Individual risk prediction: comparing Random Forests with Cox proportional-hazards model by a simulation study

Valia Baralou[1*], Natasa Kalpourtzi[1*], and Giota Touloumi[1]

[1]Department of Hygiene, Epidemiology & Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece

[*]These authors contributed equally to this work.

July 20, 2021

Machine Learning in healthcare

- Big data in healthcare
  - electronic health records
  - omics data

Machine Learning in healthcare

- Big data in healthcare
  - electronic health records
  - omics data
- Machine Learning (ML) algorithms: an alternative approach to classic statistical methods in numerous biomedical research fields, such as individual risk prediction.

Methodological approaches for individual risk prediction

- In traditional approaches, regression models have been applied to construct a points-based risk score by transforming the regression coefficients for each predictor.
  - Cox proportional hazards (Cox-PH) model
  - Weibull model

Methodological approaches for individual risk prediction

- In traditional approaches, regression models have been applied to construct a points-based risk score by transforming the regression coefficients for each predictor.
    - Cox proportional hazards (Cox-PH) model
    - Weibull model
- In recent years, ML algorithms have been employed to predict individual risk of developing a disease.
    - Artificial Neural Networks
    - Random Forest (RF)
    - Support Vector Machines

ML in survival analysis

- In the majority of the applications, the ML algorithms used ignore the time until the event occurrence.

ML in survival analysis

- In the majority of the applications, the ML algorithms used ignore the time until the event occurrence.
- Due to the presence of censored data, the standard statistical approach is to develop a regression model that takes into account the time until the event occurrence with Cox-PH model being the most popular.

ML in survival analysis

- In the majority of the applications, the ML algorithms used ignore the time until the event occurrence.
- Due to the presence of censored data, the standard statistical approach is to develop a regression model that takes into account the time until the event occurrence with Cox-PH model being the most popular.
- Random survival forest (RSF) seems to be quite popular in survival modeling of long-term outcomes.

Gap in literature

- Although a considerable number of previous studies have compared RSF's performance with Cox-PH on real data, only a few studies have used simulations for the comparison.

Gap in literature

- Although a considerable number of previous studies have compared RSF's performance with Cox-PH on real data, only a few studies have used simulations for the comparison.
- No study has compared the performance of RF that does not incorporate time-to-event information with RSF that can handle right-censored data.

Aim

In this study, we performed extensive simulation experiments to assess:

Aim

In this study, we performed extensive simulation experiments to assess:

- whether RSF outperforms Cox-PH in cases of complex relationships between the covariates and the outcome

Aim

In this study, we performed extensive simulation experiments to assess:

- whether RSF outperforms Cox-PH in cases of complex relationships between the covariates and the outcome
- whether different split criteria used by RSF significantly impact its performance

Aim

In this study, we performed extensive simulation experiments to assess:

- whether RSF outperforms Cox-PH in cases of complex relationships between the covariates and the outcome
- whether different split criteria used by RSF significantly impact its performance
- whether RF's performance is significantly inferior to the survival models in time-to-event analysis.

Illustrative example

- Data from the AMACS study ("Athens multicenter AIDS cohort study"), a large cohort study of people living with HIV (PLHIV).

Illustrative example

- Data from the AMACS study ("Athens multicenter AIDS cohort study"), a large cohort study of people living with HIV (PLHIV).
- Given the increased life expectancy in PLHIV, estimating the risk of developing a cardiovascular (CVD) event is of high importance for the assessment and management of CVD modifiable risk factors.

Illustrative example

- Data from the AMACS study ("Athens multicenter AIDS cohort study"), a large cohort study of people living with HIV (PLHIV).
- Given the increased life expectancy in PLHIV, estimating the risk of developing a cardiovascular (CVD) event is of high importance for the assessment and management of CVD modifiable risk factors.
- Well-established CVD risk scores have been developed using traditional statistical methods. However, their performance might be poor under more complex associations in PLHIV.

