European Medicines Agency (EMA) Scientific Opinion-Qualification

Critical Path Institute

Transplant Therapeutics Consortium

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel immunosuppressive therapy applications via conditional marketing authorisation

Briefing Dossier

Submitted 3 December 2021

Re-submitted 16 February 2022

Table 1. C-Path

Name
Varun Aggarwal, PhD
Hailey Davenport
William E. Fitzsimmons, PharmD, MS
Eric Frey
Amanda Klein, PharmD
Luke Kosinski, PhD
Rhoda Muse, PhD
Inish O’Doherty, PhD
Jagdeep Podichetty, PhD
Klaus Romero, PhD
Nicole Spear, MS

Table 2. Other TTC participants

Name
Paris Transplant Group (PTG)
TTC Coordinating Committee

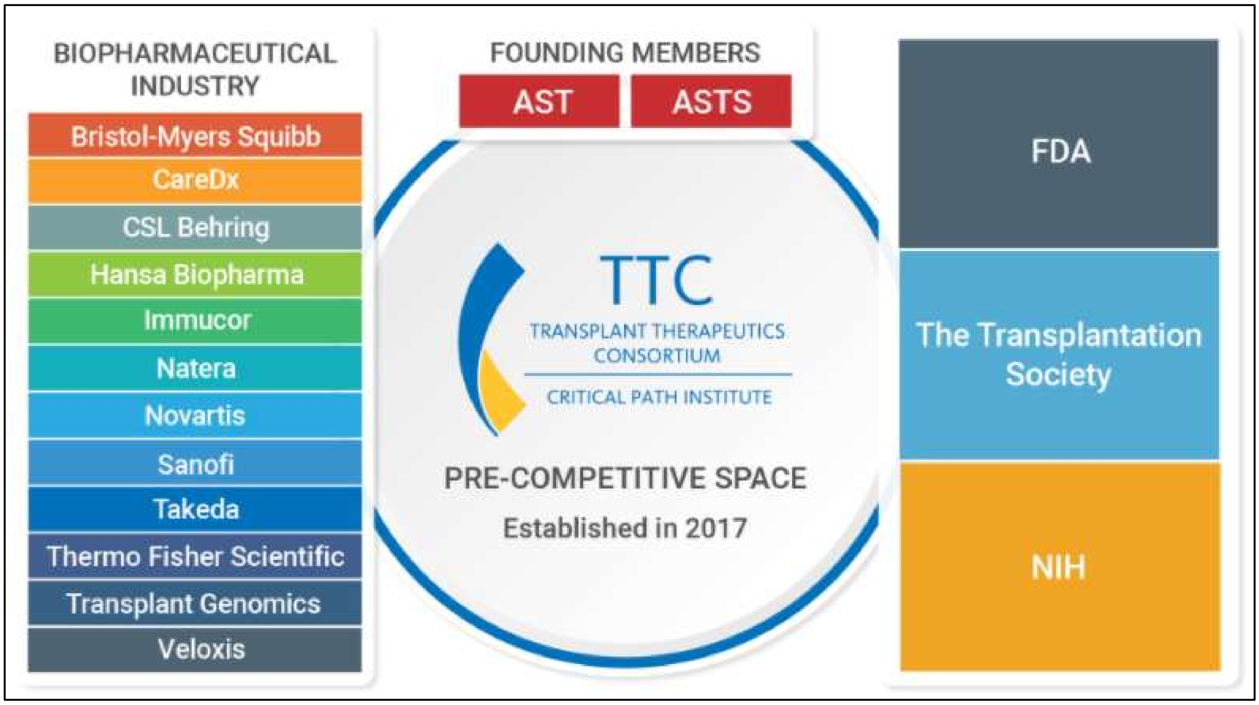


Figure 1. Transplant Therapeutics Consortium

Table 3. List of abbreviations

aAMR	Acute antibody-mediated rejection, formerly also known as acute/active AMR
Abbreviated iBox Scoring System	Abbreviated (without biopsy) iBox Scoring System (Composite Biomarker Panel)
ACE score	All-cause endpoint score
Ah	Banff lesion score, arteriolar hyalinosis
AMR	Antibody-mediated rejection
AR	Acute rejection
AST	American Society of Transplantation
ASTS	American Society of Transplant Surgeons
aTCMR	acute T cell-mediated rejection
BELA	Belatacept
BENEFIT RCT	Belatacept Evaluation of Nephroprotection and Efficacy as First-line Immunosuppression Trial
BENEFIT-EXT RCT	Belatacept Evaluation of Nephroprotection and Efficacy as First-line Immunosuppression Trial-EXTended Criteria Donor
BMJ	British Medical Journal
BMS	Bristol-Meyers Squibb
BORTEJECT RCT	A Randomized Trial of Bortezomib in Late Antibody Mediated Kidney Transplant Rejection
BPAR	Biopsy-proven acute rejection

c-statistic	Harrell’s c-statistic (Harrell, Lee, and Mark 1996)
CAN	Chronic allograft nephropathy
C.E.	Conformité Européenne
CERTITEM RCT	Fibrosis progression according to epithelial-mesenchymal transition profile: a randomized trial of everolimus versus CsA
CDRH	Center for Devices and Radiological Health
cg	Banff lesion score, presence/extent of glomerular base membrane (GBM) double contours; transplant glomerulopathy
ci	Banff lesion score, interstitial fibrosis
CI	Confidence interval
CIF	Cumulative incidence function
CIT	Cold ischaemia time
CMA	Conditional marketing authorisation
CNI	Calcineurin inhibitor
COU	Context-of-use
C-Path	Critical Path Institute
CsA	Cyclosporine
ct	Banff lesion score, tubular atrophy
cv	Banff lesion score, vascular fibrous intimal thickening
DDT	Drug development tool
DGF	Delayed graft function
DSA	Donor-specific antibody
ECD	Expanded criteria donor
eGFR	Estimated glomerular filtration rate
EMA	European Medicines Agency
ERA-EDTA	European Renal Association-European Dialysis and Transplant Association
ESOT	European Society of Transplantation
ESRD	End-stage renal disease
FDA	U.S. Food and Drug Administration
Full iBox Scoring System	Full (with biopsy) iBox Scoring System (Composite Biomarker Panel)
g	Banff lesion score, Glomerulitis
g/g	Gram per gram
HLA	Human leukocyte antigen
HR	Hazard ratio
i	Banff lesion score, interstitial inflammation
iBox	Integrative box
iBox Scoring System (Composite Biomarker Panel)	iBox Scoring System, full and abbreviated models
IFTA	Banff lesion score, interstitial fibrosis/tubular atrophy
IFU	Instructions for Use
IL-2Ra	Interleukin-2 receptor antagonist
Inserm	Institut national de la santé et de la recherche médicale

IPTW	Inverse probability of treatment weight
IQR	Interquartile range
IST	Immunosuppressive therapy
IVIG	Intravenous immunoglobulin
KM	Kaplan-Meier
MAR	Missing at random
mCAR	Missing completely at random
MDRD	Modification of Diet in Renal Disease
MOA	Mechanism of action
MFI	Mean fluorescence intensity
mGFR	Measured glomerular filtration rate
mm	Banff lesion score, mesangial matrix expansion
mTORi	Mammalian target of rapamycin signal inhibitor
OPTN	Organ Procurement and Transplant Network
PH	Proportional hazard
PNF	Primary nonfunction of the graft
ptc	Banff lesion score peritubular capillaritis
PTG	Paris Transplant Group
PVAN	Polyomavirus-associated nephropathy
rATG	rabbit antithymocyte globulin
RCT	Randomized controlled trial
RITUX ERAH RCT	One-year results of the effects of rituximab on acute antibody-mediated rejection in renal transplantation: RITUX ERAH, a multicenter double-blind randomized placebo-controlled trial
ROC	Receiver Operating Characteristic
RR	Relative risk
SAB	Single-antigen bead
SAP	Statistical analysis plan
SAWP	Scientific Advice Working Party
SCr	Serum creatinine
SD	Standard deviation
SE	Standard error
SOC	Standard of care
SRTR	Scientific Registry of Transplant Recipients
STAR	Sensitization in Transplantation: Assessment of Risk
STE	Surrogate threshold effect
t	Banff lesion score, tubulitis
TAC	Tacrolimus
TCMR	T cell-mediated rejection
ti	Banff lesion score, total inflammation
TLS	Trial-level surrogacy
TTC	Transplant Therapeutics Consortium
UACR	Urine albumin-to-creatinine ratio
UPCR	Urine protein-to-creatinine ratio
USRDS	United States Renal Data System

v	Banff lesion score, intimal arteritis

1 TABLE OF CONTENTS

1.1 List of Tables

1.2 List of Figures

1.3 List of Equations

2 EXECUTIVE SUMMARY

2.1 The objective(s) of request

The objective of this Briefing Dossier is for the Critical Path Institute's (C-Path) Transplant Therapeutics Consortium (TTC) to achieve a Qualification Opinion for a new drug development tool (DDT) for kidney transplantation through the EMA's qualification of novel methodologies for medicine drug development. This Briefing Dossier contains the proposed context-of-use (COU) statement, data source description, modeling analysis methods, and results that provide a quantitative basis to support the use of the iBox Scoring System (Composite Biomarker Panel), known as iBox Scoring System henceforth, as a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials evaluating the safety and efficacy of novel immunosuppressive therapies (ISTs). Two iBox Scoring System models have been developed and are included in this qualification submission: a full iBox Scoring System (with biopsy) and an abbreviated iBox Scoring System (without biopsy) known henceforth as the full iBox Scoring System, or the abbreviated iBox Scoring System, respectively. Additionally, a scoring system for predicting a combined endpoint including allograft failure and patient death as events), the ACE (all-cause endpoint) score, has been derived and tested in the external validation datasets (6.9 All-cause endpoint score for predicting deaths and graft losses).

The iBox Scoring System has been developed by estimating individual weights for each of the proposed components (i.e., estimated glomerular filtration rate [eGFR] calculated by the 4-variable Modification of Diet in Renal Disease (MDRD)-186 Study equation, proteinuria, kidney allograft biopsy histopathology, presence of donor-specific antibodies [DSA], and time of post-transplant iBox Scoring System risk evaluation. For the purpose of this submission, the time of post-transplant risk evaluation was fixed at one-year post-transplant. The ACE score incorporates all of the variables in the abbreviated iBox Scoring System.

2.2 The need and impact of proposed clinical novel methodology(ies)

The two major transplantation societies in the United States, the American Society of Transplant Surgeons (ASTS) and the American Society of Transplantation (AST), recognized in 2014 the need for a pathway for the development of new ISTs for transplant recipients. (Stegall et al. 2016). The two societies, along with other members of the transplant community and C-Path, created the TTC. The goal of the TTC is the goal of this proposal—to develop a path forward to accelerate the medical product development process for transplantation, with a focus on novel ISTs that are likely to improve long-term renal allograft survival. Following the Loupy et al., 2019 publication introducing the iBox risk prediction tool, AST and ASTS signed a joint letter of support in March of 2020 encouraging the Institut national de la santé et de la recherche médicale (Inserm) to share patient-level data used to derive the iBox Scoring System as per Loupy et al., 2019 with the TTC. This letter of support was written to assist the regulatory endorsement of the iBox Scoring System as a surrogate endpoint in kidney transplant clinical trials. The joint letter of support can be found in Appendix (AST-ASTS TTC Joint Letter of Support).

The historically-accepted clinical trial endpoint for multinational clinical trials of novel ISTs in kidney transplantation is the composite endpoint of equally-weighted death, graft-loss, biopsy-proven acute rejection (BPAR) and lost to follow-up at one-year post-transplantation. There are several issues with the continued reliance on this endpoint with the current standard of care (SOC) ISTs. Firstly, the incidence is low in the first year post-transplant, limiting the ability to demonstrate the superiority of a new innovative therapy. Secondly, this endpoint was originally designed to quantify the incidence of BPAR without censoring. However, this approach results in the equal weighting of transplant recipients who die compared to those with BPAR or are lost to follow-up. Lastly, the largest unmet need in transplant is improvement in the long-term survival of the transplant recipient and graft and the associated surrogate endpoints that are predictive of survival. Current IST regimens have dramatically improved short-term outcomes, with one-year graft survival rates of approximately 91% after deceased donor transplant, according to the European Renal Association - European Dialysis and Transplant Association (ERA-EDTA) 2018 Annual Report (ERA-EDTA Registry Annual Report 2018). Despite these improved short-term outcomes, long-term graft survival remains suboptimal. The 5 - and 10-year graft survival rate after deceased donor kidney transplant is 77% and 56%, respectively (Gondos et al. 2013). Consequently, there is a significant unmet need for ISTs that can help improve long-term outcomes, but developing novel therapies is challenging. One aspect of this challenge is demonstrating improved long-term outcomes, which require trials of long duration (i.e., five years or more) and contain a large number of subjects. As a result, one-to-two-year non-inferiority studies are more likely to be initiated, despite not adequately addressing the challenges of improving long-term graft survival. A strategy of using surrogate endpoints in assessing long-term outcomes has been employed in other therapeutic areas, such as oncology, diabetes, nephrology, and many rare diseases, to overcome similar challenges. Surrogate endpoints enable sponsors to seek conditional marketing authorisation (CMA) for novel agents based on clinical trials of reasonable duration (i.e., one year) that predict long-term outcomes (i.e., five years or greater) while planning and conducting studies to demonstrate longer-term therapeutic effects.

The challenges associated with developing a robust surrogate endpoint capable of accurately predicting long-term outcomes (i.e., five-year risk of graft loss) using short-term data (i.e., one-year post-transplant) are multifaceted. Two of the most significant challenges include the need to develop a reliable surrogate measure that performs across a heterogeneous subject population and the ability of the surrogate measure to demonstrate efficacy across therapies with multiple mechanisms of action (MOA). In addition, subject-level data from various sources representing a broad spectrum of subject populations and treatment settings must be aligned and curated to generate the necessary evidence to support the surrogacy claims of such a measure.

In 2019, the Paris Transplant Group (French National Institute of Health), together with 29 key opinion leaders of the transplant community from 10 referral centers from Europe and the USA, published a seminal paper on the iBox Scoring System titled: *Prediction system for risk of allograft survival in subjects receiving kidney transplants: international derivation and validation study* (Alexandre Loupy et al. 2019). The PTG designed a prospective study to identify key prognostic parameters and follow long-term outcomes of kidney transplant recipients to develop a new risk prediction model of long-term kidney allograft failure outperforming previous scoring systems.

In this publication, the iBox Scoring System is a risk prediction tool utilizing multiple clinically relevant subject features of kidney function (eGFR and proteinuria), kidney allograft biopsy histopathology, and immunological status (presence of DSA) data cross-sectionally at any timepoint post-transplantation. The component measures of the iBox Scoring System are routinely used as important factors in routine monitoring of transplant recipients to guide therapeutic interventions and for prognostic purposes. The iBox Scoring System integrates these measures to generate individualized predictions of outcomes at three, five, and seven-years post-transplant. Data prospectively collected from 4,000 consecutive subjects across four health centers in France were used to develop the iBox Scoring System, with external validation performed in cohorts from transplant centers in the U.S. (n = 1,428), Europe (n =

2,129), a phase III IST minimization trial (n = 194), a phase III trial assessing treatment of active antibody-mediated rejection (aAMR) in subjects with pre-transplant DSA (n = 38), and a phase II trial evaluating treatment of antibody-mediated rejection (AMR) in subjects with post-transplant *de novo* DSA (n = 44). The TTC, in close collaboration with the PTG, is seeking to translate the work from Loupy et al., 2019 British Medical Journal (BMJ) publication into a regulatory endpoint in hopes of streamlining drug development by facilitating clinical trials of shorter duration (i.e., one year) that can predict death-censored allograft survival.

While the underlying physiological mechanisms leading to allograft survival are complex, recent studies have shown that certain key features present relatively early after transplantation (i.e., within the first year) can accurately predict which grafts are most likely to fail at later time points (i.e., by five years). A key learning from prior efforts in the field is no one clinical feature or pathophysiological measure has the predictive power to robustly estimate long-term allograft survival (Naesens et al. 2016); (Kaplan, Schold, and Meier-Kriesche 2003); (Yilmaz et al. 2003); (Lefaucheur et al. 2010). Recent efforts that have had access to large subject cohorts with rigorous and routine clinical assessments collected at baseline and longitudinally for five to seven years have demonstrated improved predictability of long-term outcomes by assessing composites of multiple clinical features. These composite scores have focused on recipient demographics, pre-transplant measures, measures of kidney function within the first-year post-transplant, and combinations of these measures at different time points (Kaboré et al. 2017); (Shabir et al. 2014); (Gonzales et al. 2016); (Alexandre Loupy et al. 2019);(Rampersad et al. 2021).

More recently-developed composite scores have sought to predict long-term graft loss by incorporating a cross-section of the relevant pathophysiological measures of allograft survival, including kidney function, through eGFR calculated using serum creatinine (SCr) and measures of protein excreted into the urine, kidney damage as determined by pathological assessment of graft biopsy, and immune response, measured via the presence of DSA. Other composite scores have incorporated pathophysiological measures and recipient demographics (Gonzales et al. 2016); (Bentall et al. 2019). Discussion of notable risk prediction models that have informed this submission can be found in modeling analysis methodologies 5.1 (Prior knowledge).

These risk prediction scores have focused on predicting long-term allograft survival at the subject-level to inform individual clinical decision-making. However, none of these tools have been subject to independent external validation. Consequently, none of these tools have been a candidate or endorsed for use as a surrogate endpoint capable of supporting medical product registration studies or as surrogate endpoints in the context of EMA’s CMA (Menon, Murphy, and Heeger 2017). On the contrary, the iBox Scoring System showed accuracy in predicting death-censored allograft failure, which was confirmed across transplant centers worldwide, different subpopulations and clinical scenarios, as well as in randomized clinical trials (RCTs), lending its exportability to a variety of clinical trial settings.

The proposed iBox Scoring System in this submission is intended to be a surrogate endpoint for efficacy in clinical trials evaluating the safety and efficacy of novel ISTs in kidney transplant recipients as a marker for the probability of long-term allograft survival. TTC aims to improve upon the limitations of the historically utilized clinical trial primary endpoint through the development and regulatory endorsement of the iBox Scoring System capable of predicting long-term kidney transplant outcomes using measures available at one-year post-transplantation.

This effort builds on previous work in the field that has identified clinically relevant measures capable of predicting long-term allograft failure by curating data from multiple clinical trials, real-world clinical transplant center datasets, and long-term registry data. The TTC has been working closely with the PTG and the global transplant community to curate and align subject-level data to support the use of the iBox Scoring System in drug development. A key difference between the iBox Scoring System in the Loupy et al., 2019 manuscript and the iBox Scoring System as a surrogate endpoint detailed in this submission, is the time point for risk evaluation. In this submission, the COU has been defined with the risk evaluation fixed at one year post kidney transplant. While the Loupy, et al., 2019 iBox Scoring System algorithm allows the risk to be estimated at any time point post-transplant. The COU in this submission prespecified the risk evaluation at one-year post-transplant to adapt the iBox Scoring System described in Loupy et al. into a clinical trial endpoint at a fixed landmark. In order to facilitate the use of the iBox Scoring System in a multinational clinical trial, two versions of the iBox Scoring System were assessed, one version including all components as described by Loupy et al., 2019 (Full iBox Scoring System) and one version excluding pathophysiological assessment of the kidney allograft biopsy (abbreviated iBox Scoring System). Also, to adapt the Loupy et al., 2019 iBox Scoring System to be used as a one-year clinical trial endpoint, analyses were performed imputing a one-year iBox score for subjects who died or lost a graft in the first-year post-transplant.

Based on existing literature and work by the PTG, the proposed components of the iBox Scoring System model include:

- eGFR calculated by the 4-variable MDRD-186 Study equation with SCr (referred to as ‘eGFR’);
- Measurement of protein excretion into the urine through calculation of the urine protein-to-creatinine ratio (referred to as ‘proteinuria’);
- Histopathological assessment of tissue obtained by renal allograft biopsy (referred to as ‘kidney allograft biopsy histopathology’);
- Presence of DSA;
- The time of post-transplant iBox Scoring System risk evaluation. For the purpose of this submission, the time of risk evaluation was fixed at one-year post-transplant.

The multivariable Cox PH model was used to adapt the full and abbreviated iBox Scoring System models for use at one-year post-transplant as a surrogate endpoint for the five-year risk of death-censored allograft survival. Thus, this Briefing Dossier will consist of a discussion of these proposed components.

2.3 Sources of data

To acquire the subject-level data necessary to develop a novel surrogate endpoint, the TTC led an extensive global data collaboration effort across the field of kidney transplantation. To date, the TTC has acquired eleven clinical trial datasets and twenty observational datasets from clinical transplant centers, representing data from over 20,000 kidney transplant recipients in the TTC Kidney Transplant Database. A list of acquired datasets can be found in the Appendix (Revised-Transplant Therapeutics Consortium’s Kidney Transplant Database).

Datasets from relevant clinical trials of ISTs, including those in the Loupy et al. 2019 publication, and real-world data from international clinical transplant centers were prioritized for acquisition. From these 31 datasets, five contained all necessary variables collected at one-year post-transplant (i.e., eGFR, proteinuria, kidney allograft biopsy histopathology, and DSA), long-term death and graft loss follow-up of at least five years, immunosuppressive regimen information (i.e., induction and maintenance IST) to test the performance of the surrogate with all three MOA, and the documentation required to support the description of the analytical considerations for each dataset.

Datasets missing the necessary variables at one-year post-transplant or a variable necessary to calculate the model variable (as in recipient age to calculate an eGFR value) were excluded. For example, in the data for the three Novartis studies (TRANSFORM, US-92, and ELEVATE), recipient age was missing due to Novartis' anonymization procedures for data sharing. This, in turn, prohibited the calculation of eGFR values for the subjects in these studies. Moreover, US-92 and ELEVATE were missing DSA and proteinuria data, and follow-up was limited to one and two years, respectively.

The five datasets described below were therefore used for this qualification submission. These five qualification datasets consist of one derivation dataset and four validation datasets, outlined below.

