test sensitivity might allow for the earlier identification of irAEs (i.e., lower grade), resulting in more patients successfully recovering from the irAE and completing the 66 wk of IMT. Unfortunately, the low incidence of irAEs reduces the effect size of the sensitivity on cohort level.

In the currently used protocol, patients provide a blood sample every 2 wk. However, the blood samples are obtained at random time points throughout the day, depending on the patient's appointment. The accuracy of the test might be influenced by this inconsistency in timing, since it is known that some blood values fluctuate during the day because of biological variability or under the influence of external factors (e.g., food or beverages). With the introduction of a new test to the test panel, it is key to keep an eye on the influence of the test on clinical decision making. A diagnostic test will provide added value only if it provides actionable results and when a physician acts on these results, in combination with tests already used. In patients with stable disease or response, it might be detrimental to stop IMT because of an irAE in cases of a lower grade or relatively harmless irAE.

Modeling the care pathway and evaluating treatment protocols might be helpful for identifying the most optimal test strategy, based on the composition of the diagnostic kit. Moreover, it would be unfeasible and potentially unethical to evaluate all options in a trial-based setting. Herein we present how UPPAAL can be used to develop a model that emulates the clinical pathway. As with most model-based evaluations, the generalizability of the results strongly depends on the underlying data, the alignment of the model structure with real-world clinical guidelines and pathways, and assumptions about the prior knowledge of physicians using the diagnostic information. Although we do not expect significant differences in the prevalence of irAEs or progressive disease, the clinical pathways and the prior knowledge of physicians do differ between health services. Because the diagnostic accuracy largely depends on the interpretation of results by a physician, and clear guidance regarding the interpretation of the test is lacking, it is expected that there will be differences in management between physicians, not only on an international level but also on an institutional level. In this study, the diagnostic accuracy was estimated in the model calibration and reflects the average accuracy of the diagnostic path for the group of physicians involved in the treatment of the study cohort. It is likely that physicians working in another health service have different experience and prior knowledge, and the generalizability of the current study critically depends on the extent of clinical expertise and variation in prior knowledge.

Future work will involve expanding the model to identify the optimal diagnostic strategy in terms of costs and outcomes.

In conclusion, we have shown that it is worthwhile to construct a TA-based model to emulate complex clinical decisions in the management of NSCLC using UPPAAL. Based on assumptions that can be changed and adapted in the model, we calibrated the model using real-world data. The scenario analysis indicated that the effect of test accuracy on the probability of lung cancer patients treated with immunotherapy transitioning to palliative care is predominantly dependent on the test specificity. Moreover, the influence of test sensitivity is limited, and a high test specificity is important to prevent the too-early termination of IMTs.

General discussion and future perspectives

For this thesis, the framed research question was: "*is it possible to predict a response with one or more factors which directly contribute to immune response, or is an overaching surrogate marker, containing multiple factors, more predictive?".* The upcoming discussion answers this question.

Since immunotherapy is available in clinical practice, the outcome of patients with lung cancer improved in the terms of overall survival and progression free survival, however, it is an expensive treatment and the response rate remains relatively low [3, 5, 42, 133]. In order to maintain a sustainable health care system, further steps have to be taken. Besides Heathcare Technological Assessments, a better selection of patients is of paramount importance. Therefore, biomarker development is important, which allows a better therapy selection for each individual. Allthough the results in this thesis are promosing, the problem is still not solved. Especially in theirfinal stage, patients prefer to retain the best quality of life. With accurate patient selection and monitoring during treatment, it helps to prevent prescribing ineffective treatment.

Until now there are three U.S. Food and Drug Administration (FDA) approved biomarkers for the prediction on response to immunotherapy: PD-L1 (Programmed Death-Ligand 1), tumor mutational burden (TMB) and microsatellite instability high (MSI-H)/deficient mismatch repair (dMMR). The first two FDA approved markers are set out in table 1, where the seven requirements of a perfect marker are compared between the markers. Also, the other potential biomarkers presented and discussed in the previous chapters are expanded. In this thesis, most markers are compared to the current FDA approved marker PD-L1. However, a mutual comparison between all markers would be interesting as well. Table 1 is based both on literature and experience and is intended to guide the comparison, and more importantly, further discussion.