Illustrative example

- Data from the AMACS study ("Athens multicenter AIDS cohort study"), a large cohort study of people living with HIV (PLHIV).
- Given the increased life expectancy in PLHIV, estimating the risk of developing a cardiovascular (CVD) event is of high importance for the assessment and management of CVD modifiable risk factors.
- Well-established CVD risk scores have been developed using traditional statistical methods. However, their performance might be poor under more complex associations in PLHIV.
- Simulation study to assess the best performing method $\Rightarrow$ develop a robust recommendation for CVD risk prediction algorithms to be developed in the future.

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
  - different associations between the 7 predictors and the outcome
    - linear (L)
    - linear and interactions (LI)
    - non-linear (NL)
    - non-linear and interactions (NLI)

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
    - ◇ different associations between the 7 predictors and the outcome
        - ∗ linear (L)
        - ∗ linear and interactions (LI)
        - ∗ non-linear (NL)
        - ∗ non-linear and interactions (NLI)
    - ◇ different training sample sizes (500, 1000, 5000)

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
  - different associations between the 7 predictors and the outcome
    - linear (L)
    - linear and interactions (LI)
    - non-linear (NL)
    - non-linear and interactions (NLI)
  - different training sample sizes (500, 1000, 5000)
  - censoring rates (50%, 75%, 93%) and

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
  - different associations between the 7 predictors and the outcome
    - linear (L)
    - linear and interactions (LI)
    - non-linear (NL)
    - non-linear and interactions (NLI)
  - different training sample sizes (500, 1000, 5000)
  - censoring rates (50%, 75%, 93%) and
  - hazard types (increasing, decreasing, constant).

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
  - different associations between the 7 predictors and the outcome
    - linear (L)
    - linear and interactions (LI)
    - non-linear (NL)
    - non-linear and interactions (NLI)
  - different training sample sizes (500, 1000, 5000)
  - censoring rates (50%, 75%, 93%) and
  - hazard types (increasing, decreasing, constant).
- Scenarios with stronger non-linear or interaction effects (NL-unrealistic, LI-unrealistic).

Simulation scenarios

- The majority of the simulation scenarios were generated under a realistic framework, roughly based on AMACS data.
- In total, 126 simulation scenarios were set up assuming:
  - different associations between the 7 predictors and the outcome
    - linear (L)
    - linear and interactions (LI)
    - non-linear (NL)
    - non-linear and interactions (NLI)
  - different training sample sizes (500, 1000, 5000)
  - censoring rates (50%, 75%, 93%) and
  - hazard types (increasing, decreasing, constant).
- Scenarios with stronger non-linear or interaction effects (NL-unrealistic, LI-unrealistic).
- Overall, 500 replications were performed for all the scenarios.

Simulation strategy

1. Simulate the covariates (e.g., continuous, binary) from normal or truncated normal and Bernoulli distribution using parameters obtained from the AMACS data.

Simulation strategy

1. Simulate the covariates (e.g., continuous, binary) from normal or truncated normal and Bernoulli distribution using parameters obtained from the AMACS data.
2. Simulate the survival and censoring time from a Weibull distribution with different shape and scale parameters.

Simulation strategy

1. Simulate the covariates (e.g., continuous, binary) from normal or truncated normal and Bernoulli distribution using parameters obtained from the AMACS data.
2. Simulate the survival and censoring time from a Weibull distribution with different shape and scale parameters.
3. The final observed time was the minimum value of the survival time and the censoring time.

Simulation strategy

1. Simulate the covariates (e.g., continuous, binary) from normal or truncated normal and Bernoulli distribution using parameters obtained from the AMACS data.
2. Simulate the survival and censoring time from a Weibull distribution with different shape and scale parameters.
3. The final observed time was the minimum value of the survival time and the censoring time.
4. For a more realistic context, we considered administrative censoring at 10 years of follow-up.

