Empirical methods for the validation of Time-To-Event mathematical models taking into account uncertainty and variability: Application to EGFR+ Lung Adenocarcinoma.

Authors

Evgueni Jacob [1]

Angélique Perrillat-Mercerot [1]

Jean-Louis Palgen [1]

Adèle L'Hostis [1]

Nicoletta Ceres [1]

Jean-Pierre Boissel [1]

Jim Bosley [1]

Claudio Monteiro [1]

Riad Kahoul [1]

Affiliations

[1] Novadiscovery, Lyon, France

Corresponding author

Evgueni Jacob
1 Place Giovanni Da Verrazzano, 69009 Lyon
+33 6 79 40 93 62
evgueni.jacob@gmail.com

Key words

Empirical, bootstrap, log-rank, confidence interval, prediction interval, coverage, joncture, mechanistic model, knowledge based model, validation, EGFR+ Lung Adenocarcinoma.

Abstract

Over the past several decades, metrics have been defined to assess the quality of various types of models and to compare their performance depending on their capacity to explain the variance found in real-life data. However, available validation methods are mostly designed for statistical regressions rather than for mechanistic models. To our knowledge, in the latter case, there are no consensus standards, for instance for the validation of predictions against real-world data given the variability and uncertainty of the data. In this work, we focus on the prediction of time-to-event curves using as an application example a mechanistic model of non-small cell lung cancer. We designed four empirical methods to assess both model performance and reliability of predictions: two methods based on bootstrapped versions of parametric statistical tests: log-rank and combined weighted log-ranks (MaxCombo); and two methods based on bootstrapped prediction intervals, referred to here as raw coverage and the juncture metric. We also introduced the notion of observation time uncertainty to take into consideration the real life delay between the moment when an event happens, and the moment when it is observed and reported. We highlight the advantages and disadvantages of these methods according to their application context. With this work, we stress the importance of making judicious choices for a metric with the objective of validating a given model and its predictions within a specific context of use. We also show how the reliability of the results depends both on the metric and on the statistical comparisons, and that the conditions of application and the type of available information need to be taken into account to choose the best validation strategy.

Introduction

Mechanistic models and by extension knowledge-based models provide a mathematical representation of biological phenomena, and by extension physiological and pathophysiological mechanisms. Based upon knowledge in the literature describing components of biology which are integrated using fundamental laws of nature such as physical and biochemical principles, these models allow representation and analysis of complex dynamic behavior of variables seen in biology and clinical trials [1] [2]. During the past decade, mechanistic models have been progressively integrated into the pharmaceutical research and development industry workflow to provide valuable decision support in addition to conventional *in vitro* and *in vivo* approaches [3] [4].

An essential benefit of mechanistic models, when compared to statistical models or machine learning approaches, is that the model equations and associated parameters have a direct physical or biological meaning. Indeed, statistical models are based on the correlation found between variables while mechanistic ones model causality. This facilitates the overall comprehension of the process and the scientific interpretation of model results [5]. Moreover, mechanistic modeling can predict biological or physical behaviors that have not yet been reported by currently available *in vivo* or *in vitro* experiments [6] [7].

However, because of their complexity, and because this approach is more driven by knowledge, which can be considered as consolidated data, and less so by analysis of a small number of raw data from a very limited trial dataset, their credibility is often questioned compared to historical approaches, particularly their capacity to fully reproduce real-world data [8] [9]. For this reason, while the adoption of mechanistic modeling is in use at most major pharmaceutical/biotechnological companies, and its application is accelerating, trust in the relevance of such approaches for predicting novel phenomena is still a work-in-progress [10] [11] [12].

Even if the links between the variables of interest in these models are reported and justified in the literature, the range, the distribution and the correlation of their parameter values are difficult to evaluate. To overcome this problem, calibration is now a standard step in mechanistic model construction. Calibration can be defined as the search for a set of model parameter values that allows the model to reproduce a predefined set of behaviors and dynamics, observed in real life [13]. However, how can we ensure that a model calibrated on several relevant datasets is good enough to be considered as validated and credible for its intended use?

Indeed, as with statistical models, mechanistic models have to be validated in order to confirm that their predictions are reliable and accurate. To avoid tautological bias and improve model credibility, this step requires data that has not been previously used for other purposes, such as model calibration [14] [15].

Model validation is a very topical issue, and is of interest to regulatory agencies. Indeed, the ASME V&V 40 Subcommittee on Verification and Validation (ASME V&V 40) in Computational Modeling of Medical Devices developed a risk-informed credibility assessment framework including a quantitative validation phase [10], and the European Medicines Agency (EMA) has drafted a specific guidance on the reporting of PBPK models including the evaluation of the predictive performance of the drug model [16]. According to these guidelines, the context of use (CoU) of a model must also be defined, which defines the specific role and scope of the model in addressing the questions of interest [17].

The validation on retrospective data requires careful choice of appropriate metrics that take into account the nature of the measurements and the existing variability and bias [18] [19] [20]. In the case of mechanistic models, the outputs are variable dynamics over time which are, most frequently, related to discrete reported observational or experimental values. Such experimental measures show an inherent variability due to the type of instruments that were used, as well as its resolution or sensitivity level, the quality of the sample, the applied protocol, human variability in reporting results and variability between samples [21] [22] [23] [24], that also needs to be considered in the establishment of a validation strategy. In a situation of time-to-event (TTE) data, an additional difficulty can arise. Indeed, the TTE reported in real life corresponds to the moment when the event is detected by the observer, and not to the moment when it really happened. These two moments can be separated by a potentially significant period of time depending on the frequency of observations. Moreover, the model's purpose is to predict the exact time until the occurrence of the event, and not the time to the observation of the event. This concept also has to be taken into consideration during the validation process.

Another goal of the validation process is also to guarantee that the model is not overfitted, which can happen if it was calibrated using datasets with limited variability. Additionally, the validation should as well assure that the variability of predictions of the model is not excessively wide. The latter can be assessed by evaluating the prediction intervals, that is to say, the range within which future observations should fall. Therefore, an adequate validation strategy should prevent both overfitting and underfitting as designing a model with the appropriate complexity requires achieving a balance between bias and variance, as well as a control of overfitting. Otherwise, if the prediction interval is too wide, the model's outputs will lack precision and therefore the model's credibility and usefulness will be low [25].

In order to demonstrate how the validation approach is applied, a case study including TTE oncological data will be described.

In summary, in this article we address the following challenge: how to properly manage the validation process when faced with a multi-condition situation, namely:

- Discrepancy in the size of the data to be compared: indeed, on the one hand, a mechanistic model may produce a very large amount of data. On the other hand, we wish to challenge the model outputs with a limited experimental validation dataset. Issues such as excess of statistical power, discrepancy in variability and uncertainty quantification will likely arise.
- Hypotheses for the application of statistical tests are not always verified producing a lack of statistical power.
- The uncertainty linked to the occurrence of events during clinical studies: the observation time uncertainty (further detailed in the “Methods” section), that is not handled by one deterministic model.

In this article, we focus on quantitative validation, a step of the overall validation process recommended by the regulatory guidelines (ASME V&V 40 [26] and EMA [16]). We introduce multiple methods suited to validate deterministic non-linear mechanistic models including feedback loops, producing a TTE type of outcome. Importantly, these validation approaches consider both the model uncertainties and the variability of validation data.