The third FDA approved marker, MSI-H/dMMR, is not included in this comparison. This marker, or a combination of markers, shows a characteritisc of a tumor in which there is a deficiency in the mismatch repair (or dMMR). This leads to cells not being able to recognize and repair spontaneous mutations, leading to a high mutation burden and a high microsatellite instability (or MSI-H) [315]. Since this marker is mostly used in colorectal carcinomas and not (widely) used for prescribing immunotherapy for lung cancer [315, 316], this marker is not included in table 1.

FDA approved markers

PD-L1

The aim of immunotherapy is to active tumor-specific cytotoxic T lymphocytes(CTL), in order to enhance and maintain a good antitumor response [43]. Multiple cells and receptors are involved in this complicated process. Nivolumab and pembrolizumab, both PD-1 checkpoint inhibitors, interact with the PD-L1 receptor, causing such a CTL response [15, 317]. PD-L1 is a biomarker designed for the prediction on response to anti-PD-1 therapy with expressing its target. High expression of PD-L1 in a biopsy of the tumor (primary or metastases), analyzed using immunohistochemistry (IHC), suggests a good response to immunotherapy [20]. Across different trials evolving nivolumab, objective response (OR) and longer duration of response (DOR) have been registered both in PD-L1-positive and PD-L1-negative NSCLCs, even in numerically higher among positive tumors [2, 36], and no differences have been described for different levels of PD-L1 expression [2, 3, 5, 37]. The predictive role of PD-L1 expression for pembrolizumab showed a limited objective response rate (ORR) of 30%. In treatment naive patients with a tumor PD-L1 expression of >50%, the response rate was 44.8% when treated with pembrolizumab [42]. These positive results resulted in two FDA approvals for registration of pembrolizumab. Besides the limited accuracy of PD-L1, a discordance between two biopsies from the same patient has been observed. This uneven distribution of the PD-L1 positive tumor cells is a known phenomenon in clinical practice [52-57].

TMB

Tumor Mutational Burden (TMB) is defined as a total number of somatic mutations per coding area or per mega Base of a tumor genome [318] and is a potential marker for the prediction of response. In general, smokers show a higher mutational burden than never smokers [115]. The rationale is based on the observation that mostly tumors developed from chronic mutagens (cigarette smoke in lung cancer) show better responses than other tumors [112, 115, 319]. Rizvi et al investigated in two small independent cohort the correlation between tumor mutational burden and response to immunotherapy. They showed that a higher TMB was associated with clinical efficacy of pembrolizumab, with a median of 302 mutations in patients with response, compared to 148 in patients with no response (Mann-Whitney P=0.02)[112]. In the KEYNOTE-158 study, the TMB of patients was evaluated and a high TMB (>10 mutations) was compared to a non-high TMB (<10 mutations). 102 of the 790 evaluable patients showed a high TMB. Here, the overall response rate was 29%. The other 688 patients showed a response of only 6% [318]. Based on this trial, the FDA approved TMB for the treatment of adult (and pediatric) patients with unresectable or metastatic high TMB solid tumors, that progressed following prior treatment and who have no satisfactory alternative treatment options [316]. However, the disadvantages are: it is invasive (a biopsy is needed) and the test is quite expensive compared to PD-L1 [320]. Besides, different trials show a potential improved OS or PFS, but are not yet able to predict response [321].

Table 1 – Simplified overview of the different biomarkers for the prediction or monitoring of immunotherapy response in patients with NSCLC.

	FDA approved	FDA approved	Ch 3	Ch 4	Ch 5+6	Ch 7	Ch 8	Other*
Level of Evidence [322]	1A	2A	2B	2B	2B	3B	2B	2B
Requirement	PD-L1	Tumor mutational burden	Platelet-RNA	Protein classifier	eNose	Dog	Tumor markers	ctDNA
1. Understandable rationale	🟩	🟩	🟧	🟧	🟧	🟧	🟩	🟩
2. Accurately predict or monitor a responder	🟧	🟧	🟥	🟩	🟩	🟩	🟩	🟩
3. Minimally invasive	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩
4. Easy to collect and perform	🟧	🟧	🟧	🟩	🟩	🟥	🟩	🟩
5. Reproducible, robust and repeatable	🟧	🟧	🟥	🟧	🟧	🟥	🟩	🟥
6. Fast	🟧	🟧	🟧	🟩	🟩	🟩	🟩	🟧
7. Costs	🟧	🟥	🟥	🟧	🟩		🟩	🟥

*ctDNA is considered the most important other player in the field of non-invasive biomarker research and comparable to the markers in this thesis. The other FDA approved marker, MSI, is not included, since its approval does not regard lung cancer.

