h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Availability and Visualization of Predictions and their Uncertainty from Kidney Transplant Survival Models

Author: Patrick Eidemüller | Referent: Prof. Dr. Antje Jahn | Coreferent Prof. Dr. Gunter Grieser

Darmstadt University of Applied Science, Faculties of Mathematics and Natural Science & Computer Science

Motivation

Kidney transplantation is crucial for patients suffering from end-stage renal disease. Due to limited donor organ availability, accurately predicting transplant outcomes is essential for optimal organ utilization and patient care. Machine learning (ML) survival models have demonstrated great potential in providing individualized prognostic insights by capturing complex relationships between clinical variables. However, current ML models primarily focus on predictive performance without adequately addressing predictive uncertainty. This is particularly problematic in clinical decision-making, where reliance on uncertain point estimates could significantly affect patient outcomes. Additionally, despite advances in survival modeling techniques, their integration into clinical workflows remains limited, highlighting the urgent need for user-centered tools that translate complex model predictions and their associated uncertainties into practical, interpretable clinical decision-support tools.

Objectives of the Study

This study addresses key challenges in applying existing survival models to new registry data and communicating predictive uncertainty. The objectives are:

1. **Evaluation of Prognostic Models for TxReg Data**
   - Identify models (A) that allow individualized predictions.
   - Evaluate availability of variables, preprocessing, model details, and reproducibility.
   - Investigate online calculators (B) and their handling of uncertainty.
2. **Benchmarking a Selected A1 Model**
   - Select one suitable A1 model and apply it to TxReg.
   - Benchmark its predictive performance.
   - Quantify and visualize the model's uncertainty using methods found in the literature.
3. **Development of a Clinical Dashboard Prototype**
   - Design a user-friendly prototype dashboard for survival predictions and their uncertainties.

Methodology

This study systematically evaluates existing prognostic models for kidney transplant survival, applying them to data from the German Transplant Registry (TxReg). The methodological approach consists of three primary steps:

1. **Systematic Literature Review**
   - A structured screening of 50 publications was conducted to identify suitable ML-based survival models.
   - Key methodological details such as variables, preprocessing, and model design were extracted.
   - Models were categorized into type A (reproducible), A1 (reproducible and all variables available) and B (available calculators).
2. **Model Implementation and Evaluation**
   - Suitable A1 models were selected from the literature and applied to the TxReg dataset.
   - Preprocessing included covariate alignment, data preparation, and implementation of the model architecture as described in the original publications.
   - Model performance on the TxReg data was evaluated using the same metrics reported in the respective studies.
3. **Uncertainty Quantification and Visualization**
   - Uncertainty estimation techniques identified through the literature reviewâspecifically Monte Carlo Dropout and Bootstrapâwere applied to quantify predictive uncertainties.
   - A prototype of an interactive dashboard was developed to visualize individual survival predictions and their associated uncertainties, guided by best practices and visualization strategies found in the reviewed studies.

Results

Literature Review Findings: A systematic review of 50 publications on machine learning-based survival models in the context of organ transplantation revealed the following key findings:

Reproducibility: Only 10 of the 50 studies provided sufficient methodological detail covering variables, preprocessing steps, model architecture, and hyperparameters to be classified as reproducible models. This level of transparency is essential for replicability and model transfer.