We first present the methodology behind each one of those approaches, including the pre-processing of the dataset. Then, to answer the question of interest, we design four empirical methods to assess the model’s performance and the reliability of predictions: two methods based on bootstrapped versions of parametric statistical tests (log-rank and combined weighted log-ranks - MaxCombo) and two methods based on bootstrapped prediction intervals (that we named raw coverage and juncture metric). We also introduce the notion of observation time uncertainty (OTU) to take into consideration the delay between the moment when an event actually occurs, and the moment when it is witnessed and reported.

We then present an application on a clinical example. We finally discuss our results, highlight the advantages and disadvantages of these methods according to the application context, compare the performances and conclude.

Methods

The statistical approaches, which are described hereafter, are combined with two additional mathematical concepts in order to better match the actual clinical context of this application example. Thus we first introduce the bootstrap and OTU concepts and then proceed with the actual statistical validation approaches.

Bootstrap

In the context of modeling and simulation, one is not theoretically limited by the number of simulated statistical units (patients). This can be an advantage but also a drawback when using inferential statistics. Indeed, under the assumption of the same variance of the data, the statistical power will increase with the size of the sample [27]. This can lead to a misinterpretation of the results, concluding that there is a statistical difference between compared groups when there is none [28]. In order to control this statistical power, to avoid tests from being overly sensitive to negligible differences between groups, and to take into account the model uncertainty and the variance of the sampling in the simulation results, a bootstrapped version of the statistical tests is recommended [29] [30] [31]. We propose to apply a bootstrapped approach that consists in drawing a sample from the simulation outputs, of the same size as the observed population, then performing the statistical test of interest to compare the virtual sample and the corresponding observed population and storing its result. The output of the bootstrapped approach is the ratio of non-significant tests at a defined alpha risk (set to 5% in our case) out of the total number of performed tests. Note that the proportion of rejections is the estimated empirical power of the used test (bootstrap power)[34] [35] [35]. To determine how many iterations are required, preliminary tests are performed to see how long it takes for the ratio of non-significant tests to become stable. This process is then repeated *n* times, and the ratio of non-significant tests is compared to a given predefined threshold.

Observation Time Uncertainty

The mechanistic models considered here are deterministic. Because we have access by design to the model outputs at all time points, the exact time at which a simulated event takes place can be determined. This model output is named the predicted-time-to-event (PTTE). In real patients, the true TTE can only be bounded between the time of two observations. We do not know the “exact” time-to-event. Therefore, an unknown difference between the PTTE and the reported time-to-event (RTTE) exists, bounded by the time between two observations (Figure 1). This time frame is what will be called the OTU, and depends on the delay between two observations. In other words, the actual TTE could have occurred in a time period ranging from the reported RTTE to the RTTE minus the time elapsed since the previous observation period. Nevertheless, one should keep in mind that we should not expect the model to cover this entire period since there is no evidence that the whole area reflects the real time at which the event occurred. A visual representation of the OTU is presented below.

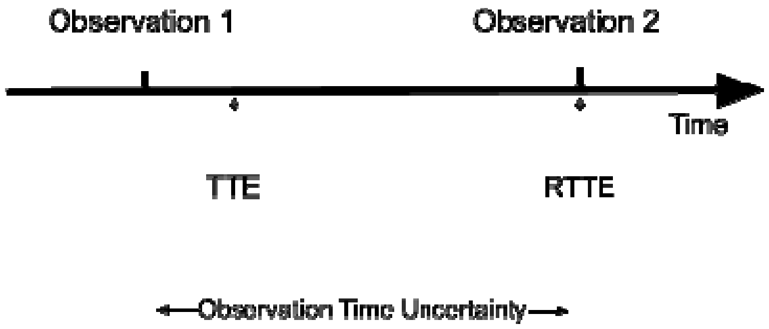


Figure 1: Representation of the observation time uncertainty. If an event happens between observation 1 and observation 2, it will only be reported at the time of observation 2. The observation time uncertainty corresponds to the time between two observations. To note, even though the TTE can theoretically happen any time between two observations, it is unknown whether this is true in a real life context. TTE: time-to-event, RTTE: reported time-to-event

The two concepts introduced above (bootstrap and observation time uncertainty) will be used in combination with the following validation approaches. Their description as well as their advantages and limits will be described.

Raw coverage

In order to perform both a quantitative and a visual validation of the computational model based on the validation dataset, a raw data coverage validation is performed. This approach consists of computing the percentage of the observed curve covered by the prediction interval of the model. In the context of simulation, there is no limit to the number of times one can run the same model, changing a few numbers of parameters and getting a new endpoint value. As a consequence, one can perform a large number of model runs so that there are more available model endpoints values than the number of endpoints values reported within the real population. Therefore, the definition of the prediction interval computed using bootstrapping has been adapted. At each iteration, a sample of simulated endpoints of the same size as the size of the real life population is taken from the set of simulated endpoints, and a Kaplan-Meier (KM) time to progression curve is estimated. This step is repeated *n* times. The 95% prediction interval of the average simulated curve is then calculated. It corresponds to the interval between the 2.5th percentile value and the 97.5th percentile value of the *n*-th simulated TTE curves. The level of coverage of the observed curve with the prediction interval of the simulated curve is then computed: for each time point, a check is performed to see if the observed curve is within the prediction interval - value is set to “True” - or not - value is set to “False” -. The percentage of “True” values is then computed. It is considered that a coverage value of at least 80% is acceptable to consider the model as validated (see Figure 2 based on generated synthetic data for illustrative purposes). The raw coverage approach has the advantage of using the raw observed data without any prior transformation, considering that the real events occurred exactly at the moment of the reported event. In addition to the computed metric, the raw coverage allows one to easily perform a graphical check of the model’s ability to reproduce the observed results.

The fact of considering that the event happened exactly at the time of the observation can also be considered as a limitation, as this is very unlikely. Indeed, the real event most certainly occurred sometime in between observation periods. Another point of concern is that the value of the raw coverage strongly relies on the width of the prediction interval. Indeed, the wider the interval, the more chances for the observed curve to be included in it. This means that if the model produces a lot of variability, then the raw coverage value will most certainly be very high.

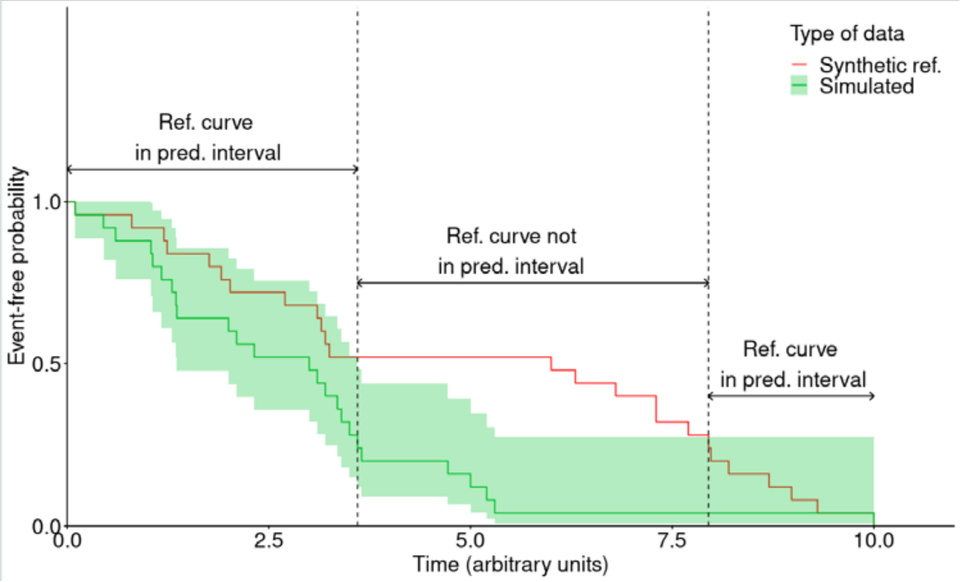


Figure 2: Representation of the raw coverage. The purpose of this example is to illustrate the raw coverage metric based on generated synthetic data. In the time interval of 0 to 10 months, the synthetic reference curve is covered by the prediction interval from t=0 to t=3.6 months, then between t=7.95 and t=10 months. This gives a raw coverage of ((3.6-0) + (10-7.95)) / (10-0) = 56.5%.