Qualification derivation dataset:

1. The qualification derivation dataset presented in this Briefing Dossier included specific adjustments to the original derivation dataset as described in Loupy et al., 2019 manuscript, (Alexandre Loupy et al. 2019), allowing the iBox Scoring System to be used as a one-year post-transplant surrogate endpoint in clinical trials. This data was received from the PTG in Paris, France, Europe consisting of the following four transplant centers:

    - Necker Hospital in Paris, France, Europe.
    - Saint-Louis Hospital in Paris, France, Europe.
    - Foch Hospital in Suresnes, France, Europe.
    - Toulouse Hospital in Toulouse, France, Europe.

Qualification validation datasets:

The qualification validation datasets presented in this Briefing Dossier contain datasets other than those used for external validation as described in Loupy et al., 2019 manuscript (Alexandre Loupy et al. 2019). The qualification validation datasets are from both transplant centers and RCTs as described below.

1. Mayo Clinic in Rochester, Minnesota, USA, North America.
2. Helsinki University Hospital in Helsinki, Finland, Europe.
3. A phase III study of belatacept-based immunosuppression regimens versus cyclosporine (CsA) in recipients of kidneys from living or standard criteria deceased donor kidneys (BENEFIT RCT) Vincenti et al., 2012.
4. A phase III study of belatacept versus CsA in recipients of allografts from extended criteria donors, those donated after cardiac death, and those with an estimated cold ischemia time (CIT) > 24 hours in duration (BENEFIT-EXT RCT) Medina-Pestana., 2012

The qualification derivation and validation datasets were aligned and curated to support the regulatory endorsement of the full and abbreviated iBox Scoring System models. These datasets were used to construct the statistical analysis plan (SAP) presented in this Briefing Dossier.

2.4 Characteristics of the proposed novel methodology

Proposed context-of-use statement

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA.

General area:

Surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications.

Target population for use of the biomarker:

Adult *de novo* kidney only transplant recipients from a living or deceased donor.

Stage of drug development for use:

All clinical efficacy evaluation stages of therapeutic interventions focused on the use of the long-term risk of allograft survival in kidney transplant recipients, including early signs of efficacy, proof-of-concept, dose-ranging, and registration studies (Phases II-IV).

Intended application:

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications via CMA. When evaluating five-year outcomes for clinical benefit and full marketing authorisation, it will be necessary to ensure that there is not a clinically meaningful decrease in transplant recipient survival with the new therapy in the clinical trial compared to the standard of care control arms.

2.5 Differences between proposed COU and the Loupy et al., 2019 publication

The original derivation dataset (Alexandre Loupy et al. 2019) was used in the derivation analysis of the full iBox Scoring System and the abbreviated iBox Scoring System. The qualification derivation dataset presented in this Briefing Dossier included specific adjustments to the originally derived formula allowing the iBox Scoring System risk evaluation at one-year post-transplantation for use in a clinical trial endpoint at a fixed landmark, further described in Methods 4.3.8 (Alignment of qualification datasets).

The qualification validation presented in this Briefing Dossier used datasets other than those used for external validation in Loupy et al., 2019 manuscript [(Alexandre Loupy et al. 2019), further described in Methods 4.3.1 (Introduction to data).

Table 4. compares and contrasts the iBox Scoring System described in Loupy et al., 2019 manuscript and the iBox Scoring System as a surrogate endpoint proposed in this Briefing Dossier for Qualification Opinion.

Table 4. iBox Scoring System as described in Loupy et al., 2019 versus iBox Scoring System proposed for Qualification Opinion

	Loupy et al., 2019	iBox Scoring System proposed for Qualification Opinion
Core components of model	1. eGFR$_{\text{MDRD}}$ 2. Proteinuria: log transformed UPCR 3. Kidney allograft biopsy histopathology 4. DSA: Semiquantitative mean fluorescence intensity (MFI) associated with DSA 5. Time of post-transplant risk evaluation: at any time from transplant	1. eGFR$_{\text{MDRD}}$ 2. Proteinuria: log transformed UPCR; imputation methodology included for datasets using other proteinuria measurements 3. Two iBox Scoring System models, one with and one without kidney allograft biopsy histopathology 4. DSA: Binary qualitative MFI associated with DSA 5. Time of post-transplant risk evaluation: one-year post-transplant
Application	Individual decision-making	Surrogate endpoint in kidney transplantation clinical trials
Derivation set	Loupy et al., 2019	Loupy et al., 2019

External validation sets	Hôpital Hôtel Dieu, Nantes, France; Hospices Civils, Lyon, France; University Hospitals, Leuven, Belgium; Johns Hopkins Medical Institute, Baltimore, MD; the Mayo Clinic, Rochester, MN; and the Virginia Commonwealth University School of Medicine, Richmond, VA	Mayo Clinic Rochester[ǂ];
		Helsinki University Hospital;
		BENEFIT RCT;
		BENEFIT-EXT RCT

Methodology	Semiparametric Cox PH model	Semiparametric Cox PH model; imputation for proteinuria and for subjects who die or lose their graft in the first year of transplant
Outcomes	Death-censored allograft survival	Death-censored allograft survival
Imputation used for sensitivity analysis in trial-level surrogacy (TLS) and for one-year endpoint definition	No	Yes
Assay documentation	Excluded	Included

[ǂ] Different dataset than in Loupy et al., 2019

2.6 Summary of the Qualification purpose, methods, and results

There is a need for new short-term endpoints in kidney transplant trials that allow demonstration of superiority of new therapies over the current SOC and translate into reductions in long-term graft loss. The availability of a surrogate endpoint is vital to stimulate innovation in immunosuppressive drug development that will serve transplant recipients by further improving short- and long-term outcomes.

Loupy et al., 2019 developed the iBox Scoring System as a risk prediction score for death-censored kidney allograft survival by estimating individual weights for each of the proposed components (i.e., eGFR, proteinuria, kidney allograft biopsy histopathology, the presence of DSA, and time of post-transplant risk evaluation). The TTC has adapted the innovative work by Loupy et al., 2019, to transform the original iBox Scoring System to a surrogate clinical trial endpoint measured at one-year post-transplant.

The following key analyses have been performed and are detailed in this submission:

- Original iBox Scoring System analyses of data by Loupy et al., 2019 have been reproduced for the full iBox Scoring System and abbreviated iBox Scoring System for the data from the PTG (derivation dataset n = 3,941 for full iBox Scoring System and n = 4,000 for abbreviated iBox Scoring System). Results 6.2 (Multivariate analysis).

- For application as an endpoint in a clinical trial at one-year, the derivation dataset from PTG was analyzed, restricting the analysis to those recipients with a full iBox Scoring System evaluation at one-year post-transplant and follow-up to five-years for graft loss (n = 1,174). The discrimination in this group was confirmed with a c-statistic = 0.85. Results 6.5.1 (Internal validation).

- Subsequently, external validation was performed in the four qualification datasets (i.e., two observational datasets from Helsinki University Hospital and Mayo Clinic Rochester and two RCTs from Bristol-Meyers Squibb (BMS), BENEFIT and BENEFIT-EXT). Results 6.5.2.1 (External validation on the qualification datasets).

- External validation was performed using discrimination (c-statistics) and calibration (observed versus predicted graft loss) methods. In all four of the qualification validation datasets using the full and abbreviated iBox Scoring System models at one year to predict five-year death-censored allograft survival, the c-statistics ranged from 0.70-0.93, and the predicted versus observed graft losses were not significantly different. These data confirmed the external validation of the full and abbreviated iBox Scoring System. Results 6.5.2.1 (External validation on the qualification datasets).

- Discrimination (c-statistics) was also included for the European validation cohort (c-statistic = 0.81, 95% confidence interval [CI] 0.78 to 0.84) and the three RCTs, [CERTITEM (c-statistic = 0.88), RITUX ERAH (c-statistic = 0.77), and BORTEJECT (c-statistic = 0.94)] described in Loupy et al., 2019 as additional data supporting this qualification submission. Results 6.5.2.2 (External validation on the European cohort and three RCTs from Loupy et al., 2019).

- The ability of the iBox Scoring System to demonstrate a treatment effect at one-year that translates into a treatment effect on death-censored five-year graft survival was assessed in two ways. First, TLS was performed but, due to insufficient data (i.e., only two prospective RCTs and a mTORi derivation subset), it was not possible to provide the precise estimation of the trial-level correlation coefficient. Study level treatment effects in the BENEFIT RCT, BENEFIT EXT RCT, and a mTORi derivation subset using a calcineurin inhibitor (CNI) free regimen, mammalian target of rapamycin (mTORi) such as sirolimus or everolimus versus CNI-based regimen data from Loupy et al., 2019 qualification derivation data for one-year iBox scores for the full and abbreviated iBox Scoring System and five-year death-censored allograft survival were also assessed. The average iBox score at one year was consistently significantly lower in the CNI-free arm (belatacept [BELA] or mTORi) compared to CNI arms. The five-year death-censored allograft survival also consistently numerically favored the CNI-free arm. At five-years in the BENEFIT RCT, death-censored allograft survival was significantly better with BELA compared to CsA. Analyses of the BENEFIT RCT included imputation of the worst-case iBox Scoring System at one-year post-transplant for recipients who died or lost their graft in the first year. This sensitivity analysis was performed to replicate the clinical trial setting where avoidance of survivor bias at one year would be necessary, and all randomized subjects would have an iBox score at one-year even if there were death or graft loss before that time. The totality of these data demonstrate that the iBox Scoring System can measure treatment effects at one-year that translate into a consistent impact on the five-year death-censored allograft survival. The lack of statistical significance on some of the five-year death-censored allograft survival analysis is related to limitations in power to detect differences based on sample size. Results 6.6.3 (Trial-level surrogacy and treatment effect analyses).

Based on these analyses, the full or abbreviated iBox Scoring System models at one-year post-transplant is a validated surrogate for the five-year death-censored allograft survival and is applicable for use in a prospective RCT with imputation for deaths and graft losses within the first year of transplant. Qualification of the iBox Scoring System as a surrogate endpoint would significantly improve upon the current standard, as it would allow drug sponsors the ability to design trials assessing the superiority, of a novel agent. As a surrogate endpoint for the long-term outcome of allograft survival, the iBox Scoring System would allow drug sponsors to seek marketing authorisation of novel agents through EMA’s CMA process while planning and conducting additional studies to demonstrate longer-term therapeutic effects, thereby significantly improving the drug development landscape by encouraging drug sponsors to engage in this therapeutic area of high unmet need. Ultimately, kidney transplant recipients will benefit from the increased drug development activity by improving access to ISTs with better short-term and long-term outcomes.

2.7 Overall goal of the present submission

The TTC presents this Briefing Dossier to request a Qualification Opinion from the Agency on the proposed COU for the iBox Scoring System at one-year post-transplant as a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications via CMA process. The TTC believes a Qualification Opinion is critical for accelerating the development of ISTs in kidney transplantation clinical trials.

3 BACKGROUND

3.1 Regulatory history

The TTC then had formal and informal interactions with EMA and U.S. Food and Drug Administration (FDA) described below.

3.1.1 Regulatory history with EMA

In March of 2020, the TTC provided a summary of this proposed qualification effort to Thorsten Vetter, Scientific Director with EMA. The summary document can be found in Appendix (Qualification of iBox Scoring System to EMA). As described in the summary correspondence to EMA, it was noted that various members of the global transplant community had independent interactions with EMA. As the TTC is a pre-competitive public-private partnership, summary documents were shared with the TTC by Novartis with feedback received from EMA regarding the application of the iBox Scoring System as a clinical trial endpoint. Additionally, the European Society of Transplantation (ESOT) shared Qualification Advice feedback received from the Scientific Advice Working Party (SAWP) regarding the application of the iBox Scoring System as a clinical trial endpoint.

EMA previously provided feedback to both ESOT and Novartis. TTC included this information in the development of this Briefing Dossier for submission, as described below:

- **Formal regulatory review of the iBox Scoring System to be considered as an endpoint:** This Briefing Dossier is submitted to EMA's pathway for Qualification of Novel Methodologies in Drug Development.

- **Provide iBox Scoring System formula:** The statistical model and the iBox Scoring System algorithm with the relative contribution of each factor of the model is provided in this Briefing Dossier (Results 6.2 (Multivariate analysis). Additionally, the patient-level data and codes for analyses are included.

- **Clarity regarding datasets designed for validation purposes and which were used for iBox Scoring System via post-hoc analysis:** The derivation and validation datasets in this qualification submission are defined and labeled throughout this Briefing Dossier. More detail can be found Methods 4.3.1 (Introduction to data).

- **Assess the technical validity of the iBox Scoring System for the proposed COU:** The COU in this submission pre-specified the iBox Scoring System risk evaluation at one-year post-transplantation for use in a clinical trial endpoint at a fixed landmark. A full and detailed description of the technical validity can be found in Results 6.5.2.1 (External validation on the qualification datasets).

- **Calibration of the iBox Scoring System in a CNI-free setting and other analyses to support hemodynamic effect not being a confounding factor:** Additional analyses were performed to assess the treatment effect (i.e., CNI-free versus CNI) in the iBox Scoring System in BELA treated transplant recipients (BENEFIT and BENEFIT EXT RCTs), and mTORi treated recipients (PTG dataset). A full and detailed description can be found in Results 6.5.2.1 (External validation on the qualification datasets).

- **Evaluation of death-censored allograft survival versus competing risks (i.e., death) to assess the extent of iBox validation:** the iBox Scoring System was designed to assess long-term risk of allograft failure. Graft failure is defined as return to dialysis or pre-emptive re-transplantation. Death of the recipient with a functioning graft was not part of the iBox design since patient death has a variety of underlying causes (e.g., malignancy, infection, cardiovascular disease), and different risk factors compared with those for graft failure. Sensitivity analyses were performed to test the performance of the iBox Scoring System in overall graft survival. A full and detailed description of the competing risk analysis can be found in Results 6.6.2 (Competing risk analysis).

- **Demonstrate that the effect observed on the surrogate outcome is strongly predictive for an effect on the true outcome (i.e., TLS):** TLS and treatment effect analyses were conducted. A full and detailed description can be found in Results 6.6.3 (Trial-level surrogacy and treatment effect analyses).

- **Develop and implement model for imputing missing iBox data/components:** By having the abbreviated iBox Scoring System model, missing biopsy data are covered since biopsy is not a factor in this model. Additionally, imputation methodology was used to account for graft loss, death, or lost to follow-up within the first year of transplantation. A full and detailed description can be found in Modeling analysis methodologies 5.5.3.1 (Imputation of iBox Scores).

C-Path and EMA held a preparatory meeting on 28 January 2022. The objectives of the meeting were to:

- Orient EMA to the Briefing Package submitted for qualification opinion of the iBox Scoring System (Composite Biomarker Panel).

- Discuss the evidence submitted for regulatory endorsement of the iBox Scoring System.

- Answer questions and/or provide clarification.

- Align on any optimizations needed for inclusion in the re-submission.

A summary of the meeting minutes can be found in Appendix (C-Path – EMA Preparatory Meeting 28 January 2022).

3.1.2 Regulatory history with FDA

On the 14th of February 2020, the TTC submitted a Letter of Intent to FDA. On the 1st of June 2020, the TTC received a favorable Determination Letter accepting the iBox Scoring System into the FDA Biomarker Qualification Program. In the Determination Letter, FDA agreed there is an unmet need, and the development of this composite scoring system to predict subjects' long-term outcomes in clinical trials will facilitate the development of novel ISTs. FDA suggested referring to the biomarker as a 'composite biomarker panel' instead of "The Integrative Box (iBox) Scoring System." To acknowledge the work of the PTG, led by Dr. Loupy, to develop the iBox Scoring System and take FDA's recommendation into action, the TTC has updated the name of the surrogate endpoint to the iBox Scoring System (Composite Biomarker Panel). Currently, the TTC is preparing a Qualification Plan for submission to FDA while the process with EMA moves forward. A full and detailed description of the FDA comments on the Letter of Intent can be found in Appendix (FDA Letter of Intent Determination Letter 2020-01-06).

The feedback from FDA in the Determination Letter was incorporated in this Briefing Dossier to enhance this Qualification Opinion submission, as described below.

- COU considerations; A full and detailed description can be found in Background 3.2 (Proposed COU statement).

- Analytical considerations; A full and detailed description can be found in Appendix (Revised-Analytical considerations).

- Clinical considerations; A full and detailed description can be found in Appendix (Revised-Clinical considerations).

- Statistical considerations; A full and detailed description can be found in Modeling analysis methodologies (Section 5).

3.2 Proposed context-of-use statement

Proposed context-of-use statement

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA.

General area:

Surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications.

Target population for use of the biomarker:

Adult *de novo* kidney only transplant recipients from a living or deceased donor.

Stage of drug development for use:

All clinical efficacy evaluation stages of therapeutic interventions focused on the use of the long-term risk of allograft survival in kidney transplant recipients, including early signs of efficacy, proof-of-concept, dose-ranging, and registration studies (Phases II-IV).

Intended application:

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications via CMA. When evaluating five-year outcomes for clinical benefit and full marketing authorisation, it will be necessary to ensure that there is not a clinically meaningful decrease in transplant recipient survival with the new therapy in the clinical trial compared to the standard of care control arm.

4 METHODS

4.1 Introduction

The purpose of this analysis is to develop a semiparametric Cox PH model, the iBox Scoring System, that describes the association between the time to death-censored allograft survival and predictor variables in kidney transplant subjects. The developed model will account for several predictors of graft loss within the defined subject population. For the purpose of this submission, the time to evaluation was fixed at one-year post-transplant. The model is intended to provide the necessary evidence to support its use as a surrogate endpoint for the five-year risk of death-censored allograft loss in kidney transplant subjects for use in clinical trials.

4.1.1 Rationale to support two iBox Scoring System models

In the Loupy et al., 2019 publication, subjects from four hospitals in the French national health care system had kidney biopsies performed post-transplantation based on either clinical indication (i.e., in the presence of renal dysfunction) or per protocol biopsies (i.e., surveillance biopsies performed in the absence of signs or symptom of rejection) to assess for histological findings consistent with acute or chronic rejection. (Ahmad 2004).

The additional information from biopsies needs to be weighed against the challenges of obtaining protocol/surveillance biopsies in all subjects within multinational, multicenter clinical trials, and the risk of performing biopsies which includes the potential for bleeding, obstruction due to clotting, renal fistulas, and hematuria. Additionally, transplant recipients have the right to decline a biopsy. By having two iBox Scoring System models, one with and the other without biopsy findings, a sponsor can assess the ability to perform surveillance biopsies and, if impractical or not feasible, design a simpler, less burdensome clinical trial, with the knowledge that both models perform well.

Table 5 and Table 6 summarize the number of subjects from the qualification datasets to support both iBox Scoring System models, a full (with biopsy) and an abbreviated (without biopsy). Additionally, the BENEFIT and BENEFIT-EXT RCTs, although included biopsies at one-year post-transplant in the protocol, had several subjects without biopsy data. This further supports the challenges of implementing protocol biopsies in a multinational, multicenter clinical trial and the benefits of having two iBox Scoring System models, one with and one without biopsy findings.

Table 5. Qualification derivation dataset to support full and abbreviated iBox Scoring System models

Dataset	Full iBox Scoring System	Abbreviated iBox Scoring System
Loupy et al., 2019 derivation	**Number of subjects**	
	n =3,941	n = 4,000

Table 6. Qualification validation datasets to support full and abbreviated iBox Scoring System models

Dataset	Full iBox Scoring System	Abbreviated iBox Scoring System
	Number of subjects	
Mayo Clinic Rochester	n = 483	n = 497
Helsinki University Hospital	n = 344	n = 344
BENEFIT RCT	n = 416	n = 515
BENEFIT-EXT RCT	n = 260	n = 357

4.2 Context-of-use

Proposed context-of-use statement

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA.

General area:

Surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications.

Target population for use of the biomarker:

Adult *de novo* kidney only transplant recipients from a living or deceased donor.

Stage of drug development for use:

All clinical efficacy evaluation stages of therapeutic interventions focused on the use of the long-term risk of allograft survival in kidney transplant recipients, including early signs of efficacy, proof-of-concept, dose-ranging, and registration studies (Phases II-IV).

Intended application:

The iBox Scoring System (Composite Biomarker Panel) used at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant subjects for use in clinical trials to support evaluation of novel IST applications via CMA. When evaluating five-year outcomes for clinical benefit and full marketing authorisation, it will be necessary to ensure that there is not a clinically meaningful decrease in transplant recipient survival with the new therapy in the clinical trial compared to the standard of care control arm.

4.3 Data

4.3.1 Introduction to data

The qualification derivation presented in this Briefing Dossier includes specific adjustments to the original derivation dataset as described in Loupy et al., 2019, allowing the iBox Scoring System to be used as a one-year post-transplant surrogate endpoint in clinical trials. (4.3.8 Alignment of qualification datasets).

The qualification validation presented in this Briefing Dossier includes datasets other than those used for external validation in Loupy et al., 2019 (Alexandre Loupy et al. 2019) to support the COU of the proposed surrogate endpoint at one-year post-transplantation in a clinical trial at a fixed landmark. However, the European validation cohort and the three RCTs described in Loupy et al., 2019 are summarized in this Briefing Dossier as additional data supporting this qualification submission. Methods 4.3.5 and 4.3.6 (Loupy et al., 2019 European validation cohort and Loupy et al., 2019 External validation in three randomized controlled trials [RCTs]).

External and independent validation of the iBox Scoring System required datasets that included the core iBox variables taken at one-year post-transplant (i.e., eGFR, proteinuria, kidney allograft biopsy histopathology, and DSA) as well as a follow-up period sufficient to evaluate long-term graft survival (i.e., at least five years). The number of available datasets that included all core variables and sufficient follow-up to at least five years was limited.

As discussed in the Executive summary 2.3 Sources of data, the five datasets had the requisite subject-level data to conduct the internal and external validation analyses in this Briefing Dossier for a Qualification Opinion submission. These datasets were acquired from clinical transplant centers (i.e., Loupy et al., 2019 derivation, Mayo Clinic Rochester, and Helsinki University Hospital) and clinical trials (i.e. [BENEFIT RCT] Vincenti et al., 2012 and [BENEFIT-EXT RCT] Medina-Pestana., 2012) representing over 5,500 *de novo* kidney transplant recipients. The subject-level data received from clinical transplant centers are inherently heterogeneous and reflect the diversity of the kidney transplant recipient population globally, as demonstrated in Table 8. In addition, the two clinical trials included in this qualification submission have the most extensive CNI-free patient-level data available with the four core variables and sufficient follow-up period.