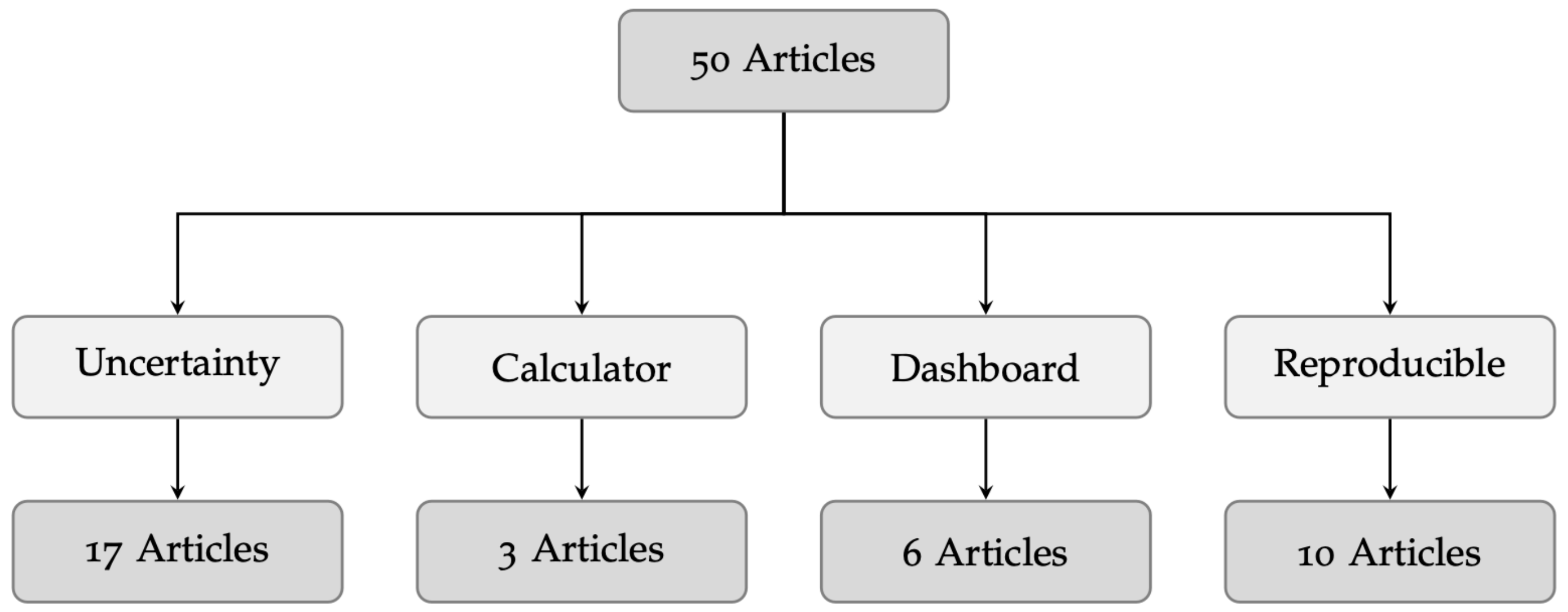
Transferability to TxReg: Among the 10 reproducible A models, only four showed a high degree of covariate overlap with the German Transplant Registry (TxReg). One model, *DeepHit* by Paquette et al., stood out for its methodological clarity and vriable alignment therefore it was selected for implementation with TxReg.

Uncertainty Quantification: 17 studies mentioned uncertainty-related concepts. However, methods such as bootstrapping or Monte Carlo dropout were rarely described in detail, and no study focused primarily on uncertainty quantification techniques.

Results

Calculator and Dashboard Availability: Only 3 studies provided web-based prognostic calculators, and 6 included dashboards or other visual tools. None of these integrated uncertainty into their prediction outputs representing a significant shortcoming for clinical use.

These findings highlight that while ML-based survival models are widely explored, crucial gaps remain in reproducibility, uncertainty modeling, and practical usability in clinical contexts.


50 Articles

Uncertainty
17 Articles

Calculator
3 Articles

Dashboard
6 Articles

Reproducible
10 Articles

Figure 1:Filtering Process of the Literature Review

Model Performance on TxReg Data: Among the 10 reproducible A models, one potential A1 model DeepHit by Paquette et al. Machine Learning Support for Decision-Making in Kidney Transplantation was selected based on variable alignment with TxReg (3 variables missing). It was applied to TxReg data alongside a Cox Proportional Hazards model (CoxPH).

- **C-Index:** The DeepHit model achieved a C-index of 0.6433 on TxReg, which is slightly lower than the original result (0.661), yet indicates acceptable discrimination and supports feasibility of model transfer.
- **Integrated Brier Score (IBS):** Although the IBS on TxReg (0.1811) was higher than in the original study (0.1528), the model still demonstrated acceptable calibration and could be feasibly applied.

Table 1:Comparison of Evaluation Metrics: Original Study vs. TxReg

Metric	Original Study (35700006)		TxReg Dataset	
	DeepHit	CoxPH	DeepHit	CoxPH
Harrell's C-Index	**0.661**	0.646	0.6433	**0.6757**
IBS	**0.1528**	**0.1543**	0.1811	0.1791

Uncertainty Quantification and Visualization

Uncertainty quantification was performed using two established methods:

- **Monte Carlo Dropout:** Produced relatively narrow confidence intervals, reflecting low model uncertainty estimates.
- **Bootstrap:** Generated wider intervals, capturing broader variability in survival predictions.

These complementary methods enabled a robust assessment of predictive uncertainty at the individual patient level. Survival functions with confidence intervals were visualized for comparison.