Juncture

The juncture approach is similar to the raw coverage in the way that it is both a mathematical and a visual validation method. It differs from the latter by the fact that it takes into account the OTU in the form of an interval. This approach aims to measure the proportion of time over the entire observation period where the observed interval and the 95% prediction interval overlap even if it is only partially. At each time point where observed data is available, a check is performed to see if the two intervals contain common values. The juncture approach metric corresponds to the ratio of time points where this condition is met, over the total number of time points. If the ratio is greater than a predetermined threshold, then, the model is considered to be validated (see Figure 3 based on generated synthetic data for illustrative purposes).

The juncture approach metric takes into consideration the OTU and for this reason, this approach does not rely on the assumption that the event occurred exactly at the time it was reported. Similarly to the raw coverage, it is easy to identify the periods of time during the observation period where the simulation outputs successfully reproduce the observed data.

One of the limits of the metric associated with this approach is that it is strongly dependent on the width of both intervals, that is to say on the variability initially included in the computational model, which can come from the data used to calibrate it, as well as on the size of the OTU. Indeed, the larger the time in between observations, the wider the interval, and vice versa. Moreover, with the juncture approach, even a slight overlap of the two intervals is enough to be considered satisfactory, at a given time point. This means that overall, even if a small fraction of the observed data is covered by the simulated outputs over the entire observation period, then the entire prediction will be considered as validated, even if the latter is shifted up or down compared to the observations. Similarly to the raw coverage, the juncture approach does not rely on a statistical test, with a p-value as an output, but instead on an arbitrary value between 0 and 100%. Similarly to the raw coverage, a value above 80% is considered to be acceptable to judge the model predictions as validated.

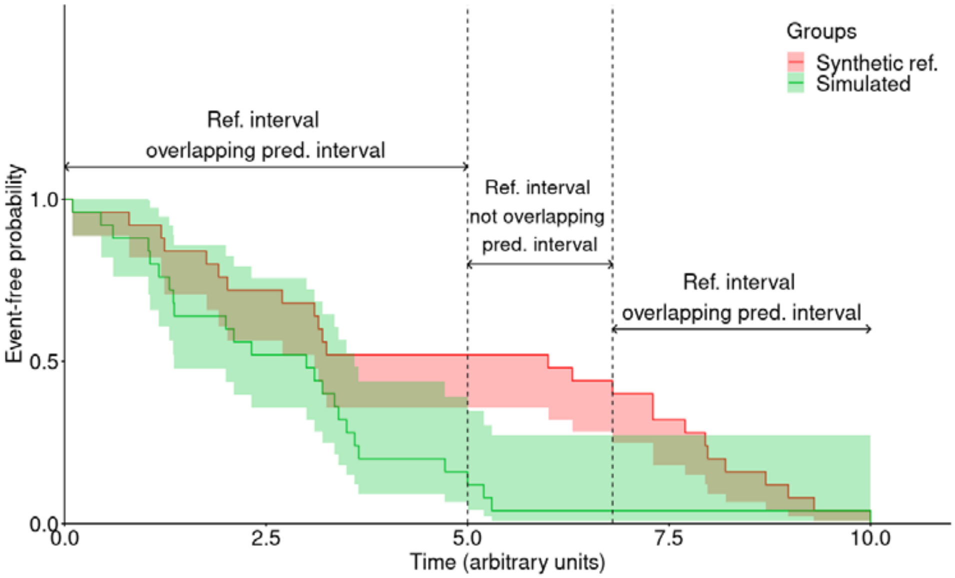


Figure 3: Representation of the juncture metric. The purpose of this example is to illustrate the "juncture" metric based on generated synthetic data. In the time interval of 0 to 10 months the synthetic reference interval overlaps, at least partially with the prediction interval from t=0 to t=5 months, then between 6.8 and 10 months. This results in a juncture of ((5-0) + (10-6.8)) / (10-0) = 82%.

Bootstrapped log-rank test

The log-rank is a well known and widely used test to compare survival curves [35] [36]. Its statistic is based on the computation of the difference between the observed and expected number of events in one of the groups at each observed event time. These differences are then added up to get an overall summary across all-time points where there is an event. The log-rank does not rely on the proportional hazards assumption, that is to say the risk associated with the event of interest remains proportional in both compared groups over the course of the follow-up period. It is a valid test of the null hypothesis of equality of survival functions but it is likely to be less powerful in case of violation of the proportionality of risk assumption and is in this case more effective for detecting the alternative hypothesis [37] [38].

Log-rank's assumptions are the following: the degree of censoring should not be related to the outcome, and the events should have really happened at the reported time.

The log-rank test is integrated into a bootstrapped approach, and is tested first on the raw experimental data. It is then tested again with an OTU sampled from a uniform distribution U(-OTU, 0) being assigned to each real patient, at each iteration.

A given number of bootstrap iterations are performed and the ratio of significant tests at a given alpha risk level is assessed. If this proportion does not exceed a certain predetermined threshold, the model is considered to be validated.

At each iteration proportional hazards assumption is checked, for exploratory purposes [39] [40].

The advantage of the log-rank based validation approach is that it relies on a statistical test frequently used to assess differences between two samples when it comes to TTE data, making its results easy to understand. The proposition to combine the log-rank test with a bootstrap approach, with a sampling of a number of model runs comparable to the number of real patients, prevents it from being excessively sensitive to differences between groups because of an excess of statistical power induced by a very large number of statistical units.

However, because the statistical power of the log-rank test is affected by the proportional hazards assumption, its results might not be considered reliable for samples where the assumption is not met [37]. This implies that if the number of samples where the proportional hazards assumption is not met is high, the ratio of significant tests can be biased. A method more suited for the situations where the proportional hazard hypothesis is not met is introduced in the next section.

In the case where the OTU is not taken into account, the TTE curve based on the sample taken from the simulated data is directly compared to the raw observed data, implying that the reported (RTTE) and the real TTE are equal, which can be considered as a strong assumption.

Bootstrapped weighted log-rank tests combination

Several statistical methods have been developed to better manage the risk of type 1 error and to optimize statistical power in a situation where the proportional hazard assumption is not met [37] [41] [42] [43]. One of these methods is the use of a combination of weighted log-rank tests, called the MaxCombo approach [44] [45]. This approach consists in the use of the Flemming-Harrington family of weights (FH(rho, gamma), rho, gamma ≥ 0). The combination of weights that is used is the following:

- FH(0,0) corresponding to a regular non-weighted log-rank
- FH(1,0) for a log-rank putting more weight on early differences
- FH(1,1) where weights are put on mid-observation differences
- FH(0,1) where late differences are given more weight

Similarly to the log-rank, this approach is also bootstrapped. At each iteration of the bootstrap, all four tests are performed, and the one with the highest z-score, that is to say, the test with the weights showing the largest difference between KM curves, is selected. Given the fact that four tests are performed at once, a Bonferroni correction is applied to the p-value of this test.

