EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

13 April 2023
EMADOC-1700519818-1064889
Executive Director

Letter of Support for the statistical adjustment on deep learning prognosis covariates obtained from histological slides

On 23/10/2020 and 16/06/2022, the Applicant Owkin France requested scientific advice for their statistical adjustment on deep learning prognosis covariates obtained from histological slides pursuant to Article 57(1)(n) of Regulation (EC) 726/2004 of the European Parliament and of the Council.

'Statistical adjustment on deep learning prognosis covariates obtained from histological slides' is proposed as a method using the predictions of two deep-learning models based on baseline histology digitized slides as prognostic biomarkers for the adjustment of efficacy analysis on overall survival of life-prolonging drugs in randomized phase 2 and phase 3 clinical trials.

For the initial advice, a discussion meeting with the Applicant took place on 06/04/2021.

On 06/05/2021, the SAWP agreed on the initial advice to be given to the Applicant.

On 20/05/2021, the CHMP adopted the initial advice to be given to the Applicant.

For the follow-up advice, a discussion meeting with the Applicant took place on 29/11/2022.

On 09/02/2023, the SAWP agreed on the follow-up advice to be given to the Applicant.

On 23/02/2023, the CHMP adopted the follow-up advice to be given to the Applicant.

This letter of support is based on the qualification advice provided to the Applicant to encourage further work to enable a future qualification of the proposed methodology.

Background and Context of Use

The Applicant has trained neural networks using digitized pathology imaging from mesothelioma and HCC patients in order to predict overall survival in those two populations. The resulting models are referred to as MesoNet and HCCnet. Training of the network is performed with an algorithm named CHOWDER or with a variant called SCHMOWDER. It uses large pre-trained models to extract information from images and focuses on small regions of the biopsy that are most relevant to prediction. The two cohorts used to train neural networks are the MESOBANK cohort (including 2,981 pleural mesothelioma patients from multiple French institutions) and a cohort from the Henri Mondor hospital (including 194 HCC patients who underwent surgical resection). In both cases, the models brought a significant improvement in predictive performance for OS when added to currently used stratification factors. External validation was performed using testing datasets from The Cancer Genome Atlas (TCGA) with 56 mesothelioma patients and a digitized cohort of the National Virtual Mesothelioma Bank (138 patients), as well as 328 HCC patients who underwent surgical resection.

The proposed Context of Use is:

In the general area of baseline histology information for use in statistical analysis in clinical trial, we propose that the predictions of two deep-learning models based on baseline histology digitized slides be used as prognostic biomarkers for the adjustment of efficacy analysis on overall survival of life-prolonging drugs in randomized phase 2 and phase 3 clinical trials.

In particular, we propose that Mesonet predictions be used as an adjustment of efficacy analysis on overall survival of life-prolonging drugs in first-line malignant pleural mesothelioma patients in randomized phase 2 and phase 3 clinical trials.

We also propose that HCCnet predictions be used as an adjustment of efficacy analysis on overall survival of life-prolonging drugs in the adjuvant setting for resected hepatocellular carcinoma patients in randomized phase 2 and phase 3 clinical trials.

This adjustment will be done either by including deep-learning prognostic score in the Cox regression, by stratifying the analysis on a discretized version of the score or by including the covariates in another method of analysis.

Discussion of the applications of the Artificial Intelligence models

A main potential strength of the artificial intelligence (AI) models is the gain in prognostic performance compared to an adjustment with covariates used in current practice in clinical trial settings. This gain in performance could translate to gains in statistical power. As the AI models are automated, they are not impacted from inter-individual variability of pathology reads from humans compared with the grading or subtyping done by pathologists. The interpretability features of the CHOWDER or SCHMOWDER algorithms, highlighting image tiles used for prediction and associated respectively with good or bad prognosis, allows insights into information used by the AI models and avoids the drawbacks of non-transparent AI models. The proposed AI model architecture includes an 'attention mechanism' that focuses on the parts of the input that includes tumoral information and information most predictive for survival and imposes the models to use only a limited number of tiles for predictions. This can contribute to robustness of predictions regarding the size of the biological specimen as well as to artefacts.

The evidence base for the AI model approach as new technology is limited compared to traditional approaches, as e.g., histological subtyping of mesothelioma tumours. There is also potential for sensitivity to technical biases (e.g., staining protocol, scanner) that can lead to reduced performance in a new setting. While it is noted that the validation approach was done partly in a multicentre setting, no prospective approach to validation was used and calendar time impact on performance was not analysed in the validation approach. Therefore, uncertainty with regard to generalisability of the AI models to future multicentre pivotal clinical trial settings when using the proposed adjusted analysis with covariates for a primary endpoint remains. The Applicant is encouraged to use their approach in future trials and do additional prospective validation.