Cox proportional hazards model

Cox-PH (Cox, 1972) is expressed by the hazard function, which is the risk of an event occurring at time t as follows:

$$h(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p},$$

where t represents the survival time, $h(t)$ the hazard function, $x_1, x_2,$ ..., x_p the values of the p covariates, $\beta_1, \beta_2, \ldots, \beta_p$ the associated coefficients and $h_0(t)$ the baseline hazard function, which is unspecified.

Random Forest

1. Draw B bootstrap samples from the training data.

Random Forest

1. Draw B bootstrap samples from the training data.
2. Grow a decision tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split into two daughter nodes using the candidate variable that best separates the data based on a split criterion, such as Gini impurity.

Random Forest

1. Draw B bootstrap samples from the training data.
2. Grow a decision tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split into two daughter nodes using the candidate variable that best separates the data based on a split criterion, such as Gini impurity.
3. Grow the tree to full size until a stopping criterion is met (i.e., all terminal nodes contain no less than a given number of observations).

Random Forest

1. Draw B bootstrap samples from the training data.
2. Grow a decision tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split into two daughter nodes using the candidate variable that best separates the data based on a split criterion, such as Gini impurity.
3. Grow the tree to full size until a stopping criterion is met (i.e., all terminal nodes contain no less than a given number of observations).
4. Combine the predictions of all trees through either averaging for probability estimation or majority voting for classification.

Random Survival Forest

- RSF (Ishwaran et al., 2008) is an extension of Breiman's RF (Breiman, 2001) that can handle right-censored survival data. It is an ensemble of a large number of survival trees which unlike RF, uses a split rule so as to maximise the survival difference between daughter nodes.

Random Survival Forest

- RSF (Ishwaran et al., 2008) is an extension of Breiman's RF (Breiman, 2001) that can handle right-censored survival data. It is an ensemble of a large number of survival trees which unlike RF, uses a split rule so as to maximise the survival difference between daughter nodes.
- In this study, we applied the following split criteria:
  - the traditional log-rank test (RSF-LR) (Ishwaran et al., 2008)
  - the log-rank score (RSF-LRS), which splits the nodes using a standardized log-rank statistic (Hothorn and Lausen, 2003)
  - the maximally selected rank statistics (RSF-MSR), which performs a binary split that adjusts for the multiple testing of multiple possible split points (Wright et al., 2017).

Performance evaluation

- Training samples of three different sizes ($N = 500$, 1000, 5000) were used to train the models and models' performance was evaluated on equal-sized test data generated independently from the training data.

Performance evaluation

- Training samples of three different sizes ($N = 500$, 1000, 5000) were used to train the models and models' performance was evaluated on equal-sized test data generated independently from the training data.
- To reduce the variability of the models' performance estimates, models' performance was evaluated on test data of size $N = 10,000$ generated independently from the training data.

Performance evaluation

- Training samples of three different sizes ($N = 500$, 1000, 5000) were used to train the models and models' performance was evaluated on equal-sized test data generated independently from the training data.
- To reduce the variability of the models' performance estimates, models' performance was evaluated on test data of size $N = 10,000$ generated independently from the training data.
- We compared models' performance with two popular criteria:
  - ◇ time-dependent area under the receiver operating characteristic (ROC) curve (AUC)
  - ◇ Integrated Brier score (IBS)

Time-dependent AUC

- The time-dependent AUC (Uno et al., 2007) at a given time point t denotes the probability that a randomly selected patient who experiences the event at time t has higher predicted risk than a randomly selected patient without the event until that time point.

Integrated Brier Score

- The Brier score at a given time point t (Graf et al., 1999) is defined as follows:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\{[0 - \hat{S}(t|\mathbf{X}_i)]^2 \frac{I(\tau_i \leqslant t, \delta_i = 1)}{\hat{G}(\tau_i)} + [1 - \hat{S}(t|\mathbf{X}_i)]^2 \frac{I(\tau_i > t)}{\hat{G}(t)}\}$$

where N is the test sample size, $I(\cdot)$ the indicator function and $\hat{G}(t)$ the Kaplan-Meier estimate of the censoring distribution.