The weighted log-rank combination approach is a more robust version of the standard log-rank based one, usable even in the situation where the two compared survival curves cross, implying the proportional hazards assumption is not met. By drawing sub-samples, from the very large simulated population, of the same size as the observed population (a reasonable sample size), we control the occurrence of excessive statistical power related to large samples, preventing the test from being overly sensitive to neglectable differences between the two groups.

As out of the four tests performed at each iteration, it is the one with the highest z-score that is selected, this approach tends to find more differences than a standard Log-rank because more weight is put on the time period where the distance between KM curves is at its maximum. For this reason, the validation acceptance threshold has to be defined accordingly, and should eventually be set lower when compared to other approaches. Similarly to the bootstrapped Log-rank test, this approach is launched twice, first without the OTU, and a second time with a random OTU assigned to real patients at each iteration.

The advantages and limits of the 4 validation methods as well as their variants with OTU are summarized in the table 1 below.

Method	Advantages	Limits
Raw coverage	Based on the reported and non pre-processed data A graphical check can easily be performed to assess the quality of the coverage	Does not take into account the OTU Strongly dependent on the width of the predicted interval Does not rely on a statistical test
Juncture	Takes into consideration the OTU A graphical check can easily be performed to see how well the observed and predicted intervals overlap	Strongly dependent on the width of both observed and predicted intervals A minimal overlap between the two intervals is enough to consider the predictions as validated for a given time point Does not rely on a statistical test
Bootstrapped log-rank (without OTU)	Based on a statistical test frequently used in a TTE context Combined with a bootstrap approach to avoid an excess of statistical power	Does not take into account the OTU Credibility of the result if the proportional hazards assumption is not met
Bootstrapped log-rank (with OTU)	Based on a statistical test frequently used in a TTE context Combined with a bootstrap approach to avoid an excess of statistical power Takes into consideration the OTU	Credibility of the result if the proportional hazards assumption is not met
Bootstrapped combination of weighted log-ranks (without OTU)	Based on an improved version of the log-rank test, more robust in case of non-proportional hazards Combined with a bootstrap approach to avoid an excess of statistical power	Does not take into account the OTU Can be overly sensitive to minor differences because of its design
Bootstrapped combination of weighted log-ranks (with OTU)	Based on an improved version of the log-rank test, more robust in case of non-proportional hazards Combined with a bootstrap approach to	Can be overly sensitive to minor differences because of its design

	avoid an excess of statistical power Same as above Takes into consideration the OTU	

Table 1: Summary of the characteristics of the validation methods (OTU: Observation Time Uncertainty, TTE: Time-To-Event)

Application example: validation of a mechanistic model of lung adenocarcinoma under gefitinib treatment

The methods that were presented in the previous section were assessed and tested on a knowledge-based mechanistic model of the tumor evolution of patients with lung adenocarcinoma.

This model, named the *In Silico* Epidermal growth factor receptor Lung Adenocarcinoma (ISELA), evaluates tumor growth and progression in patients harboring a mutation on the Epidermal Growth Factor Receptor (EGFR), and relies on a mechanistic representation of the lung adenocarcinoma (LUAD) evolution from specific EGFR mutations to clinical outcome. It includes shrinkage in response to the administration of a first generation tyrosine kinase (TKI) drug called gefitinib. This model was calibrated with publicly available data [46] [47] [48] [49] [50] [51], and details regarding the calibration of tumor growth are given in a paper published by Palgen *et al.* [52]. It should be noted that this model is not designed to predict mortality from any cause, but rather developed to predict time to tumor progression (TTP), which was deduced from progression-free and overall survival curves.

In this application context, we focus on the TTP clinical endpoint and will apply our validation strategy to ensure the ISELA model's accuracy on a dataset that was not previously used in the calibration process: the one extracted from Maemondo et *al.* and not previously used for calibration purposes [53]. This study compares the effect of gefitinib versus chemotherapy on NSCLC (of which 90.4% are LUAD) with mutated EGFR. The trial described in the article, and called NEJ002, took place in Japan and gefitinib was used as the first-line treatment. About 90% of the analyzed population had stage IIIb or IV cancers. In this study, gefitinib (250 mg/d) was orally administered once daily, until disease progression, development of intolerable toxic effects, or withdrawal of consent. The progression-free survival (PFS) and the overall survival (OS) curves were manually extracted for patients treated with gefitinib.

Pre-processing of the datasets

A gap was identified between the model output and the dataset related endpoint. While the model represents TTP, which is a clinical endpoint that censors out the patients that die, the dataset extracted from Maemondo et *al.* focuses on PFS and OS. In both clinical endpoints, a patient's death prior to disease progression is therefore an event and is not censored out.

To be able to compare the model TTP to the experimental dataset, the endpoint disease progression was derived from clinical PFS and OS: we manually extracted the KM curves of PFS and OS and their corresponding censored events, and deduced the list of PFS and OS TTEs.

Under the assumption that patients who died before disease progression are characterized by the same time to event in the PFS and OS sets, we are able to filter out PFS events that correspond to patients' death. Indeed, by removing from the PFS values all TTEs that are equal in PFS and OS datasets with a small tolerance due to manual extraction uncertainty, one is left with the TTEs where events are disease progression only. The reduced dataset was named NEJ002 TTP. We consider as equal any PFS and OS values that differ from maximum 2 days.

The NEJ002 TTP dataset is composed of 74 patients, corresponding to 68% of the original dataset, a percentage which seems plausible, considering that the remaining 32% correspond to either censoring, or dead patients. Nevertheless, the exact number was not reported in Maemondo et *al.*. Among the removed data points, 24 correspond to censored events and 10 to death preceding disease progression. Removal of those data points leads to a shift of the curve towards the left. It should be noted nonetheless that the overall linear slope is unchanged (Figure 4).

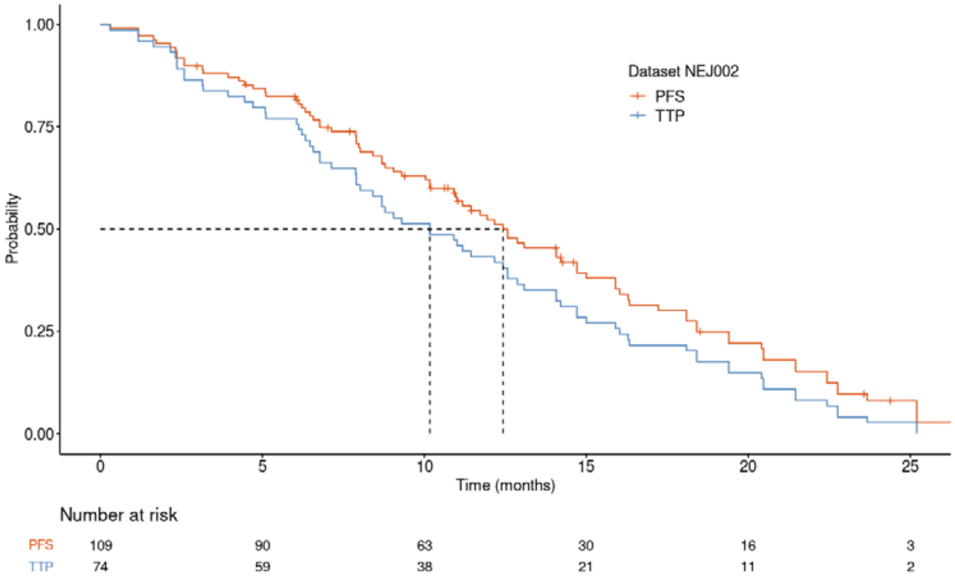


Figure 4: Probability of progression-free survival (red curve) and tumor non-progression (blue curve) from the NEJ002 dataset respectively before and after removal of dead and censored

patients. The dashed line highlights the impact of data-processing on time corresponding to the median probability. Median PFS (12.43 months) and TTP (10.17 months) are represented with dotted lines. PFS data manually extracted from Maemondo et al., processed and plotted in R.