The qualification derivation and validation datasets were curated and used to develop and validate, respectively, the full and abbreviated iBox Scoring System. A semiparametric survival modeling approach was used to develop the full and abbreviated iBox Scoring System. The variety in datasets were fundamental to developing sufficient evidence to support the biological plausibility, causality, universality, proportionality, and specificity of the marker. As described in Methods 4.4 Data exclusions, it was expected that some data received would be excluded from the final modeling due to a number of commonly encountered issues known to be associated with efforts to achieve alignment across datasets.

The summary-level characteristics of the qualification derivation and validation datasets selected for the overall analysis to support both iBox Scoring System models are shown in Table 7-12:

- Table 7 and Table 8: Overview of the qualification derivation and validation datasets.
- Table 9 and Table 10: Subject characteristics across qualification datasets for full and abbreviated iBox Scoring System models.
- Table 11 and Table 12: Core composite features across the qualification validation datasets for full and abbreviated iBox Scoring System models.

Table 7. Overview of qualification derivation dataset

Data name	Data type	Geography	Median follow-up after transplantation	Full iBox Scoring System	Abbreviated iBox Scoring System
Loupy et al., 2019 Derivation	Transplant centers	Europe	7.0 years	n = 3,941	n = 4,000

Table 8. Overview of qualification validation datasets

Data name	Data type	Geography	Median follow-up after transplantation	Full iBox Scoring System	Abbreviated iBox Scoring System

Mayo Clinic Rochester	Transplant center	North America	7.6 years	n = 483	n = 497
Helsinki University Hospital	Transplant center	Europe	8.5 years	n = 344	n = 344
BENEFIT	RCT	International*	7.0 years	n = 416	n = 515
BENEFIT-EXT	RCT	International**	7.0 years	n = 260	n = 357
Total				n = 1,503	n = 1,713

* North America, South America, Europe, Australia, Africa, and Asia

** North America, South America, Europe, Australia, and Africa

Table 9. Subject characteristics across qualification datasets for Full iBox Scoring System

	Qualification derivation dataset	**Qualification validation datasets**			
	Loupy et al., 2019 Derivation **n = 3,941**	**Mayo Clinic Rochester** **n = 483**	**Helsinki University Hospital** **n =344**	**BENEFIT RCT*** **n = 416**	**BENEFIT-EXT RCT*** **n = 260**
Recipient demographics					
Age (years) (mean, S.D.)	49.8, 13.68	50.0, 13.58	52.0, 13.06	42.1, 13.81	54.0, 12.66
Race (No., %)	Not documented	Black, 3% Nonblack, 97%	Black, 0.3% Nonblack, 99.7%	Black, 6% Nonblack, 94%	Black, 12% Nonblack, 88%
Sex (mode, %)	Male, 61%	Male, 56%	Male, 65%	Male, 69%	Male, 67%
Transplant characteristics					
Donor age (years) (mean, S.D.)	51.6, 16.35	43.6, 12.95	52.2, 14.16	39.5, 11.44	54.4, 14.29

Donor type (mode, %)	Deceased, 83%	Deceased, 21%	Deceased, 96%	Deceased, 40%	Deceased, 100%
Previous kidney transplant (yes, %)	15%	20%	10%	3%	0%
Total No of HLA-A/B/DR mismatches (mean, % >3)	3.82, 62%	3.41, 51%	2.38, 10%	3.16, 38%	3.34, 47%
CIT (hours) (mean)	16	4	22	7	21
DSA at time of transplantation (yes, %)	18%	10%	5%	6%	7%
DGF (yes, %)	26%	4%	37%	14%	49%
Induction					
Induction (yes, %)	94%	100%	10%	100%	100%
IL-2Ra (%)	44%	22%	100%	100%	100%
Lymphocyte-depleting agent (%)	56%	73%	0%	0%	0%
Baseline maintenance IST					
Maintenance (yes, %)	100%	100%	100%	100%	100%
CNI-based (%) **% TAC, % CsA**	91% 71%, 29%	99.6% 99.8%, 0.2%	100% 27%, 73%	32% 0%, 100%	33% 0%, 100%
CNI-free (%)	7%	0.4%	0%	68%	67%

% Bela, % mTORi	0%, 60%	0%, 0%		100%, 0%	100%, 0%
mTORi and CNI (%)	2%	0%	0%	0%	0%

* All IST included mycophenolate and corticosteroids

Table 10. Subject characteristics across qualification datasets for Abbreviated iBox Scoring System

	Qualification derivation dataset	**Qualification validation datasets**			
	Loupy et al., 2019 Derivation n = 4,000	Mayo Clinic Rochester n = 497	Helsinki University Hospital n = 344	BENEFIT RCT* n = 515	BENEFIT-EXT RCT* n = 357
Recipient demographics					
Age (years) (mean, S.D.)	49.8, 13.70	50.0, 13.69	52.0, 13.06	42.7, 13.68	55.0, 12.82
Race (No., %)	NA	Black, 3% Nonblack, 97%	Black, 0.3% Nonblack, 99.7%	Black, 8% Nonblack, 92%	Black, 12%, Nonblack, 88%
Sex (mode, %)	Male, 61%	Male, 56%	Male, 65%	Male, 69%	Male, 66%
Transplant characteristics					
Donor age (years) (mean, S.D.)	51.7, 16.33	43.6, 12.87	52.2, 14.16	40.0, 11.77	55.6, 13.92
Donor type (mode, %)	Deceased, 83%	Deceased, 21%	Deceased, 96%	Deceased, 41%	Deceased, 100%
Previous kidney transplant (yes, %)	15%	20%	10%	3%	0%

Total No of HLA-A/B/DR mismatches (mean, % >3)	3.82, 62%	3.42, 51%	2.38, 10%	3.19, 38%	3.38, 50%
CIT (hours) (mean)	16	4	22	8	21
DSA at time of transplantation (yes, %)	18%	10%	5%	6%	6%
DGF (yes, %)	26%	4%	37%	14%	46%
Induction					
Induction (yes, %)	94%	100%	10%	100%	100%
IL-2Ra (%)	44%	23%	100%	100%	100%
Lymphocyte-depleting agent (%)	56%	73%	0%	0%	0%
Baseline maintenance IST					
Maintenance (yes, %)	100%	100%	100%	100%	100%
CNI-based (%) **% TAC, % CsA**	91% 71%, 29%	100% 99.8%, 0.2%	100% 27%, 73%	33% 0%, 100%	32% 0%, 100%
CNI-free (%) **% BELA, % mTORi**	7% 0%, 61%	0.4% 0%, 0%	0%	67% 100%, 0%	68% 100%, 0%
mTORi and CNI (%)	2%	0%	0%	0%	0%

* All IST included mycophenolate and corticosteroids

Table 11. Core composite features across qualification validation datasets at one-year post-transplant for full iBox Scoring System

		Qualification validation datasets			
		Mayo Clinic Rochester Total (n = 483) No failure[‖] (n = 465) Failure[ǂ] (n = 18)	Helsinki University Hospital Total (n = 344) No failure[‖] (n = 323) Failure[ǂ] (n = 21)	BENEFIT RCT* Total (n = 416) No failure[‖] (n = 404) Failure[ǂ] (n = 12)	BENEFIT-EXT RCT Total (n = 260) No failure[‖] (n = 248) Failure[ǂ] (n = 12)
eGFR (ml/min/1.73m^2) (mean, SD)	Total	55.92, 14.67	61.41, 20.04	66.01, 19.14	51.1, 16.03
	No failure[‖]	56.75, 14.10	62.34, 19.38	66.17, 18.95	51.97, 15.47
	Failure[ǂ]	34.69, 13.47	47.03, 24.64	60.37, 25.17	33.03, 17.44
Log transformed UPCR estimate (g/g) (mean, SD)	Total	0.20, 0.47	0.21, 0.36	0.22, 0.32	0.29, 0.50
	No failure[‖]	0.15, 0.19	0.20, 0.32	0.21, 0.31	0.28, 0.50
	Failure[ǂ]	1.42, 1.87	0.44, 0.70	0.50, 0.45	0.42, 0.42
DSA **Yes, %**	Total	12%	6%	4%	5%
	No failure[‖]	11%	5%	4%	4%
	Failure[ǂ]	28%	19%	17%	17%
Kidney allograft biopsy histopathology					
IFTA score **0-1, 2, 3 (%)**	Total	90.5, 7.5, 2	95.9, 3.5, 0.6	94.5, 3.4, 2.2	89.2, 6.6, 4.2
	No failure[‖]	91, 7, 2	96.3, 3.4, 0.3	95, 3, 2	90, 6, 4

	Failure[ǂ]	72, 22, 6	90, 5, 5	92, 8	83.3, 8.3, 8.3
g + ptc score **0-2, 3-4, 5-6 (%)**	Total	90, 9, 1	100, 0, 0	99.3, 0.5, 0.2	99.2, 0.4, 0.4
	No failure[‖]	91, 8, 1	100, 0, 0	99.2, 0.5, 0.3	99.2, 0.4, 0.4
	Failure[ǂ]	66.7, 27.8, 5.6	100, 0, 0	100, 0, 0	100, 0, 0
i + t score **<3, ≥ 3 (%)**	Total	94, 6	97, 3	94, 6	97, 3
	No failure[‖]	94, 6	98, 2	95, 5	97, 3
	Failure[ǂ]	89, 11	86, 14	67.7, 33.3	83, 17
cg score **<1, ≥ 1 (%)**	Total	94, 6	98.3, 1.7	100, 0	99.5, 0.38
	No failure[‖]	95, 5	98, 2	100, 0	99.6, 0.4
	Failure[ǂ]	66.7, 33.3	95, 5	100, 0	100, 0

* Serum creatinine values >30.5 were assumed to be recorded as umol/L; these extreme values were converted to mg/dL after confirming with BMS.

[‖] No failure in these tables refer to the number of transplant recipients who did not experience graft loss by five years post-transplant.

[ǂ] Failure in these tables refers to transplant recipients whose graft failed by five years post-transplant.

Table 12. Core composite features across qualification validation datasets at one-year post-transplant for abbreviated iBox Scoring System

	Qualification validation datasets			
	Mayo Clinic Rochester Total (n = 497)	Helsinki University Hospital Total (n = 344)	BENEFIT RCT* Total (n = 515)	BENEFIT-EXT RCT Total (n = 357)

		No failure[‖] (n = 477) Failure[ǂ] (n = 20)	No failure[‖] (n = 323) Failure[ǂ] (n = 21)	No failure[‖] (n = 500) Failure[ǂ] (n = 15)	No failure[‖] (n = 334) Failure[ǂ] (n = 23)
eGFR (ml/min/1.73m^2) (mean, SD)	Total	55.89, 14.58	61.41, 20.04	65.81, 18.67	51.39, 16.26
	No failure[‖]	56.61, 13.98	62.34, 19.38	66.01, 18.52	52.31, 15.23
	Failure[ǂ]	38.72, 18.02	47.03, 24.64	59.26, 22.68	38.00, 23.84
Log transformed UPCR estimate (g/g) (mean, SD)	Total	0.20, 0.46	0.21, 0.36	0.22, 0.36	0.30, 0.52
	No failure[‖]	0.15, 0.19	0.20, 0.32	0.22, 0.35	0.29, 0.50
	Failure[ǂ]	1.29, 1.82	0.44, 0.70	0.43, 0.42	0.50, 0.69
DSA **Yes, %**	Total	11%	6%	4%	4%
	No failure[‖]	11%	5%	4%	4%
	Failure[ǂ]	30%	19%	13%	9%

* Serum creatinine values >30.5 were assumed to be recorded as umol/L; these extreme values were converted to mg/dL after confirming with BMS.

[‖] No failure in these tables refer to the number of transplant recipients who did not experience graft loss by five years post-transplant.

[ǂ] Failure in these tables refers to transplant recipients whose graft failed by five years post-transplant.

4.3.2 Analytical considerations

The laboratories that analyzed the biochemistry (serum creatinine), urinalysis (proteinuria), and DSA assays, as well as the histopathology laboratories that prepared the biopsy samples, have maintained accreditation and/or certifications during the entirety of the data collection period. Each laboratory determined performance specifications and was responsible for the quality of the results generated from the assays. The serum creatinine, proteinuria, and DSA assays were granted 510(k)-clearance by FDA Center for Devices and Radiological Health (CDRH) and/or have received Conformité Européenne (C.E.) markings by the European Economic Area (EEA). The assays were implemented according to the manufacturers' instructions for use (IFU). The analytical considerations documentation summarizing the assays and laboratory certification/accreditation document can be found in the analytical considerations document included in this Briefing Package. Copies of the laboratory certification/accreditation documentation are also included in this Briefing Package. C-Path has reviewed the documentation and deemed that the analytical methods were robust, reliable, and fit-for-purpose.

4.3.3 Qualification derivation dataset

The qualification derivation dataset included 4,000 subjects for the abbreviated iBox Scoring System and 3,941 subjects for the full iBox Scoring System. Subjects were over 18 years old when they were prospectively enrolled at the time of transplantation from a living or deceased donor between 1 January 2005 and 1 January 2014 at one of the following four French transplant centers. These French centers include Necker Hospital, Paris, France (n = 1,473), Saint-Louis Hospital, Paris, France (n = 928), Foch Hospital, Suresnes, France (n = 714), and Toulouse Hospital, Toulouse, France (n = 885). Subjects with grafts that never functioned (i.e., Primary nonfunction of the graft [PNF]) were excluded. All subjects provided written informed consent at the time of transplantation. For application as an endpoint in a clinical trial at one-year, additional analyses were conducted on the qualification derivation dataset, restricting the analysis to those recipients with an iBox Scoring System evaluation at one-year post-transplant and follow-up to five-years for graft loss. Summarized in Table 13.

Table 13. Summary of the qualification derivation dataset

	Total number of subjects n	No. of subjects with abbreviated iBox Scoring System n, %	No. of subjects with full iBox Scoring System n, %	No. of subjects with abbreviated iBox Scoring System at one-year post-transplant n, %	No. of subjects with full iBox Scoring System at one-year post-transplant n, %
Loupy et al., 2019 Derivation **2005-2014**	4,000	4,000 (100)	3,941 (99)	1,180 (30)	1,174 (30)

Table 14 provides an overview of the baseline maintenance IST regimen for the 3,941 subjects in the qualification derivation dataset for the full iBox Scoring System. All subjects in this cohort were known to have been prescribed maintenance IST. In addition, most subjects were on a CNI-based regimen (91%), with tacrolimus (TAC) being the most commonly prescribed CNI medication (71%). The breakdown of baseline maintenance IST regimen information for the 4,000 subjects in the abbreviated iBox Scoring System is consistent with these findings.

Table 14. Baseline maintenance IST regimens in the qualification derivation dataset for full iBox Scoring System

Baseline maintenance IST regimens **n = 3,941 (100% on baseline IST)**	
CNI-based	n = 3,590 (91%)

TAC
CsA
	n = 2,549 (71%)
	n = 1,041 (29%)

BELA	n = 0 (0%)

mTORi
Everolimus
Sirolimus
	n = 171 (4%)
	n = 62 (36%)
	n = 109 (64%)

CNI + mTORi	n = 68 (2%)
Mycophenolate and corticosteroids only	n = 112 (3%)

The anonymized data collected was prospectively entered at the time of transplantation and at the time of post-transplant graft biopsies (Alexandre Loupy et al. 2019). For subjects with multiple biopsies, risk evaluation was performed using results from the first post-transplant biopsy. The cut-off date for data collection was March 2018. All graft losses occurring before the data cut-off date were included in the analysis. A full description of the 31 candidate variables considered for inclusion in the full iBox Scoring System, as described in Loupy et al., 2019 (Alexandre Loupy et al. 2019), can be found in Methods 4.3.3.3 (Model variables).

In the Loupy et al., 2019 publication, the time of risk evaluation varied between subjects, but the laboratory and biopsy measurements for each subject were taken on the same day. The heterogeneity of evaluation time was accounted for as a covariate. In this Briefing Dossier, the time of risk evaluation was fixed at one-year post-transplant for all subjects, consistent with the presented COU statement. The TTC seeks to describe the entirety of the data (regardless of the time of risk evaluation) used in the model derivation. Of the 4,000 subjects who received a kidney transplant at one of the four French transplant centers described above between 2005-2014, there were 3,941 subjects evaluated for the full iBox Scoring System and 4,000 subjects were evaluated for the abbreviated iBox Scoring System. The proposed components of the iBox Scoring System include:

1. eGFR [SCr]: calculated with SCr and is based on the MDRD equation; A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations). Current equations used to estimate the GFR incorporate race as a variable, such as the MDRD-186 Study equation used in this qualification submission. C-Path recognizes the recent literature by Inker et al. 2021 supporting a revised CKD-EPI equation without race input. This recent publication found that eGFR calculations without race were more accurate and with smaller differences between race groups than current eGFR equations. C-path explored the CKD-EPI 2021 equation as well as MDRD-175 Study equation (both shown in Table 2 of the Analytical Considerations document) and found that the results presented in section 6 of this dossier are nearly identical no matter which eGFR equation is used.
2. Measurement of protein excretion into the urine ('proteinuria'): calculated with urine protein-to-creatinine ratio in gram per gram (g/g); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).

3. Histopathological assessment of tissue obtained by renal allograft biopsy, using the Banff 2015 scoring criteria ('kidney allograft biopsy histopathology'); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
4. Presence of ('DSA'). Additionally, the presence of the DSA was refined into a qualitative binary category based on MFI values. A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).

4.3.3.1 Cross-sectional time point used for analysis set

In the Loupy et al., 2019 publication, time of risk evaluation varied between subjects but the laboratory and biopsy measurements for each subject were taken on the same day. The heterogeneity of evaluation time was accounted for as a covariate.

In this Briefing Dossier, all of the PTG data were used for the qualification derivation and internal validation. Additional analyses of the PTG data were performed where the time of risk evaluation was fixed at one-year post-transplant for all subjects, consistent with the presented COU statement.

4.3.3.2 Time of graft loss, censoring, and competing risks

A variable was derived, denoted T_event and defined as either the time from risk evaluation to graft loss or the time from risk evaluation to the last recorded visit day for individuals with no recorded graft loss time. For individuals with no recorded graft loss, T_event was considered the right-censored time since the event of graft loss was unobserved. Specific to the derivation cohort, graft loss was defined as a subject's definitive return to dialysis or preemptive kidney transplantation, as defined as eGFR less than 10ml/min/1.73m^2.

Survival analysis may be confounded by competing risks, i.e., events that preclude the event of interest. If a death with a functioning graft occurred, graft failure was not observed. To ensure death did not confound the analysis of graft failure, a supplementary competing risk analysis was performed, described in Modeling analysis methodologies 5.5.2 (Competing risk analysis), assessing the impact, if any, of death on the model of graft failure.

4.3.3.3 Model variables

There were 31 candidate variables considered for inclusion in the multivariable Cox PH model, as described in Loupy et al., 2019 (Alexandre Loupy et al. 2019). These candidate variables are commonly and routinely collected in kidney transplant centers worldwide (Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009).

These 31 variables were categorized into five groups: recipient characteristics, transplant characteristics, functional characteristics, post-transplantation histopathology variables, and post-transplantation immunological variables, each assessed at the time of risk evaluation.

- Nine variables were excluded in backward elimination due to clinical considerations; these can be found in Methods 4.3.3.4 (Clinical considerations).

- The rationale for the categorical breakdown of candidate variables in the univariate analysis can be found in Methods 4.3.3.5 (Categorical breakdown of candidate variables in the univariate analysis).

- Rationale for the categorical breakdown of candidate variables in multivariate analysis can be found in Methods 4.3.3.6 (Categorical breakdown of candidate variables in multivariate analysis).

- Rationale for exclusion of specific individual Banff lesion scores excluded from univariate and multivariate analyses can be found in Methods 4.3.3.7 (Individual Banff lesion scores excluded from univariate and multivariate analyses).

Recipient characteristics [ERA-EDTA, AJT, KDIGO 2009]: Recipient age (per one-year increment) and sex.

Transplant characteristics [ERA-EDTA, AJT, KDIGO 2009]: Donor age (per one-year increment), donor sex, donor type (living or deceased), donor history of hypertension, donor history of diabetes, donor creatinine concentration (< 1.5 mg/dL or ≥ 1.5 mg/dL), expanded criteria donor (ECD), previous kidney transplant, CIT (< 12 hours, 12-24 hours, ≥ 24 hours) CIT citations [(Summers et al. 2013), (Debout et al. 2015), (Aubert et al. 2015), (Peters-Sengers et al. 2019)], Thymoglobulin™ (anti-thymocyte globulin [rabbit]) induction immunosuppression, number of human leukocyte antigen (HLA)-A/B/DR mismatches, delayed graft function (DGF), pre-existing DSA.

Allograft functional variables [references: KDIGO 2009]: eGFR and log proteinuria. eGFR is measured in ml/min/1.73m^2 and proteinuria is measured using spot estimation (log transformed urine protein-to-creatinine ratio) with a result given in g/g of creatinine. Proteinuria was log-transformed because of the skewed distribution to ensure the normality of continuous parameters in the Cox PH model.

SCr allows an eGFR calculation using the 4-variable MDRD-186 Study equation. (Levey et al. 2006).

Post-transplantation allograft structural histopathology variables (A. Loupy et al. 2017): Allograft biopsies were scored and graded from 0 to 3 according to the Banff 2015 criteria for allograft pathology for the following histological factors: Interstitial fibrosis/tubular atrophy (IFTA score)*, vascular fibrous intimal thickening (cv score), arteriolar hyalinosis (ah score), interstitial inflammation and tubulitis (i score and t score), transplant glomerulopathy (cg score), intimal arteritis (v score), C4d graft deposition, microcirculation inflammation (g score and ptc score).

Additional diagnoses provided by the biopsy included: polyomavirus-associated nephropathy (PVAN) (Drachenberg and Papadimitriou 2006), nephropathy recurrence (W. H. Lim, Shingde, and Wong 2019), aAMR (A. Loupy et al. 2017), and acute T cell-mediated rejection (aTCMR) (A. Loupy et al. 2017). All biopsies were graded by trained nephro-pathologists according to the international Banff 2015 criteria for allograft pathology (A. Loupy et al. 2017). Biopsies performed before 2015 were re-scored by nephro-pathologists at their respective sites.