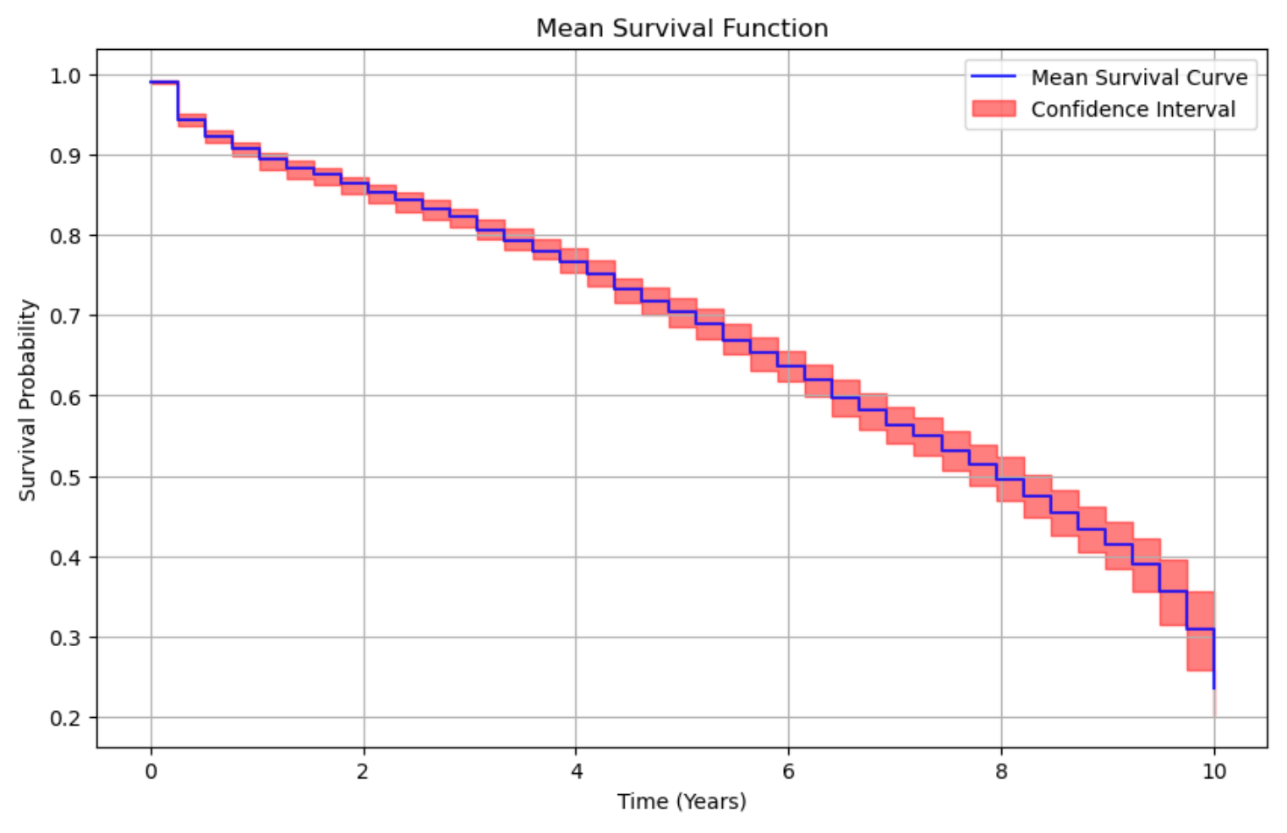


Figure 2:Bootstrap Survival Function