The statistical validation methods described previously were applied to compare the ISELA simulation results to the NEJ002 TTP dataset. For all situations where a bootstrap approach was used, 5000 iterations were performed (cf. Appendix 1), while for approaches based on the Log-rank test, the alpha risk level was set at 5%. The time between visits being 2 months, the OTU used ranged between -2 and 0 months.

Note that the ISELA model represents the tumor growth from which we can deduce the TTP, and only right censoring can be represented by the model.

Results and discussion

Results on entire population

According to the initially defined CoU, the validation approaches were applied to the data corresponding to the entire population extracted from the Maemendo et *al.* article and based on the NEJ002 trial. The results are shown in Fig. 5 and summarized in Table 2.

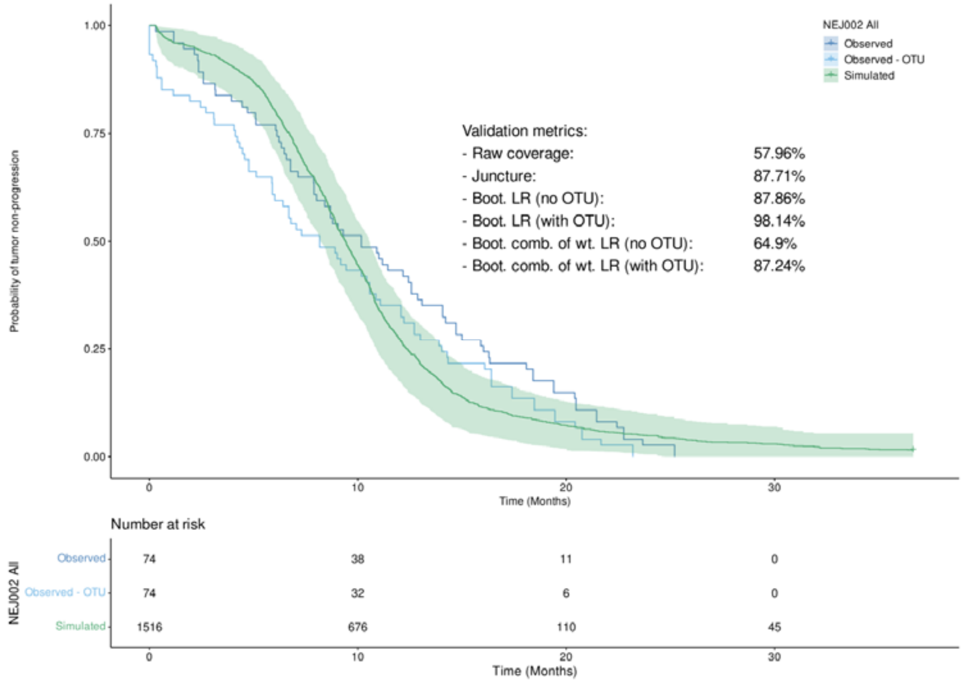


*Figure 5: Observed and simulated Kaplan-Meier curves computed on the **full** dataset. The 95% bootstrapped prediction interval of the simulated curve is represented by the green area. (Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))*

Method	**Metric value**	**Ratio of of samples where PH assumption is not met**
Raw coverage	57.96%	NA
Juncture	87.71%	NA
Bootstrapped LR (no OTU)	87.68%	4.68%
Bootstrapped LR (with OTU)	98.14%	12.92%
Bootstrapped weighted Log-rank combo (no OTU)	64.9%	4.62%
Bootstrapped weighted log-rank combo (with OTU)	87.24%	13.28%

Table 2: Results of the various validation methods applied to the **full** dataset. (PH = proportional hazards)

Note: the acceptability threshold was set at 80%. Given the way all four metrics are defined, the higher the value, the closer the model predictions are to the observed values according to the validation assumptions.

In this context of use, the results provided by the various validation methods vary from 57.96% to 98.14% of validation. Four methods show a metric superior to the chosen threshold of acceptance set at 80%, while the two others fail to reach it. The raw coverage, and the weighted LR based method without OTU fail to reach the validation threshold. The reason for the raw coverage metric to remain below 80% can be explained by the fact that between 2 and 6 months, the model underestimates the number of events, and then overestimates them between 12 and approximately 24 months, as shown in Figure 5. Regarding the weighted LR based approaches, they show that the model's predictions are not accurate while both bootstrapped LR metrics, with and without OTU, indicate that the model is performing well. This difference can be explained by the fact that simulated and observed curves cross, implying that the statistical power of LR tests is reduced, resulting in a lower rejection rate of the null hypothesis, and consequently, a higher validation metric.

Refinement of the context of use

According to previous results and the noticeable discrepancies between methods, in order to show how data structure and the model's CoU can have an impact on the model validation process we decided to go further through the exploration of the data. Indeed, considering the mutational status of the tumor, the data used for validation consist of a mixture of two populations. Each of these subsets was characterized by a specific EGFR mutation: exon 19 deletion (Del19) and L858R on exon 21. Those mutations had an impact on the time to progression, making the simultaneous validation on both types of patients not relevant and potentially incorrect [54] [55] [56]. Thus, In order to have a more precise assessment of the model's predictive capability, the validation process assessment was stratified according to the mutation status of patients.

After applying the validation approaches to the Del19 subset, new metrics were computed and summarized in Figure 6 and Table 3.

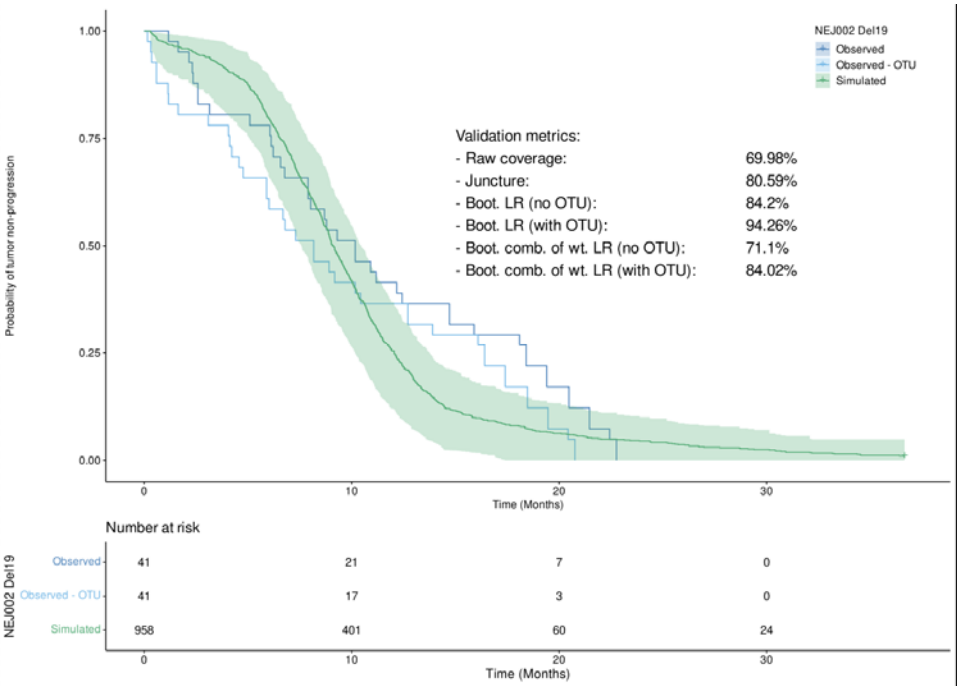


Figure 6: Observed and simulated Kaplan-Meier curves computed on the ***Del19*** *subpopulation. The 95% bootstrapped prediction interval of the simulated curve is represented by the green*

area. (Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))