** In table 1 in the introduction "cost effectiveness" is considerd an important requirement for a perfect biomarker, however, it is complicated to measure and therefore needs a study on it's own. Therefore, the first derivative, the costs of a marker, is used in this table.

Abbreviations: Ch: Chapter

The traffic light colours indicate requirement available (green), not fuly available (orange) or not available (red).

Biomarkers in this thesis

In **Chapter 3** we investigated whether platelet RNA signatures may provide classification power for nivolumab immunotherapy response prediction before start of treatment. Because platelets are involved in the immune response, we expected that this would allow us to predict whether someone would respond or not. The benefits of this biomarker would be that it is minimally invasive and easy to collect. Therefore, we analyzed and build an algorithm with the blood of 286 patients, all drawn before start of immunotherapy treatment. In a validation cohort of 107 patients, we found an area under the curve (AUC) of 0.58 (95%-confidence interval (CI) 0.45-0.7), which is not sufficient for clinical practice. As discussed before, more samples or another algorithm might improve the use of this potential biomarker. Also, nowadays the result may take days until one week, however, in the nearby future it might be possible that that time is shortenend.

All in all, it is the least promosing marker for the near future for the prediction of response. It would be interesting to know if we would be able to build a classifier for monitoring of response.

The second biomarker investigated is the use of proteins for the development of a classifier. Therefore, in **Chapter 4**, mass spectrometry (MS)-based proteomic analysis was performed on pretreatment sera derived from 289 patients with advanced NSCLC treated with nivolumab. Machine learning combined spectral and clinical data to stratify patients into three groups with good ("sensitive"), intermediate, and poor ("resistant") outcomes. Duration of response and survival were examined, which appeared to differ significantly between the three groups: significantly better OS was demonstrated for "sensitive" relative to "not sensitive" patients treated with nivolumab; HR, 0.58 (95% confidence interval, 0.38–0.87; $P = 0.009$). Our study showed that proteins used for this classifier, were associated with inflammatory processes and wound healing cascades. This is in line with all data reported so far. In this project, multiple validation cohorts where used, containing samples from other hospitals and/or intended treatment group (chemotherapy). This implies an accurately prediction test, which is reproducible, robust, and repeatable. Of note, the test-time of 15 minutes is quite fast and applicable in the clinic.

Chapter 5 and 6 show that exhaled breath analysis by the eNose discriminate between responders and non-responders to anti-PD-1 therapy in NSCLC. Ineffective anti-PD-1 therapy could potentially be prevented in 24% of the patients without erroneously withholding anyone effective treatment. These patients, classified as non-responder might be saved from unnecessary delays and start treatment with a better alternative. In the validation set, results were compared to PD-L1, which showed that the eNose outperformed PD-L1. The rationale is clear: the use of volatile organic compounds (VOCs), originating from metabolic processes represent changes in the immunological reaction. With the use of exhaled breath, it is a non-invasive biomarker. The eNose is designed for its easy use and fast result. Therefore it can be used in clinical practice. Unfortunately, the robustness depends on the sensors of the eNose, which are relativity easy out of balance, for example with temperature changes or the use of alcohol nearby. In an outpatient clinic this would not be a problem. Daily calibration is performed for the check-up of the quality of the eNose. Only the clear rationale is lacking for the eNose as a biomarker, therefore this is a very promising marker for the near future. However, these breath-based technologies have not been introduced in clinical practice yet and there are many hurdles to take prior to implementation.