* IFTA replaced chronic allograft nephropathy (CAN) in 2005 (Solez et al. 2007). IFTA is the association between two other individual Banff lesion scores, ci (interstitial fibrosis) and ct (tubular atrophy). IFTA scores were derived from ci and ct scores in the qualification validation cohort.

The IFTA score in the qualification validation cohort corresponded to the following definition, as described in Roufosse et al., 2018 (Roufosse et al. 2018).

ci score:

ci0 – interstitial fibrosis in up to 5% of cortical area.
ci1 – interstitial fibrosis in 6 to 25% of cortical area (mild interstitial fibrosis).
ci2 – interstitial fibrosis in 26 to 50% of cortical area (moderate interstitial fibrosis).
ci3 – interstitial fibrosis in >50% of cortical area (severe interstitial fibrosis).

ct score:

ct0 – no tubular atrophy.
ct1 – tubular atrophy involving up to 25% of the area of cortical tubules.
ct2 – tubular atrophy involving 26 to 50% of the area of cortical tubules.
ct3 – tubular atrophy involving >50% of the area of cortical tubules.

IFTA Banff Grade was determined based on the higher of the two lesion scores, ci and ct:

Grade 0 : ci0 and ct0.
Grade 1 : ci1 or ct1.
Grade 2 : ci2 or ct2.
Grade 3 : ci3 or ct3.

Post-transplantation recipient immunological profile variables [(Tait et al. 2013), (Schinstock et al. 2020), (Lachmann et al. 2013)]: Presence of DSA as a qualitative binary category based on MFI values.

DSA and the corresponding MFI were identified at the time of post-transplant for the full and abbreviated iBox Scoring System risk assessment. The presence of circulating DSAs against HLA-A, HLA-B, HLA-Cw, HLA-DR, HLA-DQ, and HLA-DP were determined using a single-antigen bead (SAB) assay (One Lambda, Inc., Canoga Park, CA, USA) on a Luminex™ platform. Beads with a normalized MFI, a measure of DSA presence, of greater than 1,400 were deemed positive, as supported by Reed et al., - 2013 - Comprehensive Assessment and Standardization of Solid Phase Multiplex-Bead Arrays for the Detection of Antibodies to HLA. The rationale to support this qualitative binary category based on MFI values for the full and abbreviated iBox Scoring System can be found in Appendix (Revised-Clinical considerations).

Time of post-transplant risk evaluation: Time from transplantation to full and abbreviated iBox Scoring System evaluation, expressed in years.

The final covariates included in the iBox Scoring System are described in Table 15. It was expected that covariates excluded in the iBox Scoring System were not going to be significant predictors, as described in Loupy et al., 2019 (Alexandre Loupy et al. 2019).

Table 15. Final eight covariates in the iBox Scoring System

	Notation	Description of Co-variate at Baseline	Type
1	X_{time}	Time of risk evaluation	Continuous
2	X_{eGFR}	eGFR (in ml/min/1.73m^2)	Continuous
3	$X_{Proteinuria\ (log)}$	Log transformed UPCR (g/g)*	Continuous
4	X_{IFTA}	IFTA score	Ordinal
5	X_{g+ptc}	Microcirculation inflammation (g score and ptc score)	Ordinal
6	X_{i+t}	Interstitial inflammation and tubulitis (i score and t score)	Ordinal (binary)
7	X_{cg}	Transplant glomerulopathy (cg score)	Ordinal (binary)
8	$X_{DSA-MFI}$	DSA MFI	Ordinal (binary)

*Proteinuria values of 0 will have a small positive value added to prevent undefined values

4.3.3.4 Clinical considerations for exclusion of variables

The following nine variables were excluded in backward elimination due to clinical considerations described below.

Expanded criteria donor

ECD is transplantation from a donor aged sixty years or older, or over 50 years with at least two of the following conditions: hypertension history, SCr > 1.5mg/dl, or cause of death from cerebrovascular accident (Port et al. 2002). Thus, donor age, donor hypertension history, and donor creatinine concentration are features present in the diagnosis of ECD.

> Therefore, ECD was included in place of the three individual measures of (1) donor age, (2) donor hypertension history, and (3) donor creatinine concentration.

Previous kidney transplant

There is a historical precedent in the literature demonstrating a high correlation between the presence of DSA at the time of retransplant in subjects with a previous, failed kidney transplant. [(Dunn et al. 2011), (Lefaucheur et al. 2013)].

> Therefore, DSA present at the time of retransplant was used in place of (4) previous kidney transplant.

Delayed graft function

DGF was heterogeneously defined across the French centers in the derivation cohort and thus was not considered in backward elimination. Additionally, DGF is not a baseline covariate known at the time of transplant but is an early post-transplant event. CIT is a well-defined, quantitative, and continuous variable with literature correlation to DGF. [(Summers et al. 2013), (Debout et al. 2015), (Aubert et al. 2015), (Peters-Sengers et al. 2019)].

> Therefore, CIT was used in place of (5) DGF.

Banff 2015 diagnoses

The Banff classification includes individual semiquantitative histologic indices for specific lesions that together define diagnoses. (A. Loupy et al. 2017).

> Therefore, the Banff Individual lesion scores were used in place of the diagnosis of (6) PVAN, (7) nephropathy recurrence, (8) acute antibody-mediated rejection, and (9) acute T cell-mediated rejection.

4.3.3.5 Categorical breakdown of candidate variables in univariate analysis

Several Banff lesion scores were included as candidate variables in the univariate analysis. These include intimal arteritis (v score), vascular fibrous intimal thickening (cv score), and arteriolar hyalinosis (ah score). All scores were reduced to a binary classification for analysis, indicating either presence [≥ 1] or absence [0] of the lesion. The specific rationale for categorizing and including individual lesion scores is described below.

Intimal arteritis

The Banff lesion score, intimal arteritis (v score) received binary categorization based upon the precedent set by existing literature. The presence of a positive v lesion score (v score > 0), including low scores, has been associated with AMR and t-cell mediated rejection (TCMR). (Lefaucheur et al. 2013).

Vascular fibrous intimal thickening

The Banff lesion score, vascular fibrous intimal thickening (cv score) received binary categorization based upon the precedent set by existing literature. Per the Banff criteria, positive cv scores (cv score > 0) are a feature of chronic active AMR. (Roufosse et al. 2018).

Arteriolar hyalinosis

Per the Banff criteria, arteriolar hyalinosis (ah score) is not currently utilized in any diagnostic category. It serves a purely descriptive purpose, as written in the manuscript by Roufosse et al., 2018 (Roufosse et al. 2018). However, the absence of arteriolar hyalinosis (ah score = 0) has been associated with graft loss, likely secondary to underimmunosuppression with CNI-based regimens. [(Einecke, Reeve, and Halloran 2017), (Matos et al. 2016)]. Thus, arteriolar hyalinosis was reduced to a binary classification (presence or absence) and included in the univariate analysis.

4.3.3.6 Categorical breakdown of candidate variables in multivariate analysis

The individual Banff lesion scores, interstitial fibrosis/tubular atrophy (IFTA score), glomerulitis (g score), peritubular capillaritis (ptc score), interstitial inflammation (i score), tubulitis (t score), and transplant glomerulopathy (cg score) were included in the multivariate analysis. The specific rationale for categorizing and including individual lesion scores in the analysis is described below.

Interstitial fibrosis/tubular atrophy

Interstitial fibrosis/tubular atrophy scores (IFTA score) have been categorized into two groups by the existing literature (i.e., 0-1 and 2-3). For this analysis, the category containing scores two and three were separated into distinct categories to provide additional granularity. (Alexandre Loupy et al. 2013).

Microcirculation inflammation

Microcirculation inflammation (MI) is the sum of the Banff lesion scores glomerulitis (g score) and peritubular capillaritis (ptc score). These individual lesion scores are collinear and were therefore grouped together for mitigation. These lesion scores have been categorized into two groups by the existing literature (i.e., 0-2 and > 2 to avoid misclassification). In this analysis, categories were further broken down as follows: 0-2, 3-4, and 5-6. This was done to provide further granularity given the linear association between risk of graft loss and MI score, as well as the shape of the distribution for the lesion scores. (Sis et al. 2012).

Interstitial inflammation and tubulitis

Interstitial inflammation and tubulitis is the sum of the individual Banff lesion scores inflammation (i score) and tubulitis (t score). Scores were grouped into two categories, 0-2 and > 2, to avoid misclassification. These lesion scores were categorized based upon the precedent set by existing literature. (A. Loupy et al. 2017).

Transplant glomerulopathy

Transplant glomerulopathy consisted of the Banff lesion score glomerular basement membrane (GBM) double contour (cg score). This lesion score received binary categorization of presence [≥ 1] or absence [0], based upon the precedent set by existing literature. (A. Loupy et al. 2017).

4.3.3.7 Individual Banff lesion scores excluded from univariate and multivariate analyses

The individual Banff lesion scores, mesangial matrix expansion (mm score) and total inflammation (ti score) were excluded from analyses with rationales described below.

Mesangial matrix expansion

Mesangial matrix expansion (mm score) is "currently not used to reach a diagnostic category and is purely descriptive", per Roufosse et al., 2018. (Roufosse et al. 2018).

Total inflammation

Total inflammation (ti score) lacks sufficient evidence for its association with allograft survival based on current scientific knowledge. In addition, collinearity exists between the ti score and Banff indices, i and ci. (Sis et al. 2010).

4.3.3.8 Missing data

Missing data were anticipated to be negligible based on the results in Loupy et al., 2019 (0.01% in derivation set). These values were confirmed through data exploration of the derivation cohort. If a proportion of missing data were identified as higher than previously reported, then the nature of missingness was evaluated, i.e., Missing at Random (MAR) or Missing Completely at Random (mCAR), and appropriate methodologies were considered, such as removal or imputation.

4.3.3.9 *de novo* mTORi subjects in the derivation dataset for supplementary analyses, including subset for treatment effect analyses

There were additional analyses conducted on a subset of subjects in the derivation dataset who were on CNI-free mTORi-based therapy, sirolimus or everolimus, to better understand the performance of the full and abbreviated iBox Scoring System in different clinical scenarios and subpopulations. Additionally, these subjects were also included in the pseudo trial generation, referred to as the mTORi derivation subset, to support the TLS analysis, as described in Results 6.6.3 (Trial-level surrogacy and treatment effect analyses). Table 16 and Table 17 provide a comparison of baseline characteristics in the derivation cohort between CNI and CNI-free subjects for both iBox Scoring System models.

A review of clinical scenarios in which subjects may be prescribed an mTORi at the time of transplant is summarized below.

CNIs are recommended for initial maintenance immunosuppression by renal transplantation guidelines (Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009). Other medication classes, including the mTORis, have also been studied in the context of initial maintenance immunosuppression. However, mTORis are generally not recommended to be initiated until graft function is established and surgical wounds are healed due to their increased risk of delayed wound healing and occurrence of wound-related complications post-transplantation (Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009).

While CNI-based regimens are most frequently used, they are not appropriate in all clinical circumstances. In a meta-analysis of eight RCTs, no differences in acute rejection (AR), graft survival, or subject survival were found when mTORis were used to replace CNIs (Webster et al. 2006). Advantages of mTORis include antitumor activity and less nephrotoxicity relative to CNIs (Campistol et al. 2011). For subjects with an unacceptable risk of malignancy, or nephrotoxicity, using a mTORi might be advantageous (Weir et al. 2010).

Table 16. Comparison of baseline characteristics in the derivation cohort between CNI and CNI-free subjects for the full iBox Scoring System

	Full iBox Scoring System **n = 3,941‡**		
	CNI **n = 3,590** **TAC** **n = 2,549** **CsA n = 1,041**	**CNI-Free*** **n = 171** **Sirolimus** **n = 109** **Everolimus** **n = 62**	**CNI & mTORi** **n = 68**
Recipient characteristics			
Age (years) (mean, S.D.)	49.7, 13.66	52.8, 13.95	50.7, 14.03
Race (No. %)	NA	NA	NA

Sex (mode, %)	Male, 61%	Male, 61%	Male, 68%
Donor characteristics			
Donor age (mean, S.D.)	51.5, 16.39	55.4, 14.75	53.2, 16.34
Donor type (mode, %)	Deceased, 83%	Deceased, 90%	Deceased, 79%
Transplant baseline characteristics			
Previous kidney transplant (yes, %)	16%	3%	15%
Total No. of HLA-A/B/DR mismatches (mean, % >3)	3.8, 62%	3.8, 64%	3.6, 53%
CIT (hours) (mean)	16.2	17.7	15.3
DSA at time of transplantation (yes, %)	19%	8%	13%
DGF (yes, %)	27%	28%	18%
Induction			
Induction (yes, %)	93%	98%	94%
IL-2Ra (%)	42%	53%	77%
Lymphocyte-depleting agent (%)	58%	47%	23%

ǂ Subjects that do not meet these three regimen categories are not reflected in this Table

* No documented BELA subjects

Table 17. Comparison of baseline characteristics in the derivation cohort between CNI and CNI-free subjects for the abbreviated iBox Scoring System

	Abbreviated iBox Scoring System n = 4,000ǂ		

	**CNI
	n = 3,646
	TAC
	n = 2,581
	CsA
	n = 1,065**
		**CNI-Free*
		n = 174
		Sirolimus
		n = 112
		Everolimus
		n = 62**
			**CNI & mTORi
			n = 68**

Recipient characteristics			

Age (years) (mean, S.D.)	49.7, 13.68	53.0, 13.93	50.7, 14.03
Race (No. %)	NA	NA	NA
Sex (mode, %)	Male, 61%	Male, 61%	Male, 68%
Donor characteristics			
Donor age (years) (mean, S.D.)	51.5, 16.37	55.8, 14.84	53.2, 16.34
Donor type (mode, %)	Deceased, 83%	Deceased, 90%	Deceased, 79%
Transplant baseline characteristics			
Previous kidney transplant (yes, %)	16%	3%	15%
Total No. of HLA-A/B/DR mismatches (mean, % >3)	3.8, 62%	3.8, 64%	3.7, 53%
CIT (hours) (mean)	16.2	17.7	15.3
DSA at time of transplantation (yes, %)	19%	7.5%	13%
DGF (yes, %)	27%	28%	18%
Induction			
Induction (yes, %)	93%	98%	94%
IL-2Ra (%)	42%	52%	77%
Lymphocyte-depleting agent (%)	58%	48%	23%

ł Subjects that do not meet these three regimen categories are not reflected in this Table

* No documented BELA subjects

4.3.4 Qualification validation datasets

The qualification validation dataset consisted of datasets from two clinical transplant centers-Mayo Clinic in Rochester, Minnesota, USA, and Helsinki University Hospital in Helsinki, Finland, and two clinical trials—(BENEFIT RCT) Vincenti et al., 2012; and (BENEFIT-EXT RCT) Medina-Pestana., 2012. Table 18 summarizes the subjects included in the full and abbreviated iBox Scoring System models.

Table 18. Summary of the qualification validation datasets

	Total number of subjects N	Total No. of subjects with sufficient follow-up n, %	No. of subjects with abbreviated iBox Scoring System at one-year post-transplant n, %	No. of subjects with full iBox Scoring System at one-year post-transplant n, %
Mayo Clinic Rochester **2000-2016**	1,618	1,567 (97%)	497 (31%)	483 (30%)
Helsinki University Hospital **2006-2014**	413	413 (100%)	344 (83%)	344 (83%)
BENEFIT RCT **2004-2011**	666	491 (74%)	515 (77%)	416 (62%)
BENEFIT-EXT RCT **2004-2011**	543	408 (75%)	357 (66%)	260 (48%)
Total	3,240	2,879	1,713	1,503

4.3.4.1 Mayo Clinic Rochester

Subject-level data from Mayo Clinic Rochester in Rochester, Minnesota, USA, were used to support this regulatory submission. This dataset consisted of individuals over 18 years of age who received a kidney transplant between 2000-2016. Seventy-nine percent of all transplants were from living donors. The mean recipient age was 50 years at transplant. The majority of transplant recipients were male (56%). The most frequently prescribed induction and IST regimens were lymphocyte depleting (73%) and TAC-based (99.8%), respectively.

Of the 1,618 subjects who are represented in the dataset from Mayo Clinic Rochester, there were 483 and 497 subjects to support the full and abbreviated iBox Scoring Systems, respectively. The proposed components of the iBox Scoring System include:

1. eGFR [SCr]: calculated with SCr and is based on the 4-variable MDRD-186 Study equation; A full and detailed description of the assay and laboratory certification documentation can be found in the Appendix (Revised-Analytical considerations).
2. Measurement of protein excretion into the urine ('proteinuria'): calculated with 24-hour urine (grams per 24 hour) and/or urine albumin-to-creatinine ratio (UACR)

(grams urine albumin per grams urine creatinine); A full and detailed description of the assay and laboratory certification documentation can be found in the Appendix (Revised-Analytical considerations). The 24-hour urine and UACR values from this dataset were converted to UPCR values for use in the iBox Scoring System algorithm. This conversation methodology can be found in Modeling analysis methodologies 5.5.1 (Proteinuria conversions).

3. Histopathological assessment of tissue obtained by renal allograft biopsy, using the most recent Banff scoring criteria at the time the biopsy was reported ('kidney allograft biopsy histopathology'); A full and detailed description of the assay and laboratory certification documentation can be found in the Appendix (Revised-Analytical considerations).
4. Presence of ('DSA'). Additionally, the presence of the DSA was refined into a qualitative binary category based on MFI values. A full and detailed description of the assay and laboratory certification documentation can be found in the Appendix (Revised-Analytical considerations).

4.3.4.2 Helsinki University Hospital

Subject-level data from Helsinki University Hospital in Helsinki, Finland, Europe, were used to support this regulatory submission. This dataset consisted of individuals over 18 years of age who received a kidney transplant between 2006-2014 Ninety-six percent (96%) of all transplants were from deceased donors. The mean recipient age was 52 years at transplant. The majority of transplant recipients were male (65%). Induction therapy was not frequently prescribed (10%). Of the subjects who received induction therapy, all of them were on an Interleukin-2 receptor antagonist (IL-2Ra). All subjects received CNI-based maintenance immunosuppression at baseline, with CsA most frequently prescribed (73%).

Of the 413 subjects who are represented in the dataset from Helsinki University Hospital, there were 344 subjects to support both iBox Scoring System models. The proposed components of the iBox Scoring System include:

1. eGFR [SCr]: calculated with SCr and is based on the 4-variable MDRD-186 Study equation; A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
2. Measurement of protein excretion into the urine ('proteinuria'): calculated by dipstick proteinuria (urinary total protein; non-quantitative); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations). The dipstick proteinuria results from this dataset were converted to UPCR values for use in the iBox Scoring System algorithm. This conversation methodology can be found in Modeling analysis methodologies 5.5.1 (Proteinuria conversions).
3. Histopathological assessment of tissue obtained by renal allograft biopsy, using the most recent Banff scoring criteria at the time the biopsy was reported ('kidney allograft biopsy histopathology'); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
4. Presence of ('DSA'). Additionally, the presence of the DSA was refined into a qualitative binary category based on MFI values. A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).

4.3.4.3 BENEFIT and BENEFIT-EXT RCTs

Data included in Vincenti et al., 2012 (BENEFIT RCT) and Medina-Pestana., 2012 (BENEFIT-EXT RCT) are the two studies that led to the FDA approval of Nulojix™ (BELA) in 2011, and is indicated for prophylaxis of organ rejection in adult subjects receiving a kidney transplant. BELA, a selective T cell co-stimulation blocker, was shown to provide long-term immunosuppression, better preservation of kidney function, improved cardiovascular/metabolic risk profiles, and less toxicity compared with CsA. Both of these RCTs were conducted in *de novo* kidney transplant recipients.

BENEFIT was a three-year, randomized, active-controlled, parallel-group, multicenter (100 centers worldwide) phase III study. In this study, 666 participants receiving a living or standard criteria deceased donor kidney transplant were randomized at centers in North America, South America, Europe, Australia, South Africa, and Asia. The objectives of this study were to assess belatacept-based immunosuppression compared with CsA, a CNI-based immunosuppression, on three coprimary outcomes at 12 months after kidney transplantation. These primary outcomes included: (a) the percent of participants surviving with a functioning graft by month 12 [time frame: Day 1 to Month 12], (b) the percent of participants with a composite of measured glomerular filtration rate (mGFR) < 60 mL/min/1.73 m^2 at month 12 or with a decrease in mGFR ≥ 10 mL/min/1.73m^2 from month 3 to month 12, and (c) the percent of participants experiencing biopsy-confirmed AR post-transplant by month 12 [time frame: day 1 to month 12]. (Vincenti et al. 2012).

Of the 666 randomized participants who are represented in the BENEFIT RCT dataset, there were 416 and 515 subjects to support the full and abbreviated iBox Scoring Systems, respectively. The proposed components of the iBox Scoring System include:

1. eGFR [SCr]: calculated with SCr and is based on the 4-variable MDRD-186 Study equation; A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
2. Measurement of protein excretion into the urine ('proteinuria'): calculated by dipstick proteinuria (urinary total protein; non-quantitative); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations). The dipstick proteinuria results from this dataset were converted to UPCR values for use in the iBox Scoring System algorithm. This conversation methodology can be found in Modeling analysis methodologies 5.5.1 (Proteinuria conversions).
3. Histopathological assessment of tissue obtained by renal allograft biopsy, using the most recent Banff scoring criteria at the time the biopsy was reported ('kidney allograft biopsy histopathology'); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
4. Presence of ('DSA'). Additionally, the presence of the DSA was refined into a qualitative binary category based on MFI values. A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).

BENEFIT-EXT was a three-year, randomized, active-controlled, parallel-group, multicenter (79 centers worldwide) phase III study. This study included 543 participants receiving a kidney from a deceased donor meeting UNOS extended criteria for donation living with a kidney transplant from an ECD, or those donated following cardiac death (DCD), or with an estimated CIT > 24 hours in duration, were randomized at centers in North America, South America, Europe, South Australia, and South Africa. This study's objectives were to compare belatacept immunosuppression with CsA, a CNI-based immunosuppression, on two coprimary outcomes at 12 months after kidney transplantation. These primary outcomes included: (a) the percent of participants surviving with a functioning graft by month 12 [time frame: Day 1 to Month 12] and (b) the percent of participants with a composite of mGFR < 60 mL/min/1.73 m^2 at month 12 or with a decrease in mGFR ≥ 10 mL/min/1.73m^2 from month 3 to month 12. (Durrbach et al. 2010).