Method	**Metric value**	**Ratio of of samples where PH assumption is not met**
Raw coverage	69.98%	NA
Juncture	80.59%	NA
Bootstrapped LR (no OTU)	84.2%	12.34%
Bootstrapped LR (with OTU)	94.26%	19.24%
Bootstrapped weighted log-rank combo (no OTU)	71.1%	12.42%
Bootstrapped weighted log-rank combo (with OTU)	84.02%	19.02%

Table 3: results of the various validation methods applied to the **Del19** subset. (PH = proportional hazards)

As for the Del19 subset, the validation metrics on the L858R subset were computed and summarized in Figure 7 and Table 4.

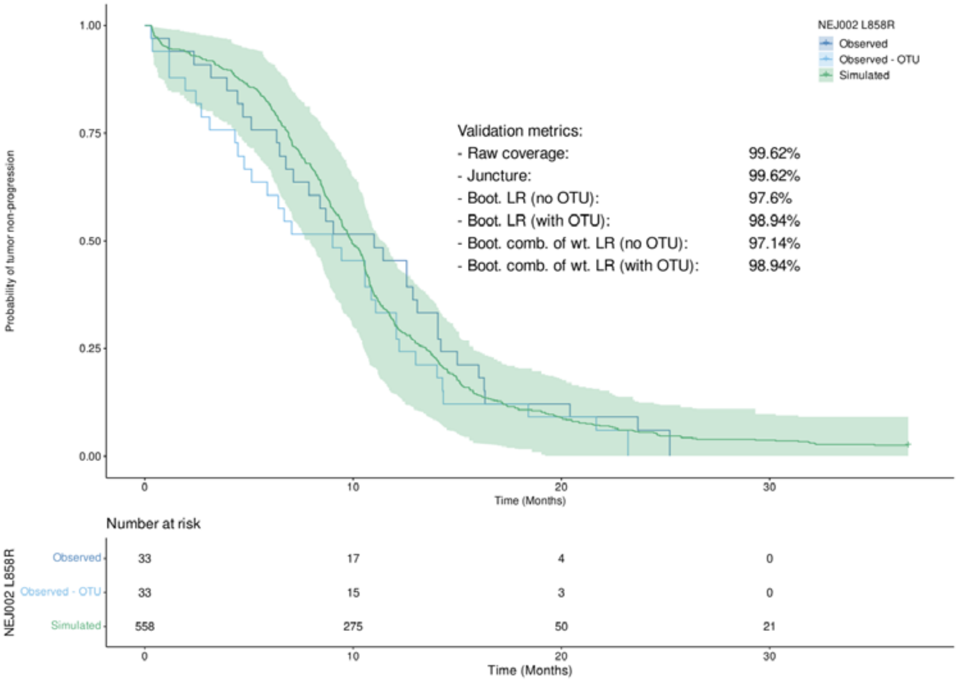


*Figure 7: Observed and simulated Kaplan-Meier curves computed on the **L858R** subset. The 95% bootstrapped prediction interval of the simulated curve is represented by the green area. (Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))*

Method	Metric value	Ratio of of samples where PH assumption is not met
Raw coverage	99.62%	NA
Juncture	99.62%	NA
Bootstrapped LR (no OTU)	97.6%	2.38%
Bootstrapped LR (with OTU)	98.94%	5.32%
Bootstrapped weighted log-rank combo (no OTU)	97.14%	2.8%
Bootstrapped weighted log-rank combo (with OTU)	98.94%	5.7%

Table 4: results of the various validation methods applied to the **L858R** subset. (PH = proportional hazards)

When applied to the Del19 subset, the raw coverage approach provided better results than on the overall population with approximately 12% more coverage of the observed curve. Regarding the juncture method, the value was lower by 7.12% for the subset. A decrease was found as well for both the bootstrapped log-rank (-3.66% without OTU, -3.88% with OTU) and the bootstrapped combination of weighted log-ranks with OTU (-3.22%). The version without the VTU increased by 6.2%. The results obtained on the Del19 subset show that there are even more differences between the validation data and the simulations than in the previous CoU. The model appears to be unable to correctly predict events in this subgroup, despite the better results obtained with the raw coverage approach, which indicate that relying on a single metric is not enough to properly evaluate the quality of the model's predictions. We note here, as an aside, that mismatches such as this help guide model improvement, allowing us to better understand the disease and treatments effects. Without such a model, these discrepancies might not even be noticed.

In the case of the L858R subset, both the raw coverage and juncture methods produced much better results than on the entire population : 99.62% for both approaches, equal to an increase of 29.64% and 19.03%, respectively. With the bootstrapped log-rank, the results were better without the OTU (+13.4%), as well as with the OTU taken into account (+4.68%). For the bootstrapped combination of weighted log-ranks, both metrics without and with OTU were better on the subset than on the global population (+32.24% and +14.92%, respectively). This demonstrates that all validation metrics can show good performances when the CoU is properly chosen. Indeed, it appears that the model is well suited to predict the events in the L858R subgroup, which was not the case for the Del19 subset.

The differences between the metrics obtained on the entire population and on the subsets are summarized in the table 5 below.

Method	**Difference between full dataset and Del19**	**Difference between full dataset and L858R**	**Difference between Del19 and L858R datasets**
Raw coverage	+12.02%	+41.66%	+29.64%
Juncture	-7.12%	+11.91%	+19.03%
Bootstrapped LR (no OTU)	-3.66%	+9.74%	+13.4%
Bootstrapped LR (with OTU)	-3.88%	+0.8%	+4.68%
Bootstrapped weighted log-rank combo (no OTU)	+6.2%	+36.24%	+32.24%
Bootstrapped weighted log-rank combo (with OTU)	-3.22%	+11.7%	+14.92%

Table 5: Differences between the results obtained on the initial dataset and the **Del19** and **L858R** subsets