Integrated Brier Score

- The Brier score at a given time point t (Graf et al., 1999) is defined as follows:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\{[0 - \hat{S}(t|\mathbf{X}_i)]^2 \frac{I(\tau_i \leqslant t, \delta_i = 1)}{\hat{G}(\tau_i)} + [1 - \hat{S}(t|\mathbf{X}_i)]^2 \frac{I(\tau_i > t)}{\hat{G}(t)}\}$$

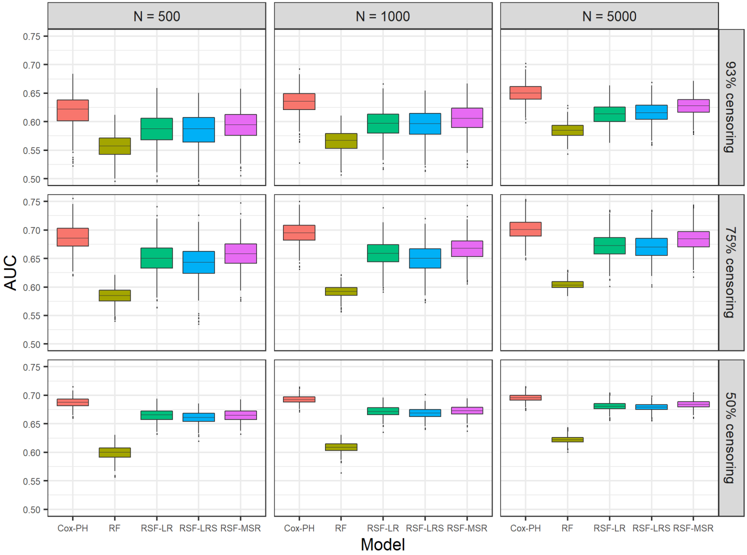
where N is the test sample size, $I(\cdot)$ the indicator function and $\hat{G}(t)$ the Kaplan-Meier estimate of the censoring distribution.

- The IBS over a certain time period $[0, t^*]$ is given by:

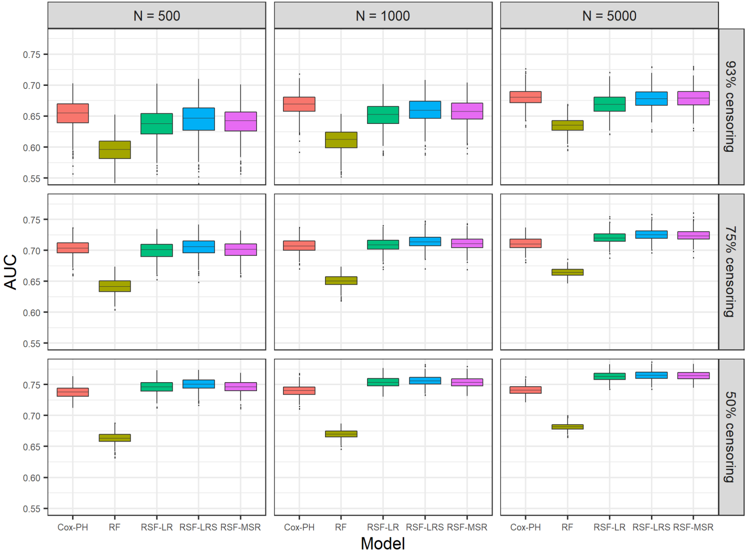
$$IBS(t^*) = \frac{1}{t^*}\int_0^{t^*} BS(t)\, dt,$$

where $BS(t)$ is the Brier score at time t. Survival models with a small IBS value are preferable.