Of the 543 randomized participants who are represented in the BENEFIT-EXT RCT dataset, there were 260 and 357 subjects to support the full and abbreviated iBox Scoring Systems, respectively. The proposed components of the iBox Scoring System include:

1. eGFR [SCr]: calculated with SCr and is based on the 4-variable MDRD-186 Study equation; A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
2. Measurement of protein excretion into the urine ('proteinuria'): calculated by dipstick proteinuria (urinary total protein; non-quantitative); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations). The dipstick proteinuria results from this dataset were converted to UPCR values for use in the iBox Scoring System algorithm. This conversation methodology can be found in Modeling analysis methodologies 5.5.1 (Proteinuria conversions).
3. Histopathologic assessment of tissue obtained by renal allograft biopsy, using the most recent Banff scoring criteria at the time the biopsy was reported ('kidney allograft biopsy histopathology'); A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).
4. Presence of ('DSA'). Additionally, the presence of the DSA was refined into a qualitative binary category based on MFI values. A full and detailed description of the assay and laboratory certification documentation can be found in Appendix (Revised-Analytical considerations).

As BENEFIT and BENEFIT-EXT were the only two RCTs included in this Briefing Dossier, pseudo trial generation using these two datasets was completed to support TLS analysis, as described in Results 6.6.3 (Trial-level surrogacy analysis).

4.3.5 Loupy et al., 2019 European validation cohort

The European centers in Loupy et al., 2019 that were part of the European validation cohort included 2,129 subjects over 18 years old prospectively enrolled at the time of transplantation from a living or deceased donor between 2002 and 2014 in three European centers: Hôpital Hôtel Dieu, Nantes, France (n = 632); Hospices Civils, Lyon, France (n = 608); and the University Hospitals, Leuven, Belgium (n = 889). The baseline characteristics of this European cohort is summarized in Table 19.

Table 19. Baseline characteristics of the European validation centers in Loupy et al., 2019

		Nantes (France) (n=632)		Lyon (France) (n=608)		Leuven (Belgium) (n=889)
	N		**N**			

Recipient characteristics						
Age (years), mean (SD)	632	50.4 (13.57)	608	46.6 (13.28)	889	53.4 (13.30)
Gender male, No. (%)	632	404 (63.92)	608	386 (63.49)	889	543 (61.08)
ESRD causes	632		608		889	
Glomerulonephritis, No. (%)		179 (28.32)		151 (24.84)		254 (28.57)
Diabetes, No. (%)		55 (8.70)		188 (30.92)		73 (8.21)
Vascular, No. (%)		53 (8.39)		49 (8.06)		37 (4.16)
Other, No. (%)		345 (54.59)		220 (36.18)		525 (59.06)
Donor characteristics						
Age (years), mean (SD)	632	53.1 (14.99)	603	44.1 (16.55)	887	47.6 (14.89)
Male gender, No. (%)	631	354 (56.10)	605	395 (65.29)	888	476 (53.60)
Hypertension, No. (%)	620	185 (29.84)	607	101 (16.64)	649	164 (25.27)
Diabetes mellitus, No. (%)	481	36 (7.48)	343	11 (3.21)	889	0
Creatinine > 132 µmol/L, No. (%)	631	80(12.68)	605	95 (15.70)	700	18 (2.57)
Donor type						
Deceased donor, No. (%)	632	576 (91.14)	608	564 (92.76)	889	834 (93.81)
Death from cerebrovascular disease, No. (%)	576	323 (56.08)	564	257 (45.57)	834	413 (49.52)
ECD, No. (%)	574	248 (43.21)	608	142 (23.36)	828	238 (28.74)
Transplant baseline characteristics						

Prior kidney transplant, No. (%)	632	101 (15.98)	608	94 (15.46)	889	127 (14.29)
Cold ischaemia time (hours), mean (SD)	632	18.75 (9.39)	599	13.68 (5.85)	862	14.37 (5.44)
DGF[a], No. (%)	630	213 (33.81)	608	102 (16.78)	889	161 (18.11)
HLA-A/B/DR mismatch, mean (SD), number	632	3.28 (1.36)	608	3.58 (1.35)	843	2.75 (1.34)

Abbreviations: ESRD: end-stage renal disease; HLA: human leucocyte antigen. [a]DGF was defined as the use of dialysis in the first postoperative week

4.3.6 Loupy et al., 2019 External validation in three randomized controlled trials

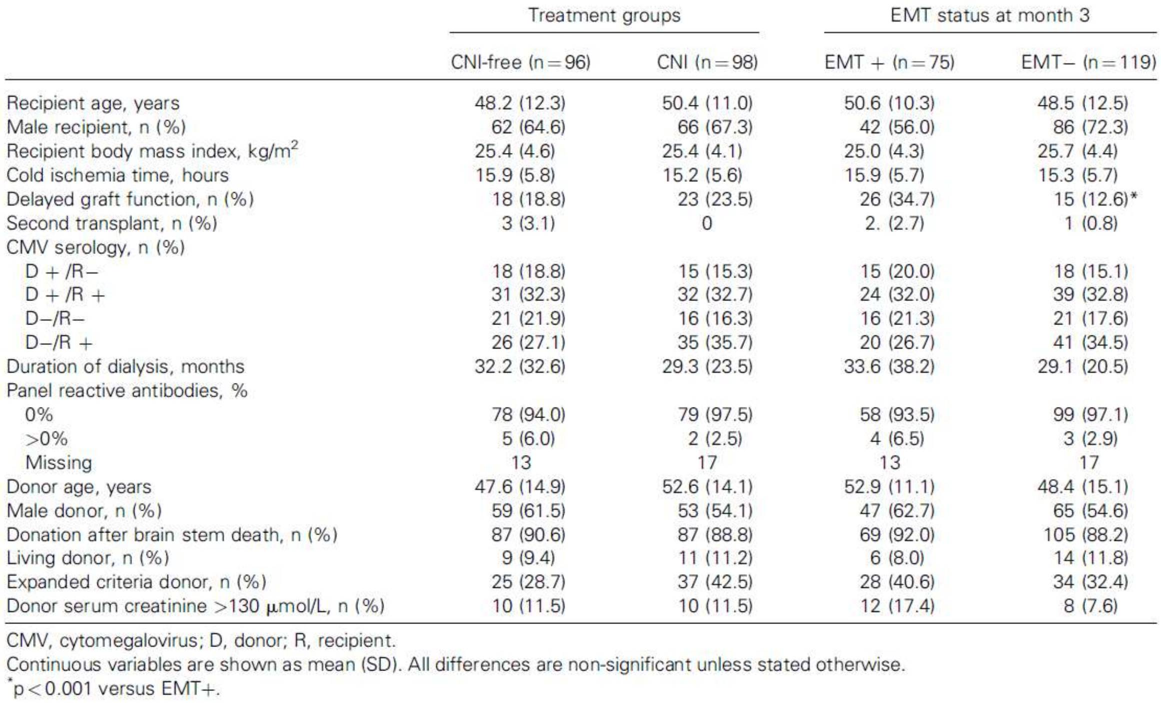
Additional external validation in Loupy et al., 2019 was conducted in three RCTs, summarized in Table 20. In particular, CERTITEM RCT was a *de novo* phase III trial. The baseline characteristics of the subjects in the CERTITEM RCT are summarized in (Rostaing et al. 2015).

Table 20. External validation in three RCTs as described in Loupy et al., 2019

STUDY	Trial #ID	Design	Clinical scenario	Target population	(n)	Time post-transplant of iBox risk score evaluation	Follow-up time post-transplant
CERTITEM (1)	**NCT 01079143**	Prospective, Randomized, open-label, multicenter trial	ISD minimization	Recipients of renal transplants from a living or deceased donor	194	Median: 0.94 years Interquartile range (IQR) (0.92-0.98)	Median: 6.62 years IQR (2.82-7.34)
RITUX ERAH (2)	**Eudra CT 2007-003213-13**	Prospective, Randomized, multicenter, double-blind, placebo-controlled trial	Treatment of ABMR (pre-existing DSA)	Recipients of renal transplants from a living or deceased donor with diagnosis of acute ABMR.	38	Median: 0.74 years IQR (0.53-1.10)	Median: 6.63 years IQR (4.03-7.69)
BORTEJECT (3)	**NCT 01873157**	Prospective, Randomized, placebo-controlled, double-blind, single-center trial	Treatment of ABMR (De novo DSA)	Recipients of renal transplants from a living or deceased donor with post-transplant de novo DSA detection	44	Median: 6.61 years IQR (4.04-15.41)	Median: 7.75 years IQR (5.32-16.41)

(1) Rostaing, L., et al. "Fibrosis progression according to epithelial-mesenchymal transition profile: a randomized trial of everolimus versus CsA." American Journal of Transplantation 15.5 (2015): 1303-1312. (2) Sautenet, B., et al. "One-year results of the effects of rituximab on acute antibody-mediated rejection in renal transplantation: RITUX ERAH, a multicenter double-blind randomized placebo-controlled trial." Transplantation 100.2 (2016): 391-399(3) Eskandary, Farsad, et al. "A Randomized Trial of Bortezomib in Late Antibody-Mediated Kidney Transplant Rejection." Journal of the American Society of Nephrology (2017): ASN-2017070818.

Table 21. Baseline characteristics of subjects in the CERTITEM RCT



	Treatment groups		EMT status at month 3	
	CNI-free (n = 96)	CNI (n = 98)	EMT + (n = 75)	EMT− (n = 119)
Recipient age, years	48.2 (12.3)	50.4 (11.0)	50.6 (10.3)	48.5 (12.5)
Male recipient, n (%)	62 (64.6)	66 (67.3)	42 (56.0)	86 (72.3)
Recipient body mass index, kg/m^2	25.4 (4.6)	25.4 (4.1)	25.0 (4.3)	25.7 (4.4)
Cold ischemia time, hours	15.9 (5.8)	15.2 (5.6)	15.9 (5.7)	15.3 (5.7)
Delayed graft function, n (%)	18 (18.8)	23 (23.5)	26 (34.7)	15 (12.6)*
Second transplant, n (%)	3 (3.1)	0	2. (2.7)	1 (0.8)
CMV serology, n (%)				
D + /R−	18 (18.8)	15 (15.3)	15 (20.0)	18 (15.1)
D + /R +	31 (32.3)	32 (32.7)	24 (32.0)	39 (32.8)
D−/R−	21 (21.9)	16 (16.3)	16 (21.3)	21 (17.6)
D−/R +	26 (27.1)	35 (35.7)	20 (26.7)	41 (34.5)
Duration of dialysis, months	32.2 (32.6)	29.3 (23.5)	33.6 (38.2)	29.1 (20.5)
Panel reactive antibodies, %				
0%	78 (94.0)	79 (97.5)	58 (93.5)	99 (97.1)
>0%	5 (6.0)	2 (2.5)	4 (6.5)	3 (2.9)
Missing	13	17	13	17
Donor age, years	47.6 (14.9)	52.6 (14.1)	52.9 (11.1)	48.4 (15.1)
Male donor, n (%)	59 (61.5)	53 (54.1)	47 (62.7)	65 (54.6)
Donation after brain stem death, n (%)	87 (90.6)	87 (88.8)	69 (92.0)	105 (88.2)
Living donor, n (%)	9 (9.4)	11 (11.2)	6 (8.0)	14 (11.8)
Expanded criteria donor, n (%)	25 (28.7)	37 (42.5)	28 (40.6)	34 (32.4)
Donor serum creatinine >130 μmol/L, n (%)	10 (11.5)	10 (11.5)	12 (17.4)	8 (7.6)

CMV, cytomegalovirus; D, donor; R, recipient.

Continuous variables are shown as mean (SD). All differences are non-significant unless stated otherwise.

*$p < 0.001$ versus EMT+.

4.3.7 Dataset for supplementary analyses

4.3.7.1 Charité – Universitätsmedizin Berlin cohort

The patient-level data used to support the dipstick to imputed-proteinuria algorithm is from a cohort of 1,387 German subjects from Charité – Universitätsmedizin Berlin with dipstick proteinuria results and UPCR values. An association of these two proteinuria measurements was assessed in order to impute a UPCR value in place of dipstick proteinuria for use in the full and abbreviated iBox Scoring System algorithm. More detail can be found in Modeling analysis methodologies 5.5.3.1 (Imputation).

4.3.8 Alignment of qualification datasets

The following criteria were evaluated for aligning data from clinical transplant centers and clinical trials to support the proposed COU, described in Methods 4.2 (Context-of-use):

- **One-year iBox Scoring System risk evaluation**: One-year post-transplant refers to all relevant features necessary to carry out a full and abbreviated iBox Scoring System evaluation for each transplant recipient at one-year post-transplant ± 28 days. A two-tailed window of 28 days was applied to the one-year post-transplant iBox evaluation in an effort to reflect the practicality of collecting the individual components of the full and abbreviated iBox in the context of a clinical trial.

- **Censoring cutoff:** Data were censored at five years + 28 days (i.e. everyone who had a functioning graft at five years + 28 days was considered as having a functional graft). A censoring cutoff at five years post-transplantation with a one-tail 28 day upper bound was applied in an effort to accommodate events in the historical data that occur proximally to a five-year cut-off.

- **Binary cutoff for presence of DSA**: The cut-off for the presence of DSA varied between laboratories. A binary cut-off of 1,400 was used for this submission based on literature precedent allowing for the inclusion and alignment of datasets.

- **Maintenance IST regimens:** Maintenance IST regimen information is limited to the drug name at baseline, defined as at time of transplant. It is understood that therapy changes, including the discontinuation of initial IST regimens, occur in clinical trials and clinical practice due to intolerance, adverse effects and/or lack of efficacy. Seven categories of drug classes were created, as described in the following section, Methods 4.3.8.1 (Therapeutics across qualification datasets).

- **Imputed-UPCR values for use in iBox calculation**: As UPCR was not ubiquitously assessed across all the qualification validation datasets, conversion algorithms were used when needed. These conversions were carried out to allow the evaluation of historical data. The TTC recommends the use of UPCR proteinuria measurement in prospective studies. The proteinuria conversions were as follows, with additional details in the Modeling analysis methodologies 5.5.1 (Proteinuria conversions):

    - 24-hour proteinuria to UPCR: This conversion was used for a portion of the subjects from Mayo Clinic Rochester.

    - Urine albumin-to-creatinine ratio (UACR) to urine protein-to-creatinine ratio (UPCR): This conversion was used for a portion of the subjects from Mayo Clinic Rochester.

    - Dipstick to UPCR: This conversion was used for the two BMS RCTs (BENEFIT and BENEFIT-EXT), and subjects from Helsinki University Hospital.

- **Reclassification of outcomes for transplant recipients based on eGFR analysis:** Differential definitions of graft loss were used across the qualification datasets. To harmonize the definition of graft loss used across the five historical datasets, TTC examined the relevant literature available for defining graft loss in the context of clinical trials. Reclassification of outcomes across the qualification datasets is described in the following section, Methods 4.3.8.2 (Reclassification of outcomes for transplant recipients based on eGFR analysis).

- **Accounting for death/graft loss within the first year of transplant:** The full and abbreviated iBox Scoring System evaluation occurs at one-year post-transplant, so transplant recipients who die or lose their graft during the first-year post-transplant cannot have an iBox Scoring System evaluation. In order to evaluate the full and abbreviated iBox Scoring System's performance as an intermediate endpoint, recipient s who die or lose their graft during the first year post-transplant must be included to avoid survivor bias. The TLS analysis evaluates the full and abbreviated iBox Scoring System as an endpoint by relating one-year iBox score differences in treatment arms to five-year death-graft survival; to avoid survivor bias, recipients who died or lost their graft in the first year of transplant were given an imputed iBox value corresponding to the worst-case scenario.

- **TLS**: Specific to TLS analysis, all graft losses occurring before data cut-off date were included in the analysis.

4.3.8.1 Therapeutics across qualification datasets

Therapeutics across datasets

Induction therapies and baseline maintenance ISTs were also assessed to support analyses for both full and abbreviated iBox Scoring System models. A comprehensive data exploration effort was undertaken to assess the level of information present in all the qualification datasets related to the use of induction and baseline maintenance ISTs, as described in Tables 22-25 below.

Maintenance immunosuppressants - Mechanisms of Action

The iBox Scoring System was tested in populations receiving immunosuppressants with various MOA. Currently, three classes of drugs are approved by Regulatory Authorities in the US and EU for maintenance immunosuppression. These include calcineurin inhibitors (CNIs) (i.e., tacrolimus and cyclosporine), mammalian target of rapamycin inhibitors (mTORi) (i.e., everolimus, sirolimus), and T cell costimulation inhibitors (i.e., belatacept). Most kidney transplant patients in the US and EU currently receive a CNI-based regimen with selected patients and centers using mTORi and/or belatacept. This qualification submission includes data on the performance of the iBox Scoring System with all three MOA. Consistent with the current standard of care, most of the patients in this submission received a CNI. C-Path believes the number of therapeutics covered in the qualification validation datasets encompasses the breadth of therapeutic combinations seen in clinical practice and clinical trials to support general surrogacy of the iBox Scoring System within the defined COU.

Induction therapies

The definition of induction therapies was consistent across qualification datasets and is consistent with the definition from the Organ Procurement and Transplantation Network (OPTN): "medications given for a short finite period in the perioperative period for the purpose of preventing AR. Though these drugs may be continued after discharge for the first 30 days after transplant, it will not be used long-term for immunosuppressive maintenance." [https://optn.transplant.hrsa.gov/resources/glossary/, KDIGO 2009]

Contemporary induction trials have largely compared lymphocyte non-depleting versus lymphocyte-depleting biological therapies. The choice of which class of biologics to use (or even in combination) depends on the transplant recipient's immunological risk profile [(M. A. Lim, Kohli, and Bloom 2017), (Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009)]. Data from the United States Renal Data System (USRDS) shows that 85% of transplant recipients in the United States currently receive induction with rabbit antithymocyte globulin (rATG) followed by alemtuzumab and basiliximab (Matas et al. 2015) Scientific Registry of Transplant Recipients (SRTR) 2019 annual data report in kidney transplantation shows induction use is up to 91.9% in kidney transplant recipients.

To best leverage the induction medication information, six categories of drug classes were created, as shown in Table 22 and Table 23:

Table 22. Induction therapies across qualification datasets for full iBox Scoring System

	Lymphocyte Nondepleting Agent (IL-2Ra)	Lymphocyte -Depleting Agent Polyclonal Antibodies	Lymphocyte -Depleting Agent Monoclonal Antibody	Other*	No induction	Missing induction information
Loupy et al., 2019 derivation	1,621	2,069	0	0	0	251
Mayo Clinic Rochester	107	273	80	22	0	1
Helsinki University Hospital	33	0	0	0	311	0
BENEFIT RCT	416	0	0	0	0	0
BENEFIT-EXT RCT	260	0	0	0	0	0

***** Examples include regimens containing methylprednisolone, Intravenous immunoglobulin (IVIG), plasmapheresis, or a combination of lymphocyte-depleting polyclonal antibodies and lymphocyte-depleting monoclonal antibody

Table 23. Induction therapies across qualification datasets for abbreviated iBox Scoring System

	Lymphocyte Nondepleting Agent (IL-2Ra)	Lymphocyte -Depleting Agent Polyclonal Antibodies	Lymphocyte -Depleting Agent Monoclonal Antibody	Other *	No induction	Missing induction information
Loupy et al., 2019 derivation	1,643	2,104	0	0	0	253
Mayo Clinic Rochester	112	278	81	24	0	2
Helsinki University Hospital	33	0	0	0	311	0
BENEFIT RCT	515	0	0	0	0	0

BENEFIT-EXT RCT	357	0	0	0	0	0

***** Examples include regimens containing methylprednisolone, IVIG, plasmapheresis, or a combination of lymphocyte-depleting polyclonal antibodies and lymphocyte-depleting monoclonal antibody

Baseline maintenance immunosuppressive therapies

The definition of maintenance IST was consistent across qualification datasets as follows: medications prescribed at the time of transplant (i.e., intent-to-treat) for the purpose of preventing AR and safely preserving allograft function long-term (i.e., for the life of the allograft). [https://optn.transplant.hrsa.gov/resources/glossary/, KDIGO 2009]

Maintenance regimens typically include two to three different agents to achieve adequate immunosuppression while minimizing the toxicity associated with individual agents [KDIGO 2009]. Data from USRDS shows most transplant recipients in the United States currently receive TAC-based therapy (in combination with mycophenolate and/or steroids, or alone) [SRTR/OPTN 2019 Annual Report, Figure KI 82]. As with induction therapies, the choice of which class of maintenance IST (or even in combination) depends on the transplant recipient's immunological risk profile [(M. A. Lim, Kohli, and Bloom 2017; Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009)].

To best leverage the maintenance medication information, seven categories of drug classes were created, as shown in Table 24 and Table 25:

Table 24. Baseline maintenance therapies across qualification datasets for full iBox Scoring System

	CNI-based (TAC, CsA)	mTORi	Selective T cell costimulation blocker (BELA)	Other*	No baseline maintenance IST	Missing baseline maintenance IST	CNI + mTORi
Loupy et al., 2019 derivation	3590	171	0	112	0	0	68
Mayo Clinic Rochester	481	0	0	2	0	0	0
Helsinki University Hospital	344	0	0	0	0	0	0
BENEFIT RCT	135	0	281	0	0	0	0
BENEFIT-EXT RCT	85	0	175	0	0	0	0

***** Including prednisone and mycophenolate

Table 25. Baseline maintenance therapies across qualification datasets for abbreviated iBox Scoring System

	CNI-based (TAC, CsA)	mTORi	Selective T cell costimulation blocker (BELA)	Other *	No baseline maintenance IST	Missing baseline maintenance IST	CNI + mTORi
Loupy et al., 2019 derivation	3646	174	0	112	0	0	68
Mayo Clinic Rochester	495	0	0	2	0	0	0
Helsinki University Hospital	344	0	0	0	0	0	0
BENEFIT RCT	169	0	346	0	0	0	0
BENEFIT-EXT RCT	116	0	241	0	0	0	0

***** Including prednisone and mycophenolate

4.3.8.2 Reclassification of outcomes for transplant recipients based on eGFR analysis in qualification datasets

Differential definitions of graft loss were used across the qualification datasets. To harmonize the definition of graft loss used across the five historical datasets, TTC examined the relevant literature available for defining graft loss in the context of clinical trials.