Discussion of statistical methodology

It is common practice in clinical trials to account for baseline factors that are relevant for the assessment of clinical endpoints (prognostic factors as well as other baseline variables) in design and analysis. Baseline variables that are prognostic for the outcome of patients explain to a certain extent the variation in the endpoint. Adjusting for a prognostic baseline variable may potentially lead to more precise estimates of the treatment effect on overall survival for the two AI models proposed. This approach is broadly supported for “true prognostic factors for which there is a commonly accepted pathophysiological rationale” by the Guideline on adjustment for baseline covariates in clinical trials (EMA/CHMP/295050/2013). A strong pathophysiological rationale for including imaging information as prognostic information is plausible, considering that prognostic value of histology information from image assessments by trained pathologists is established and including this information as baseline covariate is current practice in clinical trial settings.

To show which level of performance gain in terms of power improvement or reduction of sample size can be expected in a range of clinical trial settings, the Applicant performed simulations with a parametric time-to event model. The model uses a single adjustment covariate base on c-index allowing an estimation of the impact of the c-index on performance and cumulative incidence of the event of interest predicted. Other parameters varied in the simulation study are the size of the treatment effect, the Weibull shape of the baseline hazard function and the drop-out rate. The simulations used 5 years follow-up for subject before censoring, which can be regarded a realistic time frame. In the data set with simulated survival times, the presence of a treatment effect was tested in an unadjusted analysis and an analysis adjusted for the AI based prognostic covariate using the Wald test for the treatment coefficient in a Cox regression. The statistical power for the unadjusted analysis and the adjusted analysis was estimated.

To evaluate the added value of AI model covariate adjustment through performance gain regarding improvement in power (or reduction in sample size), the Applicant performed simulations using a simulated population of subjects based on the test data sets used for validation of the respective prognostic covariate model, the Mesobank test data for mesothelioma and the TCGA cohort. These can be regarded ‘semi-synthetic’ datasets. The simulations follow the same principles as outlined for the Cox model above. For each patient the covariates were sampled from the respective test data (Mesobank and TCGA, respectively) and the empirical survival function (defined by hazard rate and baseline survival with Kaplan-Meier estimation) is used instead of Weibull distribution. The simulated sample size is based on recent trials in the mesothelioma and early-stage HCC settings. For assessing the gain when adding the AI based prognostic covariate, the Cox model with ‘standard covariates’ in recent clinical trials according to the Applicant (histology and sex for mesothelioma, tumour staging and ECOG status for HCC) was used with and without the AI based prognostic covariate.

The simulation studies performed as outlined rely on the proportional hazard assumption. In case of non-proportional hazards there is the risk of a considerable increase in Type 1 error (Jiang H et al., Stat Med 2008) when using adjusted Cox regression. From the regulatory perspective the properties of adjusted Cox regression are unfavourable, as several assumptions including proportional hazards, correct model specification, and no interactions of covariates with treatment or other covariates are necessary. Specifically in situations where adjusted results would differ from unadjusted results concerns regarding assumptions could arise (Tangen CM and Koch CG, J Biopharm Stat 1999). The standard approach to analysis of time-to-event data in clinical trials in oncology is to use a log-rank test, with or without stratification factors, for hypothesis testing for a primary time-to-event endpoint. Cox Models with adjustment for covariates are only used for estimation of treatment effects. As alternative, the Applicant proposes to use AI based covariates in stratified analysis with Cox models or other analysis methods. A stratified analysis using discrete covariate levels may need weaker assumptions, but concerns with interpretation remain. Thresholds for a discrete covariate of the deep learning prognostic score are not proposed and may be difficult to justify. There is currently no regulatory agreement on suitability of alternative analysis methods that would allow inclusion of prognostic covariates for use as primary analysis. The application of the MesoNet and HCCNet models in simulation studies was investigated under the assumption of proportional hazards. This is considered a relevant limitation to generalisability.

Regarding risk of Type 2 error increase with inclusion of covariates with no prognostic value and a potentially underperforming model, a cautious approach should be chosen by Applicants specifically in case of small sample sizes.

Conclusions

When applying the proposed AI models for statistical adjustment in primary analysis of clinical trial data, there are relevant regulatory concerns regarding risk of increased Type 1 error with non-proportional hazards and issues with interpretability of adjusted and stratified Cox regression. While an alternative analysis method may be more appropriate, demonstration of gains by simulation relies on adjusted Cox regression and the proportional hazard assumption.

The observed results for prognostic performance of the AI models and the overall approach to training and validation with independent historical data sets are acknowledged. Strengths of the AI prognostic models supporting automated whole slice histology image evaluation and interpretability of results are also acknowledged. Uncertainty regarding generalisability of the AI models to a future multicentre pivotal clinical trial setting remains. The Applicant is encouraged to use their approach in future trials and do additional prospective validation.

Therefore, currently the method cannot be qualified by the EMA until further validation work has been performed.

EMA issues this Letter to support further development.

Sincerely,

Emer Cooke

Executive Director