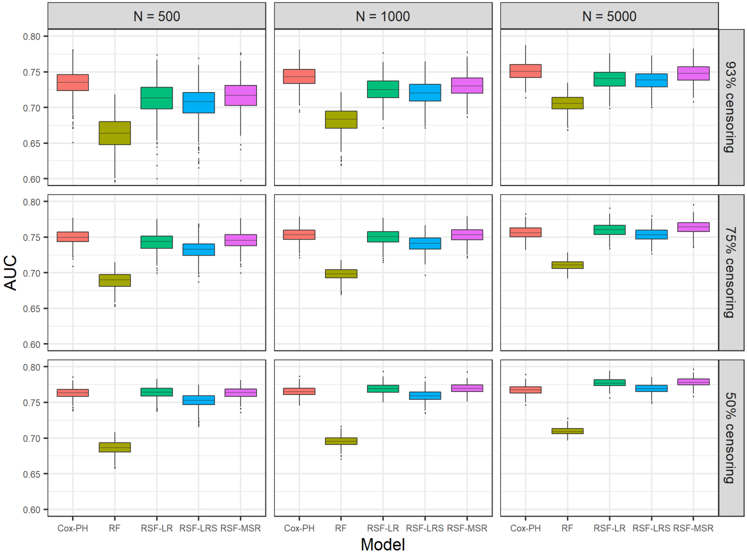
Simulation results: **Linear** scenarios



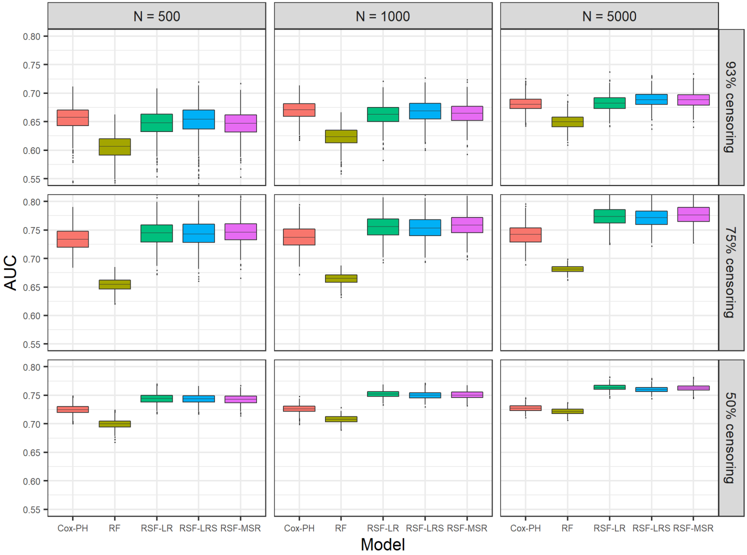
Simulation results: **Non-linear** scenarios



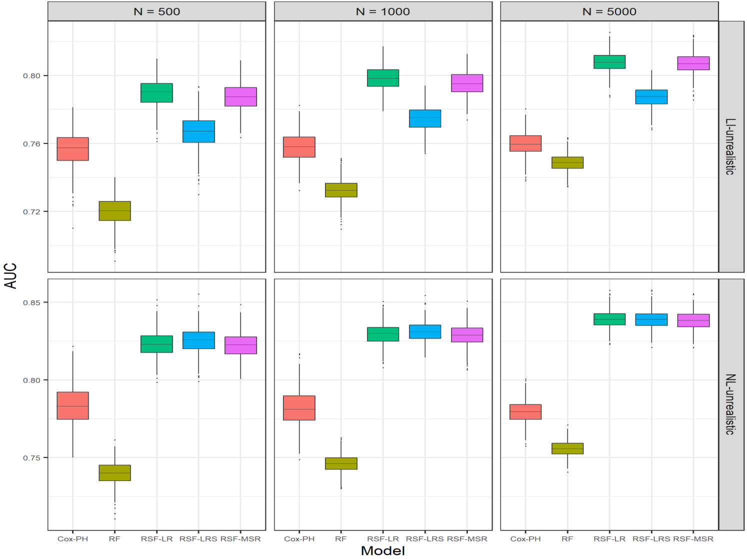
Simulation results: **Linear + Interactions** scenarios



Simulation results: **Non-linear + Interactions** scenarios



Simulation results: **unrealistic** scenarios - 50% censoring



Real data application (1)

- As an illustrative example, all models were applied and evaluated on the AMACS data to estimate the 7-year risk of developing a first CVD event in PLHIV.
- Baseline: The enrolment date in the study.