The publication titled, "International consensus definitions of clinical trial outcomes for kidney failure: 2020" was utilized and the most stringent criteria for interpreting the sustained low eGFR and sustained percent decline in eGFR were utilized. (Levin et al. 2020). The two strategies for reclassification of graft loss are described below:

Strategy 1 – Sustained low eGFR: Recipients with the following criteria needed to be met for reclassification of graft loss under strategy one:

- An eGFR value less than 15 ml/min/1.73m^2 at any time point starting at one-year post-transplant and all subsequent time points.
- No documented graft loss or death with a functioning graft at any point during study follow-up.

Strategy 2 – Sustained low eGFR and sustained percent decline in eGFR: The following criteria needed to be met for reclassification of graft loss under strategy two:

- An eGFR value at any time point beyond one-year post-transplant and all subsequent time points that showed at least fifty-seven percent decrease relative to the one-year eGFR value. A fifty-seven percent decline in eGFR approximately corresponds to a doubling of SCr and is the most well-established of these (putative) surrogates. (Levin et al. 2020).

- An eGFR value less than 25 mL/min/1.73m^2.

- No documented graft loss or death with a functioning graft at any point during study follow-up.

The TTC reviewed the qualification datasets for potential reclassification of subjects using the two strategies outlined above. The derivation dataset from the PTG did not have longitudinal eGFR data to review and therefore were excluded from potential reclassification. The two BMS studies, Mayo Clinic Rochester, and Helsinki University Hospital were reviewed for potential reclassification of graft loss. The following describes the reclassification for each of the datasets considered:

BENEFIT RCT: From the 666 subjects who were randomized in the BENEFIT RCT dataset, there were 21 subjects with a last recorded eGFR value of less than 25 ml/min/1.73m^2. Of these 21 subjects, there were ten subjects who met one or both reclassification strategies. Of these ten subjects, there were four subjects who met the COU (n = 3 for full iBox Scoring System, n = 1 for abbreviated iBox Scoring System). These four subjects were reclassified as a graft loss in the iBox Scoring System analyses. A summary of these findings is described in Figure 2.

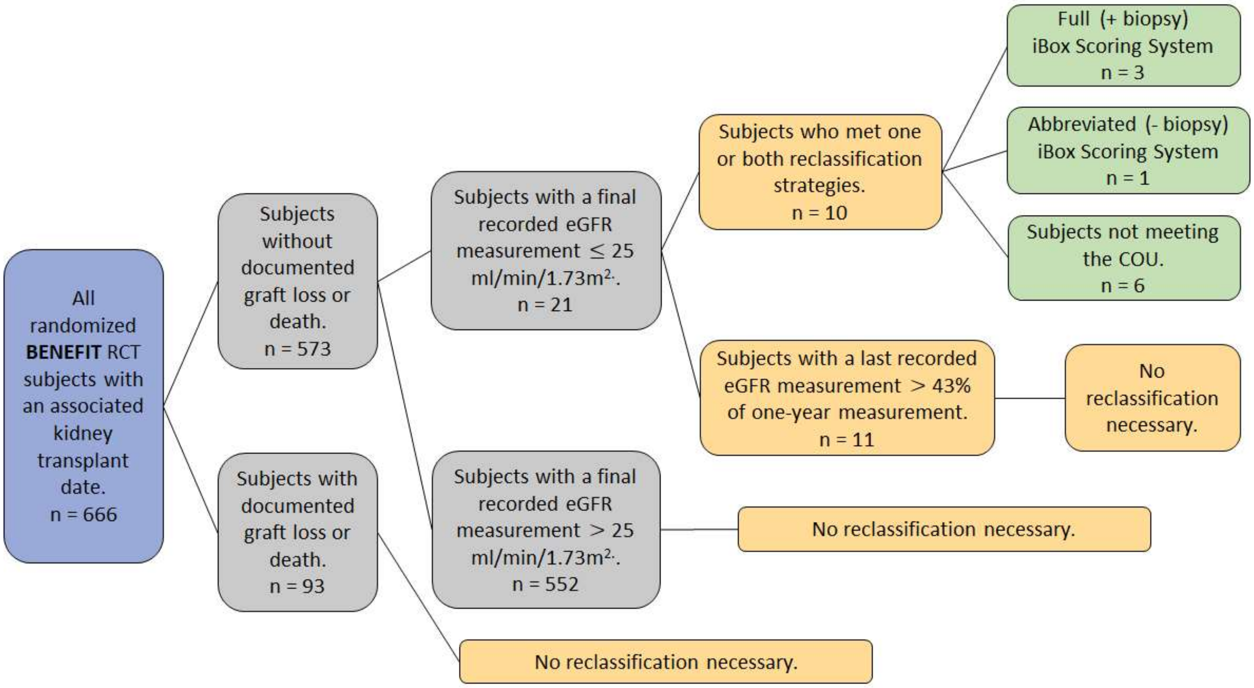

All randomized BENEFIT RCT subjects with an associated kidney transplant date. n = 666
Subjects without documented graft loss or death. n = 573
Subjects with documented graft loss or death. n = 93
Subjects with a final recorded eGFR measurement ≤ 25 ml/min/1.73m^2. n = 21
Subjects with a final recorded eGFR measurement > 25 ml/min/1.73m^2. n = 552
No reclassification necessary.
Subjects who met one or both reclassification strategies. n = 10
Subjects with a last recorded eGFR measurement > 43% of one-year measurement. n = 11
No reclassification necessary.
No reclassification necessary.
Full (+ biopsy) iBox Scoring System n = 3
Abbreviated (- biopsy) iBox Scoring System n = 1
Subjects not meeting the COU. n = 6

Figure 2. Flow chart describing reclassification of graft loss in the BENEFIT RCT.

BENEFIT-EXT RCT: From the 543 subjects who were randomized in the BENEFIT-EXT RCT dataset, there were 42 subjects with a last recorded eGFR value of less than 25 ml/min/1.73m^2. Of these 42 subjects, there were 18 subjects (including three subjects with a PNF graft) who met strategy 1 and/or strategy 2. Of these 18 subjects, there were seven subjects who met the COU (n = 3 for full iBox Scoring System, n = 4 for abbreviated iBox Scoring System). These seven subjects were reclassified as a graft loss in the iBox analyses. A summary of these findings is described in Figure 3.

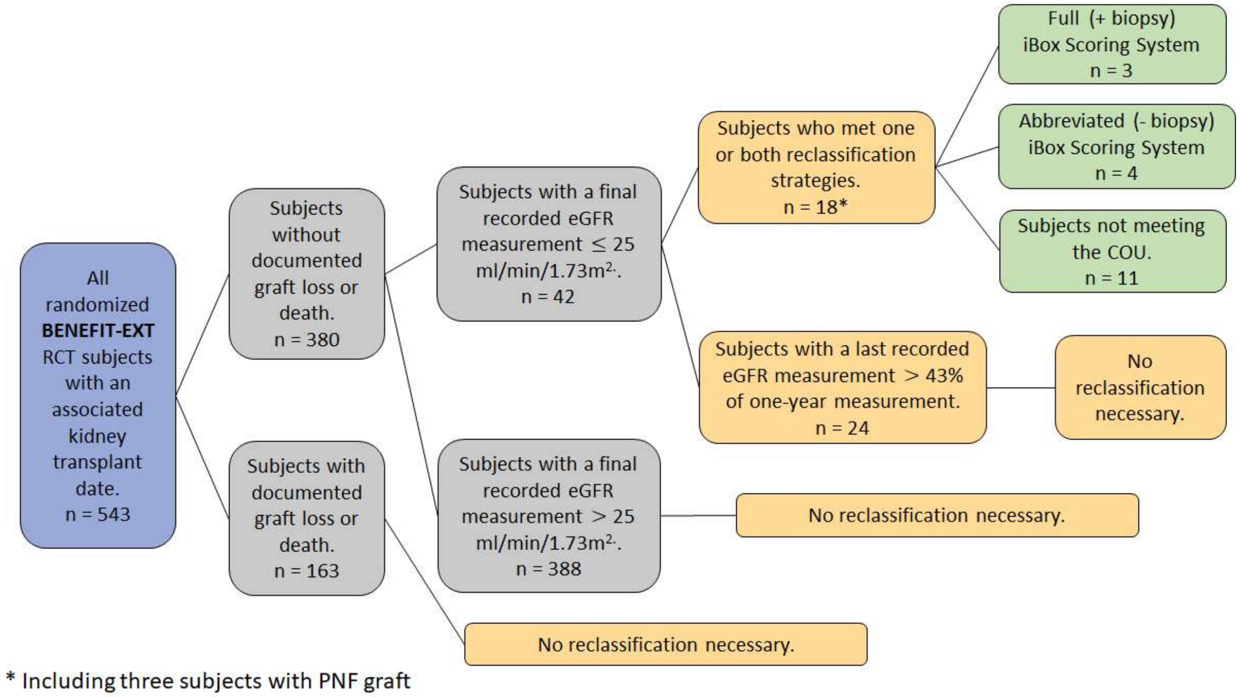

All randomized BENEFIT-EXT RCT subjects with an associated kidney transplant date. n = 543
Subjects without documented graft loss or death. n = 380
Subjects with documented graft loss or death. n = 163
Subjects with a final recorded eGFR measurement ≤ 25 ml/min/1.73m2.. n = 42
Subjects with a final recorded eGFR measurement > 25 ml/min/1.73m2.. n = 388
No reclassification necessary.
Subjects who met one or both reclassification strategies. n = 18*
Subjects with a last recorded eGFR measurement > 43% of one-year measurement. n = 24
No reclassification necessary.
Full (+ biopsy) iBox Scoring System n = 3
Abbreviated (- biopsy) iBox Scoring System n = 4
Subjects not meeting the COU. n = 11
No reclassification necessary.
* Including three subjects with PNF graft

Figure 3. Flow chart describing reclassification of graft loss in the BENEFIT-EXT RCT.

Helsinki University Hospital:

From the 413 transplant recipients in the Helsinki University Hospital dataset, there were 14 recipients with a last recorded eGFR value of less than 25 ml/min/1.73m2. Of these 14 recipients, there were 3 recipients who met strategy 1 and/or strategy 2. All three of these recipients met the COU for the full iBox Scoring System. A summary of these findings is described in Figure 4.

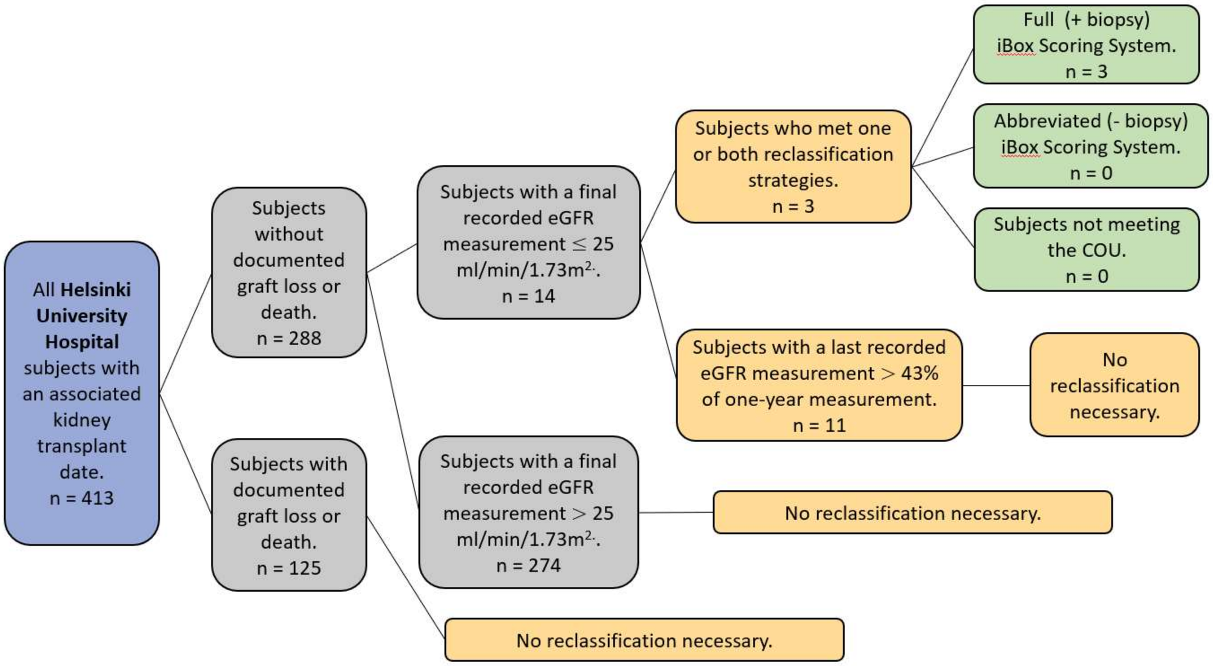

All Helsinki University Hospital subjects with an associated kidney transplant date. n = 413
Subjects without documented graft loss or death. n = 288
Subjects with documented graft loss or death. n = 125
Subjects with a final recorded eGFR measurement ≤ 25 ml/min/1.73m2.. n = 14
Subjects with a final recorded eGFR measurement > 25 ml/min/1.73m2.. n = 274
No reclassification necessary.
Subjects who met one or both reclassification strategies. n = 3
Subjects with a last recorded eGFR measurement > 43% of one-year measurement. n = 11
No reclassification necessary.
No reclassification necessary.
Full (+ biopsy) iBox Scoring System. n = 3
Abbreviated (- biopsy) iBox Scoring System. n = 0
Subjects not meeting the COU. n = 0

Figure 4. Flow chart describing reclassification of graft loss in the Helsinki University Hospital dataset.

Mayo Clinic Rochester

There were no transplant recipients in the Mayo Clinic Rochester dataset that met reclassification criteria.

4.4 Data exclusions

Criteria for data exclusion:

- Individuals without necessary features (i.e., eGFR, proteinuria, DSA, and kidney allograft biopsy histopathology) to calculate an iBox score at one-year ± 28 days post-transplant.

- Transplant recipients with PNF. (Levin et al. 2020).

- Individuals with early graft loss events (before first transplant anniversary) but had one-year full and abbreviated iBox Scoring System evaluations. This reflects two full iBox Scoring System subjects from BENEFIT RCT and three subjects (two full iBox Scoring System and one abbreviated iBox Scoring System) from BENEFIT-EXT RCT. There were no subjects with early loss graft events from Mayo Clinic Rochester or Helsinki University Hospital.

- Specific for validation, subjects in the BMS BENEFIT and BENEFIT-EXT RCTs who were discontinued from their study medication before reaching their one-year post-transplant risk evaluation.

Extreme observations of the dependent variables were double-checked for entry errors. If an entry error was confirmed (i.e., variable cannot take on such value), the observation was removed and a recheck for extreme observations was conducted. This affected minimal numbers of data points. If an entry error could not be ascertained, the observation was kept in the dataset. Lost to follow-up subjects were right-censored.

4.5 Summary of the final qualification datasets for both iBox Scoring System models

Table 26-28 summarize the final qualification datasets for both full and abbreviated iBox Scoring System models and the associated five-year outcomes by Kaplan-Meier (KM) estimates of graft survival. KM estimates give the fraction of survivors while accounting for censoring and are commonly used in survival data (Collett 2015). The first two Tables (Table 26 and Table 27) show both the number of subjects in each dataset with one-year evaluations for both versions of the iBox Scoring System; the numbers are lower for the full version because some subjects are missing biopsy information. Table 28 shows the KM estimates of transplant recipient survival, death-censored allograft survival, and overall graft survival by the end of five years. The BENEFIT-EXT RCT exhibits lower survival rates than the other datasets, which is expected due to its exclusive use of transplants coming from extended criteria donors.

Table 26. Qualification derivation dataset to support full and abbreviated iBox Scoring System models

Dataset	Full iBox Scoring System	Abbreviated iBox Scoring System
Loupy et al., 2019 derivation	**Number of subjects**	
	n =3,941	n = 4,000

Table 27. Qualification validation datasets to support full and abbreviated iBox Scoring System models

Dataset	Full iBox Scoring System	Abbreviated iBox Scoring System
	Number of subjects	
Mayo Clinic Rochester	n = 483	n = 497
Helsinki University Hospital	n = 344	n = 344
BENEFIT RCT	n = 416	n = 515
BENEFIT-EXT RCT	n = 260	n = 357

Table 28. Five-year outcomes by Kaplan-Meier estimates across qualification datasets

Qualification derivation dataset	Qualification validation dataset

	Loupy et al., 2019	Mayo Clinic Rochester	Helsinki University Hospital	BENEFIT RCT	BENEFIT-EXT RCT
Subject survival probability	0.92	0.96	0.92	0.93	0.85
Death-censored allograft survival probability	0.90	0.95	0.94	0.93	0.84
Overall graft survival probability	0.83	0.91	0.86	0.86	0.71

5 MODELING ANALYSIS METHODOLOGIES

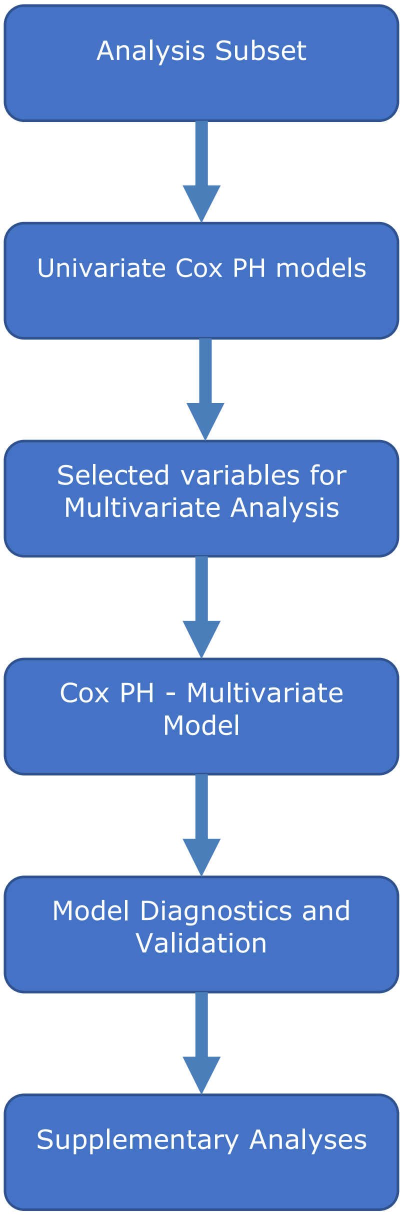

Analysis Subset
Univariate Cox PH models
Selected variables for Multivariate Analysis
Cox PH - Multivariate Model
Model Diagnostics and Validation
Supplementary Analyses

Figure 5. Modeling development workflow.

5.1 Prior knowledge

It is well established in the literature that individual markers of kidney transplant health, when used alone, are insufficient to predict long-term outcomes with acceptable accuracy. Thus, significant prior efforts have been made to develop composite scoring systems that are better able to predict long-term allograft survival. This effort is demonstrated by a 2017 meta-analysis that reviewed risk prediction models for graft failure (generally defined as dialysis, re-transplantation, or death-censored allograft survival) in kidney transplantation (Kaboré et al. 2017). This meta-analysis identified 39 risk prediction models published in the scientific literature from 2005-2015. Fourteen studies included predictors measured post-transplant, with or without pre-transplant risk factors as part of the models. These post-transplant predictors most notably included creatinine (Ho et al. 2013) or eGFR [(Hernández et al. 2005); (Moore et al. 2011)], blood pressure, and proteinuria (Foucher et al. 2010) in the weeks, months, and years following transplant. Other predictors assessed included immunological markers, carotid-femoral pulse wave velocity, transplant recipient demographics, and pathophysiological measures to enrich prediction. Previous modeling efforts have attempted to predict risk at varying times post-transplant, including short-term (one to four years) and long-term (≥ five years) outcomes.

Substantial variation exists in the definitions of outcomes, predictors, and methods used to create and validate models described in Kaboré et al. Of the 34 articles identified that developed a new model, only 13 included both internal and external validation methods. While progress has been made, none of these prognostic and predictive tools have been endorsed for use as a surrogate endpoint capable of supporting drug development. Further pursuit of an accurate and well-validated predictive model is warranted.

One notable example is the Birmingham Risk Score, developed by Shabir et al., 2014 (Shabir et al. 2014). This composite scoring system is used to predict the five-year risk of kidney transplant failure using data available at one-year post-transplant. This effort utilized clinical data from 651 subjects from Birmingham, United Kingdom, to develop a model capable of predicting death-censored and overall transplant failure at five years post-transplantation. The Birmingham Risk Score incorporates recipient sex, age, ethnicity, history of AR, one-year eGFR, serum albumin, and UACR. The model was then validated in 3 international cohorts, including 787 subjects from Leeds, United Kingdom, 736 subjects from Tours, France, and 475 subjects from Halifax, Canada. The model was determined to have adequate predictive value with a c-statistic of 0.78-0.90 for death-censored transplant failure and 0.75-0.81 for overall transplant failure. However, the Birmingham Risk Score was limited in the relatively small size of its derivation dataset and its smaller set of considered variables compared to later models.

