Poster 4287

The combination of statistical methods to compare observed and simulated data allowed to assess effectively the validity of mathematical model predictions in a context of a EGFR+ Lung Adenocarcinoma

Evgueni Jacob[1], Laura Villain[1], Nicoletta Ceres[1], Jean-Louis Palgen[1], Adèle L'Hostis[1], **Claudio Monteiro**[1], Riad Kahoul[1]
[1]Novadiscovery, Lyon, France
Contact: claudio.monteiro@novadiscovery.com

BACKGROUND

In silico models proved to be a promising tool to complement and optimize clinical trials. These models should be validated to assess their capacity to reproduce real life behaviors. Opposed to real-life clinical trials where the amount of available data might be low, *in silico* approaches give us the possibility to simulate virtual populations of thousands of patients. This data size heterogeneity might be an issue. Moreover, in real life data, patients are seen at scheduled visits, leading to an observation time uncertainty (OTU), which is not the case in simulated data.
In the context of the validation of an EGFR mutant Lung Adenocarcinoma mathematical model, we depicted the interest of using combined validation methods to assess the capacity of the model to predict the time to tumor progression, from heterogeneous clinical trials datasets. Furthermore, we demonstrated that a model is meant to be applied to a specific context of use (CoU) via an exploration of the model's prediction capacity on subsets of the original data used for the validation.

METHODS

In this context, adapted versions of usual statistical methods have been used for the analysis of time-to-event (TTE) data. Those approaches rely on the use of two additional mathematical concepts in order to better match the actual clinical context of this application example:

- **Bootstrapping**: In the context of modeling and simulation, one is not limited by the number of simulated statistical units (patients), leading to an excess of statistical power. We applied a bootstrapped approach that consists in drawing a sample from the outputs, of the same size as the observed population, then performing the statistical test or computing a prediction interval to compare the virtual sample and the corresponding observed population.
- **Observation time uncertainty**: The mechanistic models allow computing the exact time at which a simulated event takes place. In real patients, the true TTE can only be bounded between the time of two observations. This time frame named OTU depends on the delay between two visits.

These two statistical concepts were combined with 4 different methods to perform the validation step, as illustrated in Figure 1:

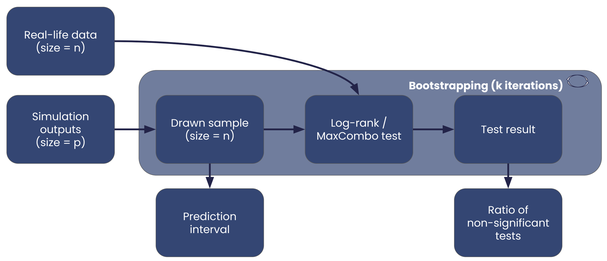

Figure 1: Flowchart of validation metrics computation

- **2 methods based on the log-rank test** where the ratio of non-significant tests at a given alpha risk level is assessed, with and without taking into account the OTU: the "default log-rank test" and the "modified test" based on a combination of weighted log-rank tests (MaxCombo approach). [1]
- **2 methods based on prediction intervals**. The "raw coverage", corresponding to the proportion of the observed TTE curve included in the prediction interval and the "juncture metric", which corresponds to the observation period proportion where the prediction interval overlaps with an interval bound between observed data and the same data shifted by the OTU. [2]

RESULTS

Results (1 - Whole population)

The validation was performed initially on the entire real-life dataset (NEJ002) [3].

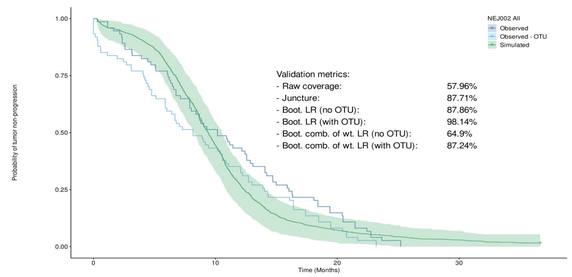

Figure 2: Observed and simulated Kaplan-Meier curves computed on the full dataset. The 95% bootstrapped prediction interval of the simulated curve is represented by the green area.
(Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))

The results displayed above showed noticeable discrepancies between validation metrics. In order to explore 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 the data used for validation consists in a mixture of two populations, characterized by a specific EGFR mutation:

- Exon 19 deletion (**Del19**)
- **L858R** on exon 21.