Real data application (1)

- As an illustrative example, all models were applied and evaluated on the AMACS data to estimate the 7-year risk of developing a first CVD event in PLHIV.
- Baseline: The enrolment date in the study.
- $N = 5,054$ participants
- 7.8% developed a first CVD event during 7 years of follow-up.

Real data application (1)

- As an illustrative example, all models were applied and evaluated on the AMACS data to estimate the 7-year risk of developing a first CVD event in PLHIV.
- Baseline: The enrolment date in the study.
- $N = 5,054$ participants
- 7.8% developed a first CVD event during 7 years of follow-up.
- To be consistent with our simulation study, we randomly split the data into a training set (70% of the participants) used for building the models and a test set (the remaining 30%) used for model assessment.

Real data application (2)

- In agreement with the results of all simulation scenarios, RF had the lowest performance in terms of both AUC and IBS.
- Similarly with the scenarios corresponding to 93% censoring rate, Cox-PH had the highest AUC compared to the RSF algorithms among which RSF-LRS performed best, while the IBS values for all the models were close.

	AUC	IBS
Cox-PH	0.758	0.048
RSF-LR	0.723	0.048
RSF-LRS	0.743	0.048
RSF-MSR	0.735	0.048
RF	0.703	0.071

Conclusions

- RSF is a promising alternative to Cox-PH as data complexity increases (i.e., non-linear relationships, high order interactions), but requires a higher number of events for training.

Conclusions

- RSF is a promising alternative to Cox-PH as data complexity increases (i.e., non-linear relationships, high order interactions), but requires a higher number of events for training.
- In all scenarios investigated (apart from the unrealistic ones) with a small number of events, the performance of Cox-PH was at least non-inferior to RSF despite deviations from linearity or multiplicativity.

Conclusions

- RSF is a promising alternative to Cox-PH as data complexity increases (i.e., non-linear relationships, high order interactions), but requires a higher number of events for training.
- In all scenarios investigated (apart from the unrealistic ones) with a small number of events, the performance of Cox-PH was at least non-inferior to RSF despite deviations from linearity or multiplicativity.
- Given the poor performance of RF in comparison with the remaining models, we strongly suggest to use models that account for survival time in time-to-event analysis.

Conclusions

- RSF is a promising alternative to Cox-PH as data complexity increases (i.e., non-linear relationships, high order interactions), but requires a higher number of events for training.
- In all scenarios investigated (apart from the unrealistic ones) with a small number of events, the performance of Cox-PH was at least non-inferior to RSF despite deviations from linearity or multiplicativity.
- Given the poor performance of RF in comparison with the remaining models, we strongly suggest to use models that account for survival time in time-to-event analysis.
- All three variants of RSF algorithm performed similarly in all scenarios investigated.

References

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999), 'Assessment and comparison of prognostic classification schemes for survival data', *Statistics in medicine* **18**(17-18), 2529–2545.

Hothorn, T. and Lausen, B. (2003), 'On the exact distribution of maximally selected rank statistics', *Computational Statistics & Data Analysis* **43**(2), 121–137.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. et al. (2008), 'Random survival forests', *Annals of Applied Statistics* **2**(3), 841–860.

Uno, H., Cai, T., Tian, L. and Wei, L.-J. (2007), 'Evaluating prediction rules for t-year survivors with censored regression models', *Journal of the American Statistical Association* **102**(478), 527–537.

Wright, M. N., Dankowski, T. and Ziegler, A. (2017), 'Unbiased split variable selection for random survival forests using maximally selected rank statistics', *Statistics in medicine* **36**(8), 1272–1284.

Acknowledgements

The present work was co-funded by the European Union and Greek national funds through the Operational Program "Competitiveness, Entrepreneurship and Innovation", under the call "RESEARCH-CREATE-INNOVATE" (project code: T1EΔK-03812).

Thank You!