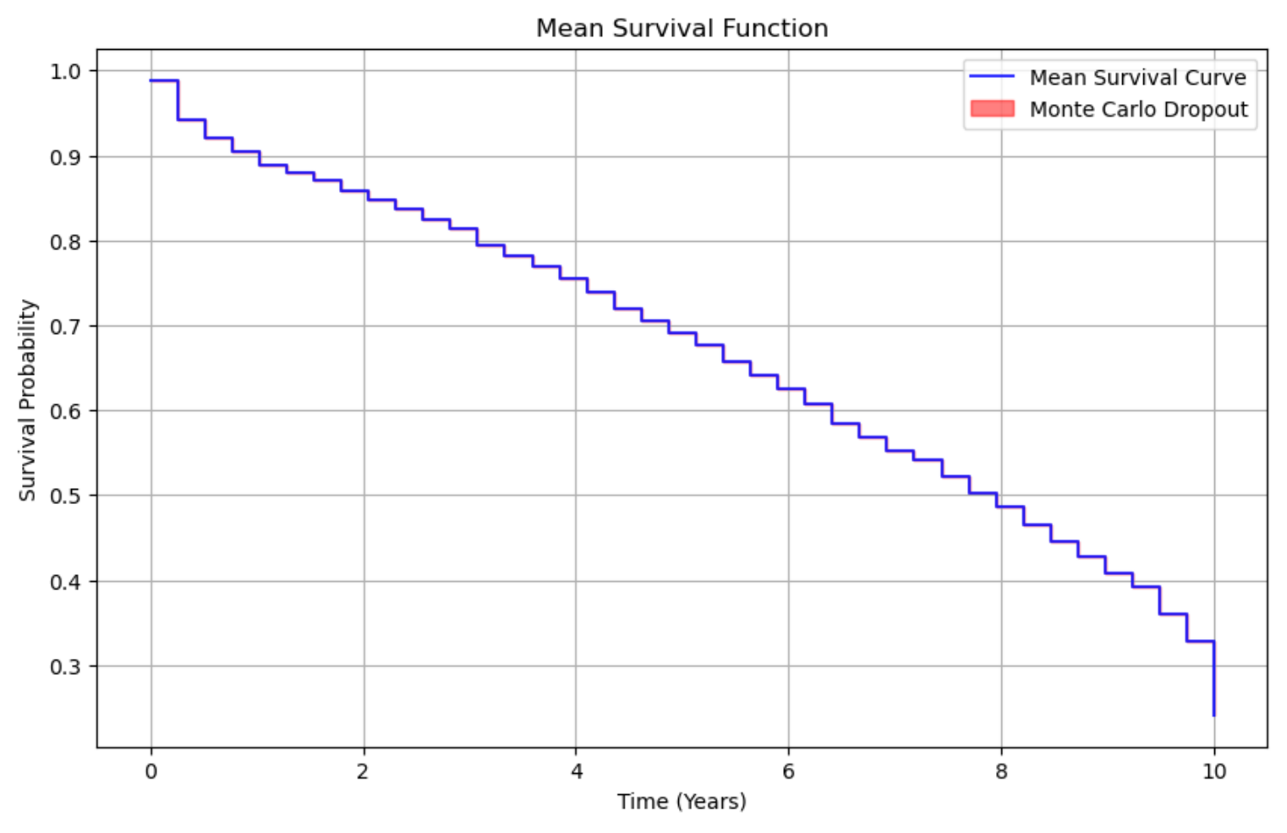
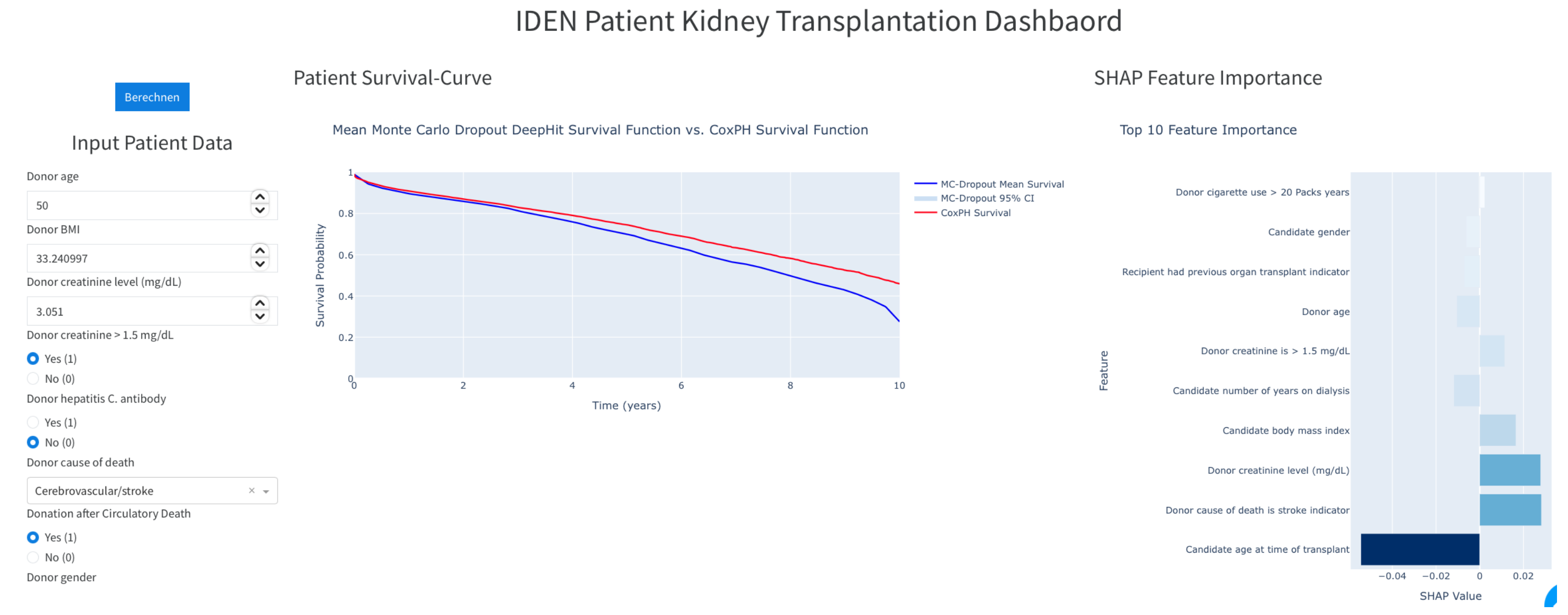