Conclusion

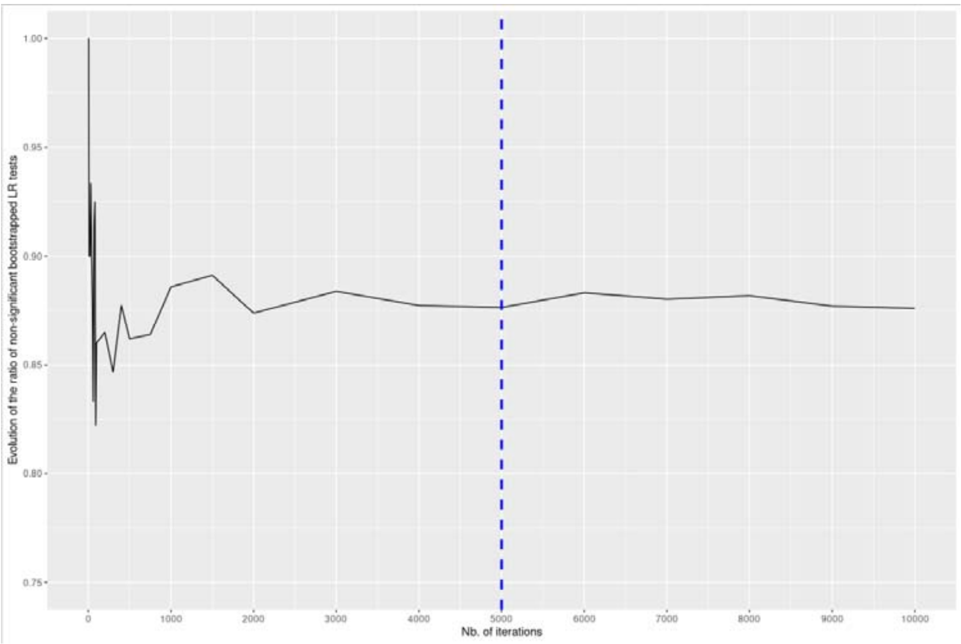
In this article, we introduced different approaches to validate mathematical model predictions on Time-To-Event data, and gave some insight on how to perform a robust validation of a mathematical model by choosing one or multiple methods to correctly evaluate the model's prediction. We have emphasized that the choice of methods and metrics is highly impactful and thus it should be made according to the context, available validation data, and to its specificities, structure and nature (single curve or interval).

We demonstrated in the application section that a model is meant to be applied to a specific context of use (CoU), as otherwise, by performing a validation on an excessively broad dataset, the whole process may fail because the model will not be able to correctly predict the events for heterogeneous subpopulations. Moreover, the importance of using multiple validation methods at once instead of relying on a single one was illustrated by the results obtained on a non-adapted CoU (*e.g.* Del19) where, by looking at only one validation metric (raw coverage in this specific case), one could wrongfully conclude that the model performed well, or at least better than within the previous CoU, while in fact, all the other metrics together demonstrated a worse performance of the model.

Indeed, the strength of the validation process comes from the combination of well selected validation metrics, as each one has its own strengths and weaknesses and conditions of application (*see* Table 1). We highlighted and suggested that simultaneously using multiple methods that rely on different statistical concepts can ensure correct evaluation of the model's performance. Nevertheless, we noted that some of the methods introduced in this article will have more weight than others in a combined approach, because some methods are relatively more robust and more prone to detect differences between observed and simulated data, for example the MaxCombo approach when the assumption of proportional hazard is not met.

It should be noted that during the validation process, it is necessary to avoid tautology, principally by using data not used for the construction of the model, and to avoid trying to validate the model by arbitrarily changing goals, but to define a priori protocol and methods in order to evaluate the model and the context of use properly.

Appendix



Annex 1: Evolution of the ratio of non-significant bootstrapped log-rank tests on the entire population, according to the number of bootstrap iterations. The ratio can be considered as stable after 5000 iterations.

References

1 - Gobburu, J. V., & Lesko, L. J. (2009). Quantitative disease, drug, and trial models. Annual review of pharmacology and toxicology, 49, 291-301.

2 - Tomlin, C. J., & Axelrod, J. D. (2007). Biology by numbers: mathematical modelling in developmental biology. Nature reviews genetics, 8(5), 331-340.

3 - Milligan, P. A., Brown, M. J., Marchant, B., Martin, S. W., Van Der Graaf, P. H., Benson, N., ... & Lalonde, R. L. (2013). Model‐based drug development: a rational approach to efficiently accelerate drug development. Clinical Pharmacology & Therapeutics, 93(6), 502-514.

4 - Courcelles, E., Boissel, J. P., Massol, J., Klingmann, I., Kahoul, R., Hommel, M., Pham, E., & Kulesza, A. (2022). Solving the Evidence Interpretability Crisis in Health Technology Assessment: A Role for Mechanistic Models?. Frontiers in medical technology, 4, 810315.

5 - Wang, Y., Zhu, H., Madabushi, R., Liu, Q., Huang, S. M., & Zineh, I. (2019). Model‐informed drug development: current US regulatory practice and future considerations. Clinical Pharmacology & Therapeutics, 105(4), 899-911.

6 - Dronne, M. A., Grenier, E., Chapuisat, G., Hommel, M., & Boissel, J. P. (2008). A modelling approach to explore some hypotheses of the failure of neuroprotective trials in ischemic stroke patients. Progress in biophysics and molecular biology, 97(1), 60-78.

7 - Gal, J., Milano, G., Ferrero, J. M., Saâda-Bouzid, E., Viotti, J., Chabaud, S., ... & Chamorey, E. (2018). Optimizing drug development in oncology by clinical trial simulation: Why and how?. Briefings in bioinformatics, 19(6), 1203-1217.

8 - Musuamba, F. T., Bursi, R., Manolis, E., Karlsson, K., Kulesza, A., Courcelles, E., ... & Geris, L. (2020). Verifying and validating quantitative systems pharmacology and in silico models in drug development: current needs, gaps, and challenges. CPT: Pharmacometrics & Systems Pharmacology, 9(4), 195.

9 - Musuamba, F. T., Skottheim Rusten, I., Lesage, R., Russo, G., Bursi, R., Emili, L., ... & Geris, L. (2021). Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. CPT: Pharmacometrics & Systems Pharmacology, 10(8), 804-825.

10 - Viceconti, M., Juárez, M. A., Curreli, C., Pennisi, M., Russo, G., & Pappalardo, F. (2019). Credibility of in silico trial technologies—a theoretical framing. IEEE Journal of Biomedical and Health Informatics, 24(1), 4-13.

11 - Kuemmel, C., Yang, Y., Zhang, X., Florian, J., Zhu, H., Tegenge, M., ... & Zineh, I. (2020). Consideration of a credibility assessment framework in model‐informed drug development: potential application to physiologically‐based pharmacokinetic modeling and simulation. CPT: Pharmacometrics & Systems Pharmacology, 9(1), 21-28.

12 - Baker, R. E., Pena, J. M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community?. Biology letters, 14(5), 20170660.

13 - Gadkar, Kapil, D. C. Kirouac, D. E. Mager, Piet H. van der Graaf, and Saroja Ramanujan. "A six‐stage workflow for robust application of systems pharmacology." CPT: pharmacometrics & systems pharmacology 5, no. 5 (2016): 235-249.

14 - Oberkampf, W. L., & Roy, C. J. (2010). Verification and validation in scientific computing. Cambridge University Press.

15 - Dahabreh, I. J., Chan, J. A., Earley, A., Moorthy, D., Avendano, E. E., Trikalinos, T. A., ... & Wong, J. B. (2017). A review of validation and calibration methods for health care modeling and simulation. Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment.

16 - European Medicines Agency. (2018). Guideline on the reporting of physiologically based pharmacokinetic (PBPK) modeling and simulation, Appendix 2 p.15/16. Retrieved February 25, 2022, from https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-reporting-physiologically-based-pharmacokinetic-pbpk-modelling-simulation_en.pdf

17 - Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., & Tshinanu, F. M. (2021). In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. Methods, 185, 120-127.

18 - Rajamanickam, V., Babel, H., Montano-Herrera, L., Ehsani, A., Stiefel, F., Haider, S., ... & Knapp, B. (2021). About model validation in bioprocessing. Processes, 9(6), 961.