In **Chapter 7,** the use of a dog to identify medical situtions has been introduced by others. In the article, referred to in this editorial, a trained dog was used for the (early) diagnosis of lung cancer. With their outstanding good scent, dogs are able to smell the small differences between patients with lung cancer and healthy individuals. The results of using a dog in clinical practice are promising, with a sensitivity of 95% and a specificity of 98% for the detection of cancer. It is a minimal-invasive method for the patient since exhaled breath is used. In this study, breath samples are used, which might be a hard to perform, however, using a 'real dog' would make it easy to perform and fast. However, for every dog the full analytic and clinical validation should be performed. Also, how to make sure their results remain stable. What if a dogs get sick and recovers (for example: COVID), should the full validation 'program' start all over again? Therefore, also the costs should be questioned, since living beings need more than just a place to be.

All in all it exemplifies the potential of breath-based diagnostics with, but will never become available due to the completely impracticle nature of this “biomarker”.

In **Chapter 8**, readily available tumor markers were used for the monitoring of response at 6 weeks. These markers were introduced decades ago in different laboratories across the world, but so far it has not been possible to use these markers in practice for predicting response. In our study, a relatively easy test was used on two conditions per value (a 50% increase and above a minimal value). In the validation set, a specificity of 91.9% and a sensitivity of 40.2% could be reached with the use of 3 tumor markers. With a relatively easy test based on two conditions per value (a 50% increase and above a minimal value). These results indicate that serum tumor markers can be used to identify patients in which treatment can be discontinued early because of its ineffectiveness. Since the markers are associated with tumor mass and therefore markers are expected to increase when progression occurs, the rationale is clear. It is minimally invasive, easy to collect, already validated for its robustness, and fast. This biomarker cannot be used to predict responses accurately but is a generic monitoring biomarker and as such can be used to stop treatment early on before radiologic or clinical progression and as such can prevent toxicity and costs. The optimal positioning is currently being studies in various prospective trials.

ctDNA as marker

Table 1 also shows another important and comparable player in the field of biomarker development, namely, the use of ctDNA. ctDNA is a byproduct of dying cancer cells and with its short half-life time it can be used for response monitoring of the tumor higher specificity then serum tumor markers [250, 323] . Goldberg et al showed in 28 patients the dynamics of ctDNA in immunotherapy, where they saw a drop of ctDNA in patients who showed a response to immunotherapy [250]. They found a strong agreement between imaging response monitoring and ctDNA response monitoring. The rationale is not specific for immunotherapy; it represents tumor load and is therefore a generic biomarker for response monitoring [324]. Since it can be measured in blood, the test can easily be performed, but it often takes one week or more. ctDNA as a marker is already widely validated. Currently, the test is expensive, and a Heath Technology Assessment should be performed to determine its place in the decision tree. Another disadvantage is a baseline mutation marker must be present for follow up. Its current development and its minimal-invaseness by using blood makes this marker interesting when compared with the other biomarkers. In the nearby future it might be possible to shorten the time to result and the costs. More research is needed to assess the robustness.

Comparison of biomarkers

In search of the perfect biomarker, the candidate markers must outperform the FDA approved markers as shown in table 1. Both approved markers have a clear rationale for the use in IO treatment. However, they are not very accurate, needs invasive testing and some come with high costs (TMB). The PDL1 en TMB biomarkers were developed as companion diagnostics and therefore been implemented early on as standard to allow the use of IO drug therapy [38, 41, 316].

Costs

The costs of the markers vary widely and goes from relatively cheap tumor markers to expensive DNA analysis. Although a cost-effective biomarker is preferred in a sustainable health care system, research is still done with relatively expensive methods. It goes without saying that it is easier for a cheap marker to be costeffective and with that the question of how useful it is to experiment with expensive methods. However, if a marker prevents patients from having severe side-effects, these benefits may still outweigh these costs.

Review

We have tried to address the question if it is possible to predict an IO response to checkpoint inhibitors while using surrogate biomarkers. Therefore two main hypothesis were formulated:
(1) It is possible to predict response with a surrogate biomarker.
(2) There are more perfect biomarkers for the prediction of response to immunotherapy in patients with advanced NSCLC.

This thesis showed that it is possible to predict response with the use of different surrogate markers (table 1), but these are still in the beginning of their development and more research needs to be done.