Building on research assessing the importance of surveillance biopsy and alloantibody data, the Birmingham-Mayo model (Gonzales et al. 2016) was developed to evaluate whether risk models were improved by the addition of biopsy histopathology and/or antibody evaluations. In this work, 1465 adults from the Mayo Clinic in Rochester, Minnesota, USA, had risk scores calculated using the Birmingham risk model. The model was then expanded to include Banff scoring criteria and validated on a cohort of 981 subjects. This process was repeated for DSA status and validated on a cohort of 622 subjects. While the addition of the presence or absence of DSA into the original model failed to improve the predictability of the model, the presence of glomerulitis or chronic interstitial fibrosis on a one-year surveillance biopsy improved the model's prediction (c-statistic=0.90), calibration, and resulted in the reclassification of the graft failure risk in 29% of subjects. The Birmingham-Mayo model has been externally validated in a high-risk cohort, performing well (c-statistic=0.784) when predicting five-year graft loss in subjects with the presence of DSA (Bentall et al. 2019). Despite these improvements over the Birmingham Risk Score, the Birmingham-Mayo model still used a limited set of subjects to derive the model.

iBox Scoring System from Loupy et al., 2019

The Birmingham-Mayo modeling approach was further enhanced by Loupy et al., 2019 (Alexandre Loupy et al. 2019). The authors leveraged the nationalized health care system in France to prospectively follow long-term outcomes of kidney transplant recipients to develop a new model capable of predicting the risk of death-censored allograft loss at 3, 5, and 7-years post-transplant. Quantitative analyses were performed on the data to identify predictors of long-term outcomes. A scoring system, termed iBox, was developed using the identified predictors, including time of post-transplant risk evaluation, eGFR, proteinuria, categorical DSA MFI, and kidney allograft biopsy histopathology. The iBox Scoring System was the first model to include DSA as a predictor, and also had four biopsy predictors (interstitial fibrosis/tubular atrophy, microcirculation inflammation, interstitial inflammation and tubulitis, and transplant glomerulopathy), making it the most comprehensive model to date. The derivation cohort used by Loupy et al., 2019 included 4,000 consecutive subjects from four centers across France with a median follow-up time of 7.65 years, a derivation dataset more than twice as large as previous models. The performance of iBox Scoring System was then evaluated in two validation cohorts (n=3,557) from the United States and Europe, also the largest validation dataset to date. Overall, model performance in Loupy et al. showed good calibration and discrimination at seven-years post-evaluation (c-statistics 0.81, 95% CI 0.79 to 0.83). Validation against three phase II or III clinical trials revealed c-statistics of 0.87, 0.82, and 0.92 in each of the three studies. Further, the risk score was shown to predict the observations of graft loss in these studies accurately. Discrimination (c-statistics) was also included for the European validation cohort and the three RCTs described in Loupy et al., 2019 as additional data supporting this qualification submission, found in Data Loupy et al., 2019 European validation cohort4.3.5and 4.3.6 (Loupy et al., 2019 European validation cohort and Loupy et al., 2019 External validation in three RCTs). The iBox Scoring System model was then assessed in multiple clinical scenarios and different subpopulations with acceptable performance characteristics in each, with c-statistics that ranged between 0.78 and 0.84.

iBox Scoring System (Composite Biomarker Panel)

This Briefing Dossier seeks to build upon Loupy et al., 2019 by converting the iBox Scoring System from a tool in individual patient-level decision making to the application as a surrogate endpoint in regulatory decision-making. The qualification derivation presented in this Briefing Dossier is as described in Loupy et al., 2019, allowing the iBox Scoring System for use as an endpoint in a clinical trial at one-year, previously discussed in the Executive summary 2.5 (Differences between proposed COU and the Loupy et al., 2019 BMJ publication). Additionally, the qualification validation presented in this Briefing Dossier used datasets other than those used for external validation in Loupy et al., 2019 (Alexandre Loupy et al. 2019), as described in Data 4.3.4 (Qualification validation datasets). However, as mentioned above, the European validation cohort and the three RCTs described in Loupy et al., 2019 are summarized in this Briefing Dossier as additional data supporting this qualification submission, found in Data 4.3.5 and 4.3.6 (Loupy et al., 2019 European validation cohort and Loupy et al., 2019 External validation in three RCTs). Several analyses were conducted (outlined in detail in subsequent sections) assessing the performance of the iBox Scoring System as a surrogate endpoint.

Several recent composite scores are described below and summarized in Table 29.

Table 29. Recent composite scores for predicting long-term allograft survival in kidney transplantation

Reference	Model Purpose	Model Description	Predictors	Dataset size

Birmingham Risk Score 2014 (Shabir et al. 2014)	Development of a risk score for predicting five-year transplant failure, based on data available 12 months post-transplantation	Cox PH model - Development of a risk score for predicting five-year transplant failure, based on data available 12 months post-transplantation	UACR, Serum albumin, eGFR, race, age	Derivation = 651
				Validation = 1,998

Birmingham–Mayo model 2016 (Gonzales et al. 2016)	Predicting Individual Renal Allograft Outcomes Using Risk Models with 1-Year Surveillance Biopsy and Alloantibody Data	Cox PH model - Risk prediction score that incorporates easily obtainable clinical factors and determines if histologic findings at 1-year surveillance biopsy and/or serum DSA status could improve the predictability of graft loss by five years	UACR, serum albumin, eGFR, race, age, Banff lesion scores (g score, ci score)	Derivation = 1,465
				Validation = 1,603

iBox Scoring System 2019 (Alexandre Loupy et al. 2019)	Prediction system for risk of allograft survival in subjects receiving kidney transplants: international derivation and validation study	Cox PH model - Application of multivariable Cox PH model for construction of an integrated score	Time from transplant to evaluation, eGFR, UPCR, Banff lesion scores (g, score, ptc score, IFTA score, i score, t score, and cg score), DSA MFI	Derivation = 4,000
				Validation = 3,557

5.2 Events of interest for modeling analyses

The primary event of interest was graft loss. Consideration of additional events of interest, namely death and lost to follow-up, depended on the type of analysis conducted. Validation analyses assessed how well the iBox Scoring System predicted graft loss specifically, and so focused on graft loss as the event of interest and death and lost to follow up were censored. The competing risk analysis, as described in the Modeling analysis methodologies 5.5.2 (Competing risk analysis) investigated the relationship between graft loss and death and therefore considered both death and graft loss events of interest. Finally, the TLS analysis investigated the correlation between a one-year iBox Scoring System surrogate on the true clinical outcome of graft loss, death, or lost to follow up, and therefore considered all three events as events of interest. Lost to follow-up subjects were right-censored.

5.3 Software

Model building, visualization, model assumptions, diagnostics, and external validation were conducted in R (version 4.0.0; Vienna, Austria, R Core Team, 2018) using the packages "survival" (Therneau 2020), "survminer" (Kassambara and Kosinski, n.d.), "dplyr" (Wickham et al. 2020), "survAUC" (Potapov, Adler, and Schmid 2015), "rms" (Harrell 2019) and "riskRegression" (Ozenne et al. 2017).

5.4 Cox proportional hazard (PH) model

The semiparametric Cox PH model relates the graft loss events with covariates,

$$h_i(t) = h_0(t)\exp\left(\sum_{j\in I}\beta_j X_{ij}\right)$$

where $h_i(t)$ is hazard function for individual i determined by a set of j covariates $\{X_{ij}\}$ and corresponding (estimated) coefficients $\{\beta_j\}$, t is the survival time and $h_0(t)$ is the baseline hazard. The use of a Cox PH model implies that the underlying baseline hazard function has no specified distribution and that the PH assumption holds, i.e., the ratio of hazards between different individuals remains constant over time. Following Loupy et al., 2019 (Alexandre Loupy et al. 2019) the estimated coefficients were used to estimate the time-varying probability of graft failure for a specified set of covariates. Additionally, the Cox PH model assumes that censoring is independent of graft survival (the competing risk analysis in Results 5.5.2 (Competing Risk analysis) confirmed the validity of this assumption).

5.4.1 Calculation of the iBox score

The iBox Scoring System is a composite score with a linear combination of parameters from the multivariate Cox PH model, as presented in Modeling analysis methodologies 5.4.1.2 (Multivariate analysis). The raw iBox score can be calculated for each subject in the qualification derivation dataset using the following equation:

Equation 1. Raw iBox score for each subject in the qualification dataset

$$\boxed{Score_i = \sum_{j=1}^{J} \hat{\beta}_j X_{ij}}$$

Where i and j refer to the ith subject and the jth subject feature, respectively, and $\widehat{\beta_j}$ (log of the hazard ratio [HR] values) is the estimated weight of the subject features X_{ij}, i.e., eGFR, UPCR, presence of DSA, and kidney biopsy histopathology (Table 30). For the full iBox Scoring System, there were 3,941 subjects who had the necessary components to calculate an iBox score; 59 subjects were missing biopsy components. For the abbreviated iBox Scoring system, all 4,000 subjects had the necessary components to calculate an iBox Score.

The component measures were assessed at 12 months post-transplantation and used to compute the iBox score. The determined weighting for each component was a coefficient in the multivariate Cox PH model.

Table 30. Calculation of the iBox score for the full iBox Scoring System

$\boldsymbol{iBox_i = \Sigma_{j=1}^{12} \widehat{b}_j X_{i,j}}$ for subject i where	
$\boldsymbol{X_{i,1}}$	Time of post-transplant risk evaluation
$\boldsymbol{X_{i,2}}$	eGFR, where eGFR is measured in ml/min/1.73m^2
$\boldsymbol{X_{i,3}}$	Log transformed (UPCR value[1]), where UPCR is measured in g/g
$\boldsymbol{X_{i,4}}$	Interstitial fibrosis/tubular atrophy (IFTA score): Categorical variable with 3 levels • IFTA score = 0-1 (reference group) • IFTA score = 2 • IFTA score = 3
$\boldsymbol{X_{i,5}}$	Microcirculation inflammation (g score and ptc score): Categorical variable with 3 levels • g and ptc score = 0-2 (reference group) • g and ptc score = 3-4 • g and ptc score = 5-6
$\boldsymbol{X_{i,6}}$	Interstitial inflammation and tubulitis (i score and t score): Categorical variable with 2 levels

	• i score and t score = 0-2 (reference group) • i score and t score ≥ 3
$X_{i,7}$	Transplant glomerulopathy (cg score): Categorical variable with 2 levels • cg score = 0 (reference group) • cg score = ≥ 1
$X_{i,8}$	DSA MFI: Categorical variable with 2 levels • MFI < 1400 (reference group) • MFI ≥ 1400

[1] For proteinuria values below 0.05 g/g are replaced by 0.05 g/g before log-transformation.

For categorical variables with more than 2 levels e.g., IFTA score, the contribution of the variables was calculated as follows: $a_1x_1 + a_2x_2$. If the IFTA score = 0 or 1, then x_1=0 and x_2=0. If the IFTA score = 2, then x_1=1 and x_2=0. If the IFTA score = 3, then x_1=0 and x_2=1. a_1 and a_2 refer to the beta coefficients for the IFTA scores = 2 and 3, respectively.

The distribution of iBox scores in the derivation dataset is shown (Figure 6). The distribution is approximately normal, with lower iBox scores indicating lower risk. Recipients who experience graft loss have a right-shifted distribution of scores, indicating higher risk.

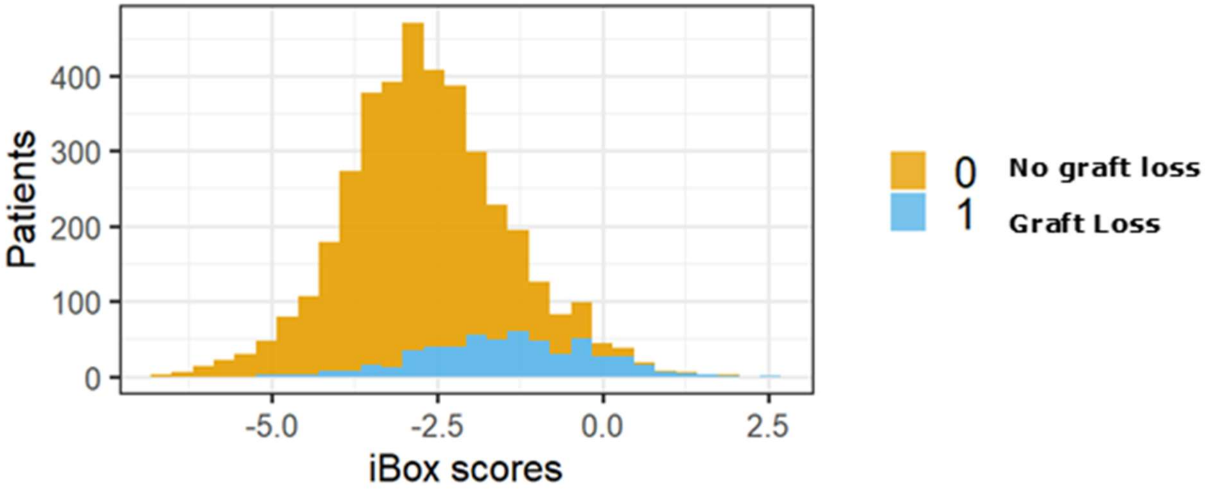


Figure 6. Distribution of iBox scores in the qualification derivation dataset.

C-Path investigated how differences in clinically meaningful changes in iBox Scoring System parameters relate to a difference in the iBox score. The results are shown in Table 31. The values for the two quantitative parameters, eGFR and proteinuria, are linear with the iBox score, so an eGFR difference between two patients of 10 ml/min/1,73m^2 is 0.46 while a difference of 5 ml/min/1.73m^2 is 0.23.

Note that the magnitude of the score difference in the table below does not necessarily give the relative importance of the parameter in a clinical trial setting. Two populations might be expected to vary in average eGFR by 10 ml/min/1.73m^2 or more, translating to a difference in risk score of at least 0.46. In contrast, they might vary in the presence of DSA MFI by 5%, translating to a difference in the iBox score of 0.05 × 0.61 = 0.03.

Table 31. Translating a clinically meaningful difference in iBox Scoring System parameters into a difference in iBox score.

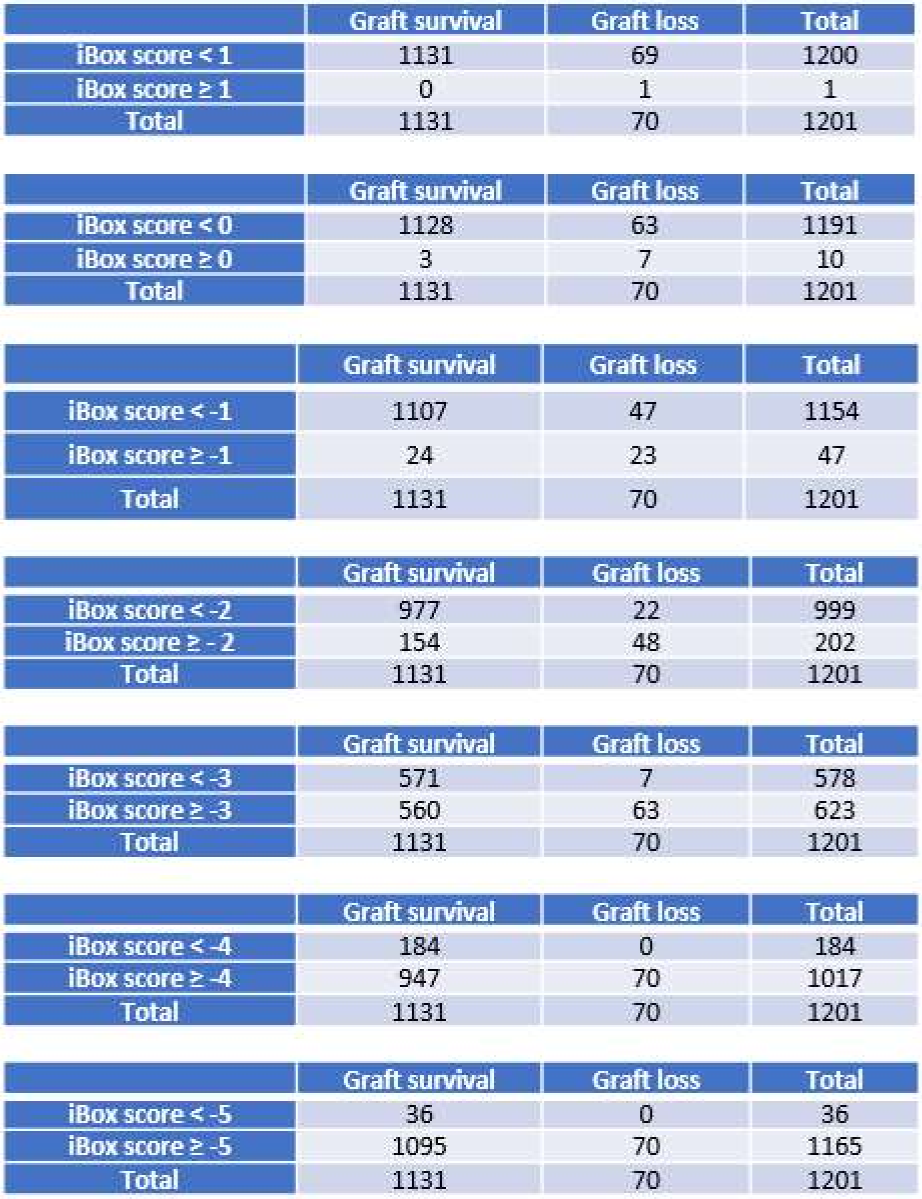
Parameter difference		Magnitude of iBox score difference
eGFR (ml/min/1.73m^2) difference		
5		0.23
8		0.37
10		0.46
Dipstick proteinuria difference	**UPCR proteinuria (log g/g) difference**	
Negative vs. Trace	0.05	0.02
Negative vs. +	0.24	0.10
Negative vs. ++	0.96	0.39
Negative vs. +++	3.11	1.27
DSA MFI difference		
<1400 vs. ≥ 1400		0.61
IFTA score difference		
< 2 vs. 2		0.14
< 2 vs. 3 or more		0.34
g and ptc score difference		
< 3 vs. 3-4		0.36
< 3 vs. 5 or more		0.61
cg score difference		
0 vs. 1 or more		0.38

Sensitivity analyses

Sensitivity analyses with iBox scores in the qualification derivation dataset were performed. The iBox scores from Figure 6 that were measured at approximately one-year post-transplant were split into binary variables at the following cutoffs: 1, 0, -1, -2, -3, -4, and -5. The number of death-censored allograft survivals and graft losses at five-years in the various cut-offs were then recorded in Table 32, and their respective KM estimates were computed in Figure 7 and Figure 8.

The 1201 subjects include patients in the derivation dataset who were part of the a) belatacept switch cohort and b) patients who were on CNI and MTOR regimen. These are excluded from Trial level surrogacy analysis but included here.

Table 32. 2 by 2 contingency tables for iBox cut-offs and five-year death-censored allograft survival.



	Graft survival	Graft loss	Total
iBox score < 1	1131	69	1200
iBox score ≥ 1	0	1	1
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < 0	1128	63	1191
iBox score ≥ 0	3	7	10
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < -1	1107	47	1154
iBox score ≥ -1	24	23	47
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < -2	977	22	999
iBox score ≥ - 2	154	48	202
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < -3	571	7	578
iBox score ≥ -3	560	63	623
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < -4	184	0	184
iBox score ≥ -4	947	70	1017
Total	1131	70	1201

	Graft survival	Graft loss	Total
iBox score < -5	36	0	36
iBox score ≥ -5	1095	70	1165
Total	1131	70	1201

The lower iBox score values always had a higher death-censored allograft survival, as expected. As the iBox score cut-off reduced (i.e., from 1 to -5), the differences in the survival rates also reduced. No graft losses in recipients with scores below -5 were observed, as shown in Figure 7 and Figure 8. The survival rates of those with iBox scores above 0/1 were much lower compared to those with iBox scores below 0/1. The difference in survival rates reduces as the cut-off reduces.

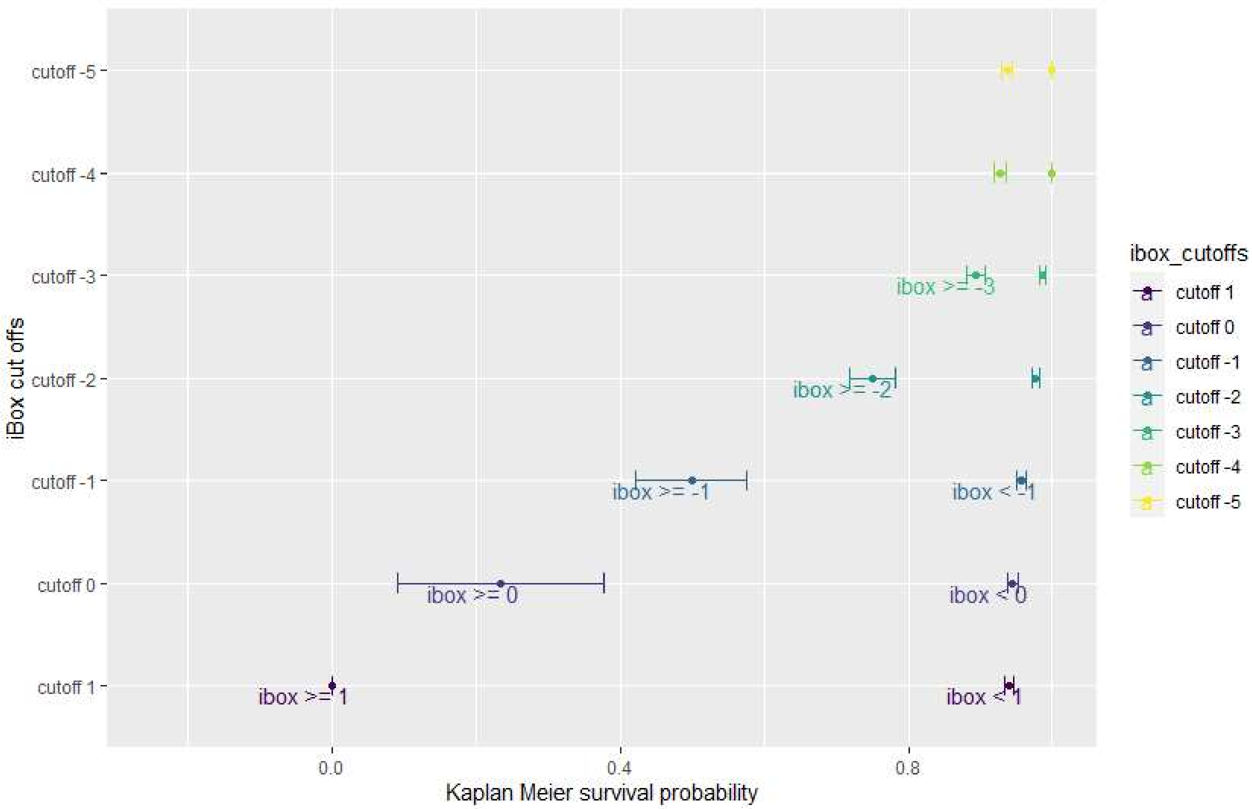


Figure 7: Kaplan-Meier survival probability ± SD for the different iBox cut-offs. For each cut-off, the error bars on the left are the probability estimates for all iBox scores above the cutoff while those on the right are for iBox scores below the given cutoff.

The stratified Kaplan-Meier plots are shown in Figure 8 and demonstrates the reduction in the survival rates differences as the cut-off reduces.