19 - Kirouac, D. C. (2018). How do we “validate” a QSP model?. CPT: Pharmacometrics & Systems Pharmacology, 7(9), 547.

20 - Hasdemir, D., Hoefsloot, H. C., & Smilde, A. K. (2015). Validation and selection of ODE based systems biology models: how to arrive at more reliable decisions. BMC systems biology, 9(1), 1-19.

21 - Altman, N., & Krzywinski, M. (2015). Points of significance: Sources of variation. Nature methods, 12(1).

22 - Blainey, P., Krzywinski, M., & Altman, N. (2014). Points of significance: replication. Nature methods, 11(9), 879.

23 - Lin-Gibson, S., Sarkar, S., Elliott, J., & Plant, A. (2016). Understanding and managing sources of variability in cell measurements. Cell Gene Ther. Insights, 2(6), 663-73.

24 - McCormack, J. P., & Holmes, D. T. (2020). Your results may vary: the imprecision of medical measurements. Bmj, 368.

25 - Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting. Nature methods, 13(9), 703-705.

26 - ASME V&V 40, 2018 Edition, November 19, 2018 - Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices

27 - Houser, J. (2007). How many are enough? Statistical power analysis and sample size estimation in clinical research. Journal of Clinical Research Best Practices, 3(3), 1-5.

28 - Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. Journal of graduate medical education, 4(3), 279-282.

29 - Zhu, W. (1997). Making bootstrap statistical inferences: A tutorial. Research quarterly for exercise and sport, 68(1), 44-55.

30 - Horowitz, J. L. (2001). The bootstrap. In Handbook of econometrics (Vol. 5, pp. 3159-3228). Elsevier.

31 - Brownstone, D., & Valletta, R. (2001). The bootstrap and multiple imputations: harnessing increased computing power for improved statistical tests. Journal of Economic Perspectives, 15(4), 129-141.

32 - Kleinman, K., & Huang, S. S. (2016). Calculating power by bootstrap, with an application to cluster-randomized trials. EGEMs, 4(1).

33 - Walters, S. J., & Campbell, M. J. (2005). The use of bootstrap methods for estimating sample size and analysing health‐related quality of life outcomes. Statistics in medicine, 24(7), 1075-1102.

34 - Wang, Z. (2019). Comparison of sample size by bootstrap and by formulas based on normal distribution assumption. Therapeutic Innovation & Regulatory Science, 53(2), 170-175.

35 - Bland, J. M., & Altman, D. G. (2004). The logrank test. Bmj, 328(7447), 1073.

36 - Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep, 50, 163-170.

37 - Royston, P., & B Parmar, M. K. (2020). A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. Trials, 21(1), 1-17.

38 - Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One, 10(1), e0116774.

39 - Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515-526.

40 - Keele, L. (2010). Proportionally difficult: testing for nonproportional hazards in Cox models. Political Analysis, 18(2), 189-205.

41 - Li, H., Han, D., Hou, Y., Chen, H., & Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One, 10(1), e0116774.

42 - Karrison, T. G. (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. The Stata Journal, 16(3), 678-690.

43 - Royston, P., & Parmar, M. K. (2016). Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Medical Research Methodology, 16(1), 1-13.

44 - Prior, T. J. (2020). Group sequential monitoring based on the maximum of weighted log-rank statistics with the Fleming–Harrington class of weights in oncology clinical trials. Statistical Methods in Medical Research, 29(12), 3525-3532.

45 - Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., ... & Cross-Pharma Non-Proportional Hazards Working Group. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. Statistics in Biopharmaceutical Research, 12(2), 187-198.

46 - Asahina, H., Yamazaki, K., Kinoshita, I., Sukoh, N., Harada, M., Yokouchi, H., ... & Nishimura, M. (2006). A phase II trial of gefitinib as first-line therapy for advanced non-small cell lung cancer with epidermal growth factor receptor mutations. British journal of cancer, 95(8), 998-1004.

47 - Yang, C. H., Yu, C. J., Shih, J. Y., Chang, Y. C., Hu, F. C., Tsai, M. C., ... & Yang, P. C. (2008). Specific EGFR mutations predict treatment outcome of stage IIIB/IV patients with chemotherapy-naive non–small-cell lung cancer receiving first-line gefitinib monotherapy. Journal of Clinical Oncology, 26(16), 2745-2753.

48 - Wu, J. Y., Wu, S. G., Yang, C. H., Gow, C. H., Chang, Y. L., Yu, C. J., ... & Yang, P. C. (2008). Lung cancer with epidermal growth factor receptor exon 20 mutations is associated with poor gefitinib treatment response. Clinical Cancer Research, 14(15), 4877-4882.

49 - Vasconcelos, P. E., Gergis, C., Viray, H., Varkaris, A., Fujii, M., Rangachari, D., ... & Costa, D. B. (2020). EGFR-A763_Y764insFQEA is a unique exon 20 insertion mutation that displays sensitivity to approved and in-development lung cancer EGFR tyrosine kinase inhibitors. JTO clinical and research reports, 1(3), 100051.

50 - Yasuda, H., Park, E., Yun, C. H., Sng, N. J., Lucena-Araujo, A. R., Yeo, W. L., ... & Costa, D. B. (2013). Structural, biochemical, and clinical characterization of epidermal growth factor receptor (EGFR) exon 20 insertion mutations in lung cancer. Science translational medicine, 5(216), 216ra177-216ra177.

51 - Sugio, K., Uramoto, H., Onitsuka, T., Mizukami, M., Ichiki, Y., Sugaya, M., ... & Yasumoto, K. (2009). Prospective phase II study of gefitinib in non-small cell lung cancer with epidermal growth factor receptor gene mutations. Lung Cancer, 64(3), 314-318.

52 - Palgen, J. L., Perrillat-Mercerot, A., Ceres, N., Peyronnet, E., Coudron, M., Tixier, E., ... & Monteiro, C. (2022). Integration of heterogeneous biological data in multiscale mechanistic model calibration: application to lung adenocarcinoma. bioRxiv.

53 - Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., ... & Nukiwa, T. (2010). Gefitinib or chemotherapy for non–small-cell lung cancer with mutated EGFR. New England Journal of Medicine, 362(25), 2380-2388.

54 - Sheng, M., Wang, F., Zhao, Y., Li, S., Wang, X., Shou, T., ... & Tang, W. (2016). Comparison of clinical outcomes of patients with non-small-cell lung cancer harbouring epidermal growth factor receptor exon 19 or exon 21 mutations after tyrosine kinase inhibitors treatment: a meta-analysis. European journal of clinical pharmacology, 72(1), 1-11.

55 - Zheng, Z., Xie, D., Su, H., Lin, B., Zhao, L., Deng, X., ... & Xie, C. (2017). Treatment outcome comparisons between exons 19 and 21 EGFR mutations for non-small-cell lung cancer patients with malignant pleural effusion after first-line and second-line tyrosine kinase inhibitors. Tumor Biology, 39(6), 1010428317706211.

56 - Lee, C. K., Wu, Y. L., Ding, P. N., Lord, S. J., Inoue, A., Zhou, C., ... & Yang, J. C. H. (2015). Impact of specific epidermal growth factor receptor (EGFR) mutations and clinical characteristics on outcomes after treatment with EGFR tyrosine kinase inhibitors versus chemotherapy in EGFR-mutant lung cancer: a meta-analysis. Journal of Clinical Oncology, 33(17), 1958-1965.