Strengths

The strength from these beforementioned studies, is based on the high number tested and at the start of treatment. This results in a uniform group and limits certain biases, such as selection bias and temoral bias, as much as possible. Also, all studies contain a training and validation group, confirming the results on a small scale. The results are accessable for other researchers who want to use our research for further projects.

Limitations

At first, all these markers are only tested in one trial, so more research needs to be done. As described in chapter 1, there are multiple steps to be taken before a marker can be used in clinical practice, for example a FDA approval. Since the current treatment landscape of NSCLC is changing, this is a challenging.

These studies are at risk for different forms of bias: sample bias, storage or patient selection bias since only patients living in the Netherlands who were able to visit the NKI (i.e. having the money and the perforamce) could participate. In two studies other centers were involved, which might reduce the chance on bias.

Some patients were included before histologic al samples were analyzed for PD-L1 expression or tested with one of the different available assays. Initially, PD-L1 testing was not required and was often missing, leading to a non-uniformity of the patient group.

Recommendations and future perspective

Biobank

The development period of PD-L1, more or less started in 2012 together with the clinical trials, was relatively short, which shows that combining biomarker and drug development is very effective [15]. However, the development of biomarkers becomes more challenging once the drugs are already available. To safe time, and therefore years of giving to much therapy, it would be recommended that the biobank, collected by pharmaceutical companies during phase 1-3 trial, should become available for other further biomarker development directly following registration of the drugs.

Publication of "negative" results

Most projects are finished within four years. The exception here is the platelet RNA. In today's research world it is all about funding and publications: the more publications, the more successful a research group. Therefore, most projects are finished when a publication has followed. However, there is a strong bias towards publication of positive results [325]. Although more attractive for its readers, it also creates an illusionary world in which either research with a negative outcome is not published, or it is written in a way that there is an outcome looks better than it is. We choose to report the results of a negative biomarker study (Chapter 3). We are convinced of the results and feel that negative studies are of great importance to be presented to the scientific community. This also prevents other researchers to embark on fruitless endeavors.

Compound biomarkers

With more and more biomarkers becoming available, the physician has to make a decision based on multiple results. This is complicated when the biomarkers all point into a different direction. In **Chapter 8**, in which we used multiple readily available tumor markers, we searched for the best combination of these markers. In 10 years' time more studies will show which combinations of markers should be used, and if only one marker needs to show a positive result, or all the markers. Because a large cohort is preferred for the search of a (combination of) biomarker(s), also modeling can be used. In **Chapter 9**, using data from 248 patients in combination with the literature, a model was built that was representative of reality regarding duration of treatment and survival. Then, this model as used to see how often we should run tests to detect side effects. Here we show that with least testing we were also able to detect a side effect on time. Using a model, the first steps could be taken for combining markers.

New biomarker development

In 2018, a few years after immunotherapy was introduced in second line treatment, the base of this thesis, the therapeutic landscape changed. Nowadays, pembrolizumab is used in first line treatment in patients with >50% PD-L1 expression. A combination of chemotherapy and immunotherapy is given in the other situations [50]. Other drug development, the use of already excisiting drugs and/or new combinations of drugs continues [328]. For example, the addition of bevacizumab (an anti-angiogenesis agent) to the current chemotherapy and immunotherapy regime, as seen in the IMpower150 study, showed an overall response rate of 61%, and is therefore quite promising [329].

There is growing evidence that there is a correlation between the microbes in the gut and lung, the so-called microbiome, and cancer development [330]. Studies have shown the influence of the microbiome on response to immunotherapy [331-333], and could therefore be potential biomarkers [45].

This continuous change may complicate the development of new biomarkers in the context patient treatment, but helps us to better understand all processes involved. To address this problem, an overarching standardized operation protocol (SOP), in which paired tumor biopsies blood and microbiome samples are collected, biomarker research can continue.

Collaboration

Last but not least, a well-known recommendation that should not be forgotten is collaboration. In this thesis we started a collaboration with another department (chapter 8) or another hospital (chapter 3, 5, 6), with an available laboratory to develop and validate a biomarker. I do believe that this is an important step for biomarker development and the reason for its success. Therefore, it is recommended to focus on two sorts of collaborations: (1) between professions and (2) between universities/hospitals to solve these clinically relevant questions.