Those mutations had an impact on the time to progression (TTP), making the simultaneous validation on both types of patients not relevant and incorrect. 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.

Results (2 - Del19 mutation only)

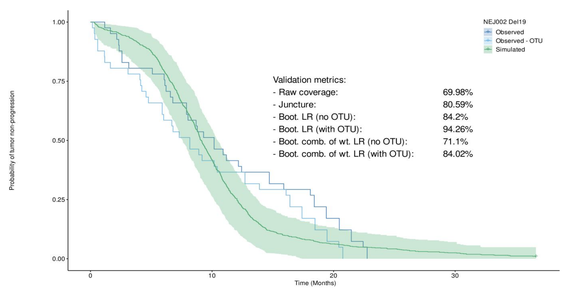

Figure 3: Observed and simulated Kaplan-Meier curves computed on the Del19 subpopulation. The 95% bootstrapped prediction interval of the simulated curve is represented by the green area.
(Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))

Results (3 - L858R mutation only)

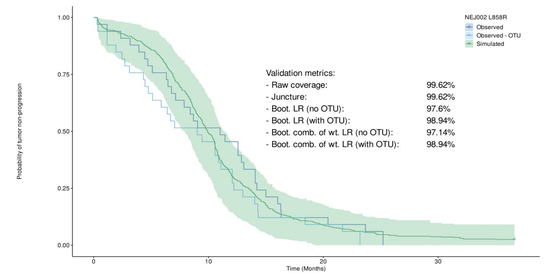

Figure 4: Observed and simulated Kaplan-Meier curves computed on the L858R subset. The 95% bootstrapped prediction interval of the simulated curve is represented by the green area.
(Boot. = Bootstrapped, LR = log-rank test, comb. of wt. LR = combination of weighted log-rank tests (MaxCombo))

DISCUSSION & CONCLUSION

The results showed that the originally unsatisfactory results obtained on the whole population were related to the fact that the model was better suited to predict the TTP in patients with L858R mutation, but not Del19 one (cf. table 1).

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 1: Comparison of validation metrics computed on full dataset and on mutation-specific datasets.

The validation process is of utmost importance to assess the level of credibility of a model and to refine its CoU. The simultaneous use of multiple metrics highlights the eventual flaws in model predictions and/or the misuse of validation datasets, that would not have been detected by a single approach. With this application, we therefore showed that using combination of validation approaches can provide relevant insights for model evaluation.
In addition to qualitative validation steps, as outlined in the V&V40 [4], quantitative model validation ensures that the model predictions are reliable for a given CoU. Once validated, the model can be used to explore hypotheses, and simulate virtual clinical trials to inform their real-life counterparts.

ACKNOWLEDGMENTS

The authors would like to thank E. Peyronnet, H. Darré, A. Perrillat-Mercerot, F. Hammami, B. Martin, D. Lefaudeux, C. Couty, A. Nativel and P. Masson who helped develop the model at nova and Janssen-Cilag France for the support to the lung adenocarcinoma modeling project.

REFERENCES

1. 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.
2. Jacob, E., Perrillat-Mercerot, A., Palgen, J. L., L'Hostis, A., Ceres, N., Boissel, J. P., ... & Kahoul, R. (2022). Empirical methods for the validation of Time-To-Event mathematical models taking into account uncertainty and variability: Application to EGFR+ Lung Adenocarcinoma. bioRxiv, 2022-09.
3. 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.
4. ASME V&V 40, 2018 Edition, November 19, 2018 - Assessing Credibility of Computational Modeling Through Verification and Validation: Application to Medical Devices