Figure 3:Monte Carlo Survival Function

Result

To support clinical interpretation, a **prototype dashboard** was implemented, visualizing individualized survival curves along with uncertainty intervals. Its design was inspired by best practices from the literature and aims to enhance transparency while reducing overreliance on point estimates.


IDEN Patient Kidney Transplantation Dashbaord

Input Patient Data

Donor age: 50
Donor BMI: 33.240997
Donor creatinine level (mg/dL): 3.051
Donor creatinine > 1.5 mg/dL: ● Yes (1) ○ No (0)
Donor hepatitis C. antibody: ○ Yes (1) ● No (0)
Donor cause of death: Cerebrovascular/stroke
Donation after Circulatory Death: ● Yes (1) ○ No (0)
Donor gender

Figure 4:Interactive Dashboard for Survival Prediction and Uncertainty Visualization

Discussion & Limitations

Discussion: This study demonstrates that machine learningâbased survival models, such as DeepHit, can be transferred to independent datasets like the German Transplant Registry (TxReg). While the C-index was comparable to the original study, increased prediction error (IBS) indicates reduced calibrationâlikely due to differences in variable availability and data distributions. Uncertainty quantification using Monte Carlo dropout and bootstrapping enabled more transparent interpretation of patient-specific predictions and highlighted variability in confidence estimation. The developed dashboard integrates these findings into a visual interface that improves interpretability and supports clinical decision-making.

Limitations: Several challenges were encountered:

- **Model Reproducibility:** Many reviewed studies lacked standardized reporting of model architecture, preprocessing, and hyperparameters, limiting reproducibility.
- **Transferability:** Even reproducible models required adjustments due to unmatched or missing variables in TxReg, influencing performance.
- **Uncertainty Methods:** Only two techniques (dropout, bootstrapping) were explored. Broader comparison with other methods like Bayesian networks or conformal prediction was beyond scope.
- **Dashboard Evaluation:** The prototype dashboard was developed without clinician feedback, and further validation is necessary to ensure usability in practice.

Conclusion & Outlook

This study demonstrates that while machine learning based survival models show promising results for individualized predictions in kidney transplantation, significant challenges persist in model reproducibility, transferability, and uncertainty communication. Out of 50 reviewed publications, only a few met the standards for reproducibility, and just one model DeepHit by Paquette et al. Ewas transferable to the TxReg dataset after adaptation. Despite slight performance reductions, it remained functional, highlighting both the potential and limitations of model reuse across contexts. The integration of Monte Carlo Dropout and Bootstrap for uncertainty quantification provided valuable insights into prediction confidence and supported the development of an interactive dashboard for clinicians. However, the dashboard remains a prototype and requires further validation through user feedback and clinical testing.

Future work should focus on improving model documentation standards to enhance transferability, integrating advanced uncertainty quantification techniques directly into model training, and co-designing dashboards with medical professionals to improve clinical usability and trust.

References

[1] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993. ISBN: 9780412042317.

[2] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. 2016, pp. 1050–1059.

[3] Changhee Lee et al. "Deephit: A deep learning approach to survival analysis with competing risks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.