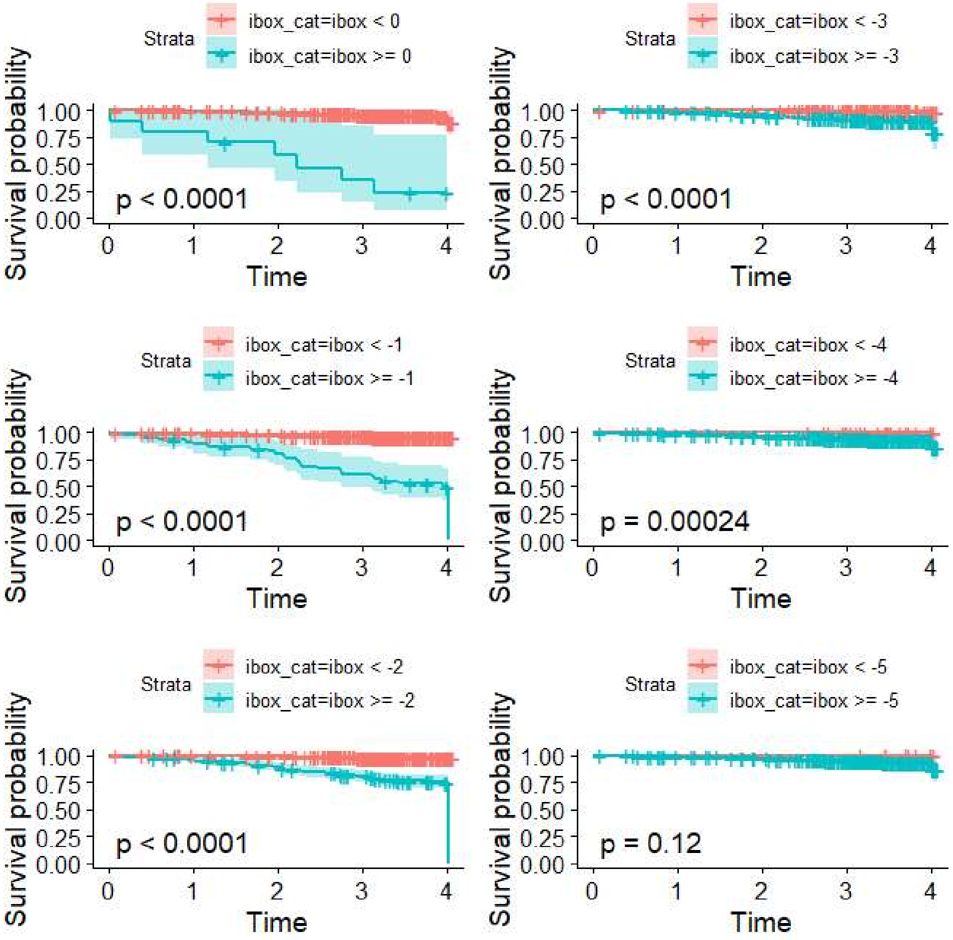


Figure 8: Kaplan-Meier plots for the various iBox Scoring System cut-offs.

The sensitivity and specificity of the iBox score to predict death-censored allograft loss at various threshold cut-off values is shown in Table 33. These values were used in the Receiver Operating Characteristic (ROC) Curve in Figure 9 illustrating the clinical utility of the iBox score with optimal predictive sensitivity and specificity between −2 and −3.

Table 33. Summary of iBox cut-offs and their respective sensitivity and specificity

Cut-off	Sensitivity	Specificity
+1	0.014	1.000
0	0.100	0.997
-1	0.329	0.979
-2	0.686	0.864
-3	0.900	0.505

-4	1.000	0.163
-5	1.000	0.032

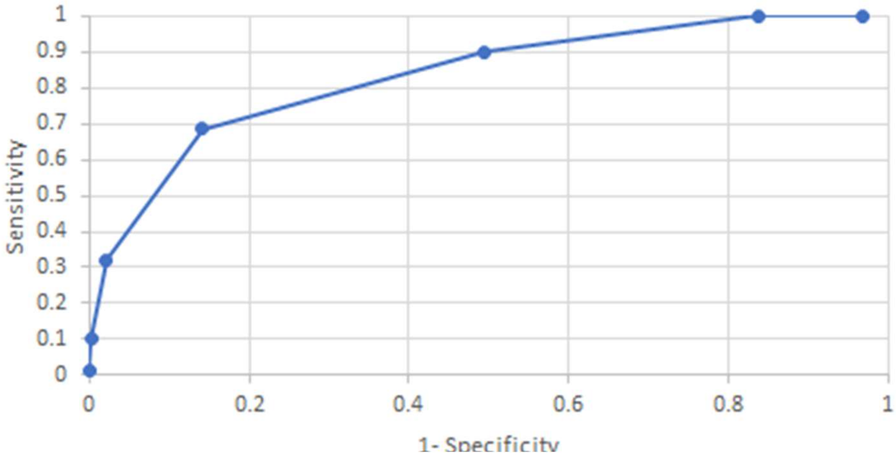


Figure 9. ROC Curve for iBox Scoring System at one-year in the qualification derivation dataset.

5.4.1.1 Univariate analysis

Univariate analysis was performed by estimating a Cox PH model for the covariates listed in Methods 4.3.3.3 (Model variables). The 'coxph' function in the 'survival' R package was used for Cox PH analysis (Therneau 2020). Covariates with no significant univariable association (P-value ≥ 0.1) with death-censored kidney allograft survival were not considered for backward elimination. The P-value was computed using the Wald test, which evaluates whether the covariate coefficient is statistically different from zero.

5.4.1.2 Multivariate analysis

Candidate variables that were not removed in the univariate analysis and clinical considerations were included in the multivariate Cox PH model. The variables included in the final model were selected through backward elimination. Covariates with no significant association (P-value ≥ 0.05) with death-censored kidney allograft survival were dropped from the multivariable model. An abbreviated iBox Scoring System was estimated from the full iBox Scoring System.

5.4.1.3 Model diagnostics

To assess if the PH assumption is satisfied, the log-graphic method was used to test the proportionality hazard assumption. To investigate whether non-linear relationships exist between the continuous covariates and the log hazard of graft loss, martingale residuals were used. For categorical covariates, the goodness of fit was assessed by plotting the error bars around the mean of the martingale residuals. The property of a martingale residuals is that if the estimated model is the true model, then the mean is equal to zero.

5.4.1.4 Model validation

Risk prediction tools like the full and abbreviated iBox Scoring System models are validated by assessing their discrimination, their ability to rank individuals from lower to higher risk, their calibration, and their ability to accurately predict absolute risk level (Crowson, Atkinson, and Therneau 2016). The full model, as the original model derived by Loupy et al. (2019), was first evaluated in the data it was trained on, i.e., the qualification derivation dataset (internal validation). The abbreviated model was treated as a modification of the full model and not validated internally. Once internal validation was complete, both the full and abbreviated models were validated on external datasets that were not part of the training data (external validation) (Collett 2015). The following sections explain the methodologies used to assess both internal and external validation. Given that the full and abbreviated iBox Scoring System is trained primarily on subjects receiving CNI-based maintenance immunosuppressive therapies, performance of the iBox Scoring System in subjects receiving CNI-free maintenance therapies was explored in model validation.

5.4.1.4.1 Internal validation

Harrell's c-statistic (Harrell, Lee, and Mark 1996) (c-statistic) was used to measure the iBox Scoring System's discriminatory ability. The c-statistic gives the probability that, for any two randomly selected individuals, the individual with the shorter survival time has the higher model-predicted hazard of death (Collett 2015). A c-statistic value of 0.5 indicates a discriminatory ability no better than random chance, while a value of 1.0 indicates perfect discriminatory ability (Collett 2015). A c-statistic value of 0.7 or greater indicates good discriminatory ability (Collett 2015).

The iBox Scoring System was derived using data from four different transplant centers in France. To evaluate whether the baseline HR is different between treatment centers, the TTC built a stratified Cox PH model to see if it significantly improves the c-statistic over the non-stratified model. If the two c-statistics were not significantly different, then the baseline HR was assumed to be effectively constant across treatment centers. The TTC also verified whether the eight predictors in the final full iBox Scoring System remained independently associated with allograft survival in the stratified model.

Understanding iBox Scoring System performance on clinically relevant subpopulations and scenarios has the potential to assist with clinical trial development, particularly in the area of inclusion criteria. The TTC further examined internal c-statistics by exploring how the full iBox Scoring System model performs on different clinically relevant subpopulations and scenarios of the qualification derivation dataset using c-statistic values. The TTC evaluated whether subpopulations had significantly different c-statistics than the full derivation population by comparing each subpopulation's 95% CI (calculated as c-statistic estimate ± 1.96 × standard error [SE]) to the full derivation population estimate.

5.4.1.4.2 External validation

External validation was performed using the qualification validation datasets to quantify the full and abbreviated models' predictive power. The discrimination ability of the full and abbreviated iBox Scoring System was assessed as described previously using c-statistic. The full and abbreviated iBox Scoring Systems' calibration on the external datasets was also assessed. Calibration measures whether the model is accurately assessing the absolute risk level (Crowson, Atkinson, and Therneau 2016), which was evaluated here by checking whether observed events match predicted. Calibration was evaluated using a Poisson model

(Crowson, Atkinson, and Therneau 2016). Full details of the Poisson calibration method are presented in full in the Crowson et al. (2016) paper, but a brief synopsis is as follows:

A cumulative hazard function $H(t)$, which can be calculated by integration from a hazard function $h(t)$, can be interpreted as the expected number of events experienced by time t. The calibration method described by Crowson et al. (2016) takes advantage of this property to assess the accuracy of the iBox Scoring System models for the external dataset using the following Poisson regression model:

Equation 2. Poisson Regression Model for assessing accuracy of the iBox Scoring System

$$\boxed{\log(E[Y_i]) = \alpha + log\left(\widehat{H}_{iBox}(t_i, iBox_i)\right),}$$

where Y_i is the number of events experienced by the i^{th} subject of the dataset (in our case 0 if the subject was censored and 1 if the subject experienced an event) during the observation period (from time 0 to t_i), $E[Y_i]$ is the expected number of events if this Poisson model is true, α is the model intercept, and $\widehat{H}_{iBox}(t_i, iBox_i)$ is the cumulative hazard at time of event or censoring t_i as predicted by the full and abbreviated iBox Scoring System for subject i as a function of its iBox score $iBox_i$. Here $log\left(\widehat{H}_{iBox}(t_i, iBox_i)\right)$ is used as an offset (a term where the coefficient is fixed to one) in the Poisson regression model.

The property mentioned above implies that $\alpha = 0$ if the iBox Scoring System model exactly predicts the number of events. Therefore, $\hat{\alpha}$ represents calibration-in-the-large, the degree to which the expected number of events predicted by the iBox Scoring System for the dataset subjects match the expected number of events predicted by the Poisson model (the latter of which is estimated using the actual number of observed events in the external dataset). Statistical significance is evaluated using the SE on this intercept term.

The TTC generated a visual representation of the calibration on the survival scale. This visual was done by first calculating, for a hypothetical subject with survival function predicted from the iBox Scoring System model equal to $S_{iBox} = s$, the survival function predicted from the estimated Poisson model $\left(S_{Poisson} = s^{\left(e^{\widehat{\alpha}}\right)}\right)$ and then plotting $S_{Poisson}$ versus S_{iBox} for s varying from 0 to 1. The CI for $S_{Poisson}$ when $S_{iBox} = s$ can be calculated by applying the delta method to get the CI for α for the non-linear transformation $s^{(e^{\cdot})}$.

As mentioned above, $\alpha = 0$ if the iBox Scoring System model is true; therefore, perfect calibration corresponds to the identity line $(S_{Poisson} = S_{iBox})$. If the confidence band overlaps with an identity line, that implies $S_{Poisson} = S_{iBox}$ within a reasonable margin of error and model calibration is suitable.

5.5 Supplementary analyses

Proteinuria conversions, competing risk, and TLS supplementary analyses were conducted to support the qualification effort, as outlined in Table 34.

Table 34. Overview of supplementary analyses

Supplementary Analysis	Objective	Require CNI-Free	Datasets
Proteinuria conversion	Define proteinuria measure to be used across external validation datasets and appropriate conversion methodology	No	Mayo Clinic Rochester, Helsinki University Hospital, BENEFIT RCT, and BENEFIT-EXT RCT
Competing risk	Examine how the incidence of death affects iBox Scoring System model estimation	No	PTG qualification derivation dataset
TLS	Demonstrate and quantify that the relationship between the treatment effect on the surrogate (iBox Scoring System) and clinically meaningful outcome (graft survival)	Yes	BENEFIT RCT, BENEFIT-EXT RCT, mTORi derivation subset

5.5.1 Proteinuria conversion

In Loupy et al., 2019, UPCR was used to assess proteinuria in the derivation cohort. However, UPCR was not the proteinuria measure used consistently across the validation datasets. Yet to be published, work by this group has demonstrated the performance of the full and abbreviated iBox Scoring System with alternate measures of proteinuria as summarized in the Table 35.

Table 35. Proteinuria measurements across the qualification datasets

Dataset	Proteinuria measurement
Derivation	
Loupy et al., 2019	UPCR
Validation	
Mayo Clinic Rochester	24-hour, UACR
Helsinki University Hospital	Dipstick proteinuria
PTG	UPCR
BENEFIT RCT	Dipstick proteinuria
BENEFIT-EXT RCT	Dipstick proteinuria

Some transplant recipients from the Mayo Clinic Rochester dataset had two proteinuria measurements, 24-hour and UACR, leaving a question of which measurement to use. Because 24-hour proteinuria does not require a conversion to UPCR outside of a change of units, 24-hour proteinuria was used whenever both measurements were present. Analyses were conducted supporting the conversion from various proteinuria measurements to UPCR for use in the full and abbreviated iBox Scoring System.

1. 24-hour proteinuria to UPCR: A full and detailed description of the conversion methodology and results can be found in Results 6.6.1 (Proteinuria conversions).

2. UACR to UPCR: A full and detailed description of the conversion methodology and results can be found in Results 6.6.1(Proteinuria conversions).

3. Dipstick proteinuria to UPCR: A full and detailed description of the conversion methodology and results can be found in Results 6.6.1(Proteinuria conversions).

5.5.2 Competing risk analysis

In the Loupy et al., 2019 publication, the risk of death was not considered, so the TTC investigated whether death affects the estimation of graft loss. The TTC used two methods for identifying whether the competing risk of death affects the full and abbreviated iBox Scoring System's predictions of graft loss. First, cumulative incidence functions (CIF) of graft loss that do and do not account for death were compared. CIFs show the increasing incidence of some event over time and account for censoring. A CIF of graft loss that does not account for death will censor individuals who die, essentially treating death as uninformative to the graft loss incidence. However, if death is informative (i.e., because people who die cannot experience graft failure), then censoring death may result in an overestimate of the true incidence of graft loss. If the incidence of graft loss is overestimated, the CIF that treats death as uninformative should be greater than the one that accounts for death as an informative event, and if not, the two curves should be statistically indistinguishable. Second, a Fine-Gray subdistribution survival model (Austin, Lee, and Fine 2016) was built that accounts for death and compared to the iBox Scoring System, which is a Cox survival model that does not account for death. If graft loss incidence is overestimated, the iBox Scoring System should overestimate the hazard of death compared to a Fine-Gray subdistribution model, because it censors deaths under the assumption they are uninformative. But if the two models are highly similar (i.e., their parameter estimates are within a CI of each other's SEs), then the iBox Scoring System is not overestimating the incidence of graft loss and is appropriate. CIFs were plotted in R using the cmprsk (Gray 2020) R package, and a Fine-Gray subdistribution survival model was built in R using the cmprsk (Gray 2020) and riskRegression (Gerds et al. 2020) R packages.

A full and detailed description of the competing risk results can be found in Results 6.6.2 (Competing risk analysis).

5.5.3 Trial-level surrogacy analysis

TLS analysis was performed to show if a treatment effect on the full and abbreviated iBox Scoring System is predictive of treatment effect on graft survival using RCT results, based on previous work on trial and individual-level surrogacy by Alonso 2016, Bujkiewicz 2019, Daniels 1997 and Blumenthal 2015.

It is well recognized (Baker 2003) that a positive relationship between a potential surrogate endpoint (in this Briefing Dossier, the full and abbreviated iBox Scoring System) and a true clinical outcome (e.g., death-censored allograft loss) does not necessarily imply that a positive difference between treatment groups in the surrogate will translate into a positive difference between treatment groups in the true clinical outcome. This necessitates the need to conduct the TLS analysis. The TLS analysis was conducted in two stages:

1. Estimated the treatment effect for each trial on full and abbreviated iBox Scoring System and graft loss.

2. Computed the correlation coefficient and/or the surrogate threshold effect (STE).

With the existing datasets, three datasets were proposed for the analysis based on the availability of a treatment and control arm. Two of these datasets were the BENEFIT and BENEFIT-EXT RCTs, but neither of the trials was prospectively designed to assess five-year graft survival. Due to the paucity of the RCT data with five-year follow up and the iBox Scoring System features at one-year post-transplant, a third dataset was constructed retrospectively from a subset of subjects in the derivation dataset. This third TLS dataset, referred to as "mTORi derivation subset," consisted of subjects who were on a CNI-free mTORi-based therapy, sirolimus or everolimus, versus CNI-based therapy at the time of transplant with full and abbreviated iBox Scoring System evaluations at one-year post-transplant, consistent with the proposed COU. This mTORi derivation subset used propensity score techniques to reweight subjects on the two arms and randomization emulation to reduce potential confounding issues that can be present when examining non-RCT data. This provided three RCT datasets (two prospective and one retrospective) with CNI and CNI-free arms for TLS analyses.

Three different versions of TLS analyses were performed for the full and abbreviated iBox Scoring System models, as described below and in Table 36:

A. Analysis of five-year death-censored allograft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant.

B. Analysis for five-year death-censored allograft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant with the addition of subjects that died/withdrew/lost their graft within the first year of transplant.

C. Analysis for five-year overall graft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant with the addition of subjects that died/withdrew/lost their graft within the first year of transplant.

For analysis B and C, the two BMS RCTs were the data sources used while in analysis A the two BMS RCTs were used as well as the qualification derivation dataset. Below is a summary of the inclusion criteria for three different versions (A-C) of TLS analyses.

Table 36. Three different versions of TLS analyses used for full and abbreviated iBox Scoring System models

TLS analyses	Event definition	Follow up time	Include study med discontinuation (yes or no)
A. Death-censored allograft loss without imputation	All subjects that experience graft loss beyond one-year post-transplant	Time from risk evaluation (one year) to graft status	No

B. Death-censored allograft loss with imputation	All subjects that experience graft loss before/after one year	Time from transplant to graft status	Yes
C. Overall graft loss with imputation	All subjects that die/experience graft loss before/after one year	Time from transplant to death/graft status	Yes

5.5.3.1 Imputation of iBox scores

In analyses B and C, the subjects who died/withdrew/lost their graft before the first year of transplantation have missing iBox score values. These subjects were assigned imputed iBox score values calculated according to the equation in section 5.4.1 (Calculation of the iBox Score)

$$Score_i = \sum_{j=1}^{J} \hat{\beta}_j X_{ij}$$

corresponding to the worst-case scenario as follows:

1. eGFR value set at 0 ml/min/1.73m^2.

2. log UPCR value set at the max dipstick-imputed score. Imputation methodology for dipstick proteinuria to UPCR is described in Modeling analysis methodologies, 5.5.1 (Proteinuria conversion).

3. IFTA score set at maximum value of three.

4. Microcirculation inflammation (g score and ptc score) set at maximum categorical value of > 4.

5. i + t score set at a maximum categorical value of ≥ 3.

6. cg score set at maximum categorical breakdown of ≥ 1.

7. DSA MFI set at maximum binary qualitative cut-off of ≥ 1,400.

8. Time from transplant to evaluation is fixed to 1.

The imputed iBox score was 2.79 for the full iBox Scoring System. The imputed iBox score was 1.48 for the abbreviated iBox Scoring System. These imputed iBox scores for the full and abbreviated iBox Scoring System are derived as shown in Table 37. Note that the coefficients i.e., $\hat{\beta}_j$'s used to derive the full and abbreviated iBox Scoring System models shown in Table 37 can be found in Table 40 and Table 41, respectively.

Table 37. Imputed iBox score calculation for the full and abbreviated iBox Scoring System models

	Imputed iBox score calculation
Full iBox Scoring System	0.0791+0.4069*log(3.236)+0.3432+0.6079+0.2886+0.3848+0.6080 = **2.79**
Abbreviated iBox Scoring System	0.1150+ 0.4652*log(3.236)+0.8164 = **1.48**

5.5.3.2 Step 1: Computation of treatment effects

The goal of Step 1 in TLS analysis was to generate the treatment effects and SEs on the full and abbreviated iBox Scoring System at one-year post-transplant and on the five-year death-censored graft survival for the pseudo trials as well as the correlation coefficients between the two treatment effects within a pseudo trial.

5.5.3.2.1 BENEFIT and BENEFIT-EXT RCTs

Subjects from the two RCTS who had full and abbreviated iBox Scoring System measurements at one-year post-transplant and known graft status up to 5 years post-transplant were selected for analyses.

For all TLS analyses versions (A-C), the treatment effect ($\hat{\theta}_i$) (log-HR), and its variance ($\hat{\sigma}^2_{\hat{\theta}_i}$) were calculated from the log-rank test.

Given an example outcome at time (j<5 years; where j=1,..J are the distinct event times before five years) from a study as shown in

Table 38, the log-HR is computed as:

Equation 3. Log Hazard Ratio

$$\log hazard\ ratio = \frac{\sum_{j=1}^{J} Observed\ events\ _{Belatacept,j}\ - Expected\ events\ _{Belatacept,j}}{\sum_{j=1}^{J} Var_{Belatacept,j}}$$

Where expected events at time j were defined as follows:

Equation 4. Expected events and Variance for the log-rank test

$$Expected\ events\ _{Belatacept,j} = \frac{C_j * G_j}{N_j} \text{ and } Var_{Belatacept,j} = \frac{C_j * G_j * T_j * R_j}{N_j^2(N_j - 1)}$$

Table 38. Example study outcomes at a fixed time j> one year (number of events per arm)

Treatment	# of graft losses at j-years post-transplant	# at risk at j-years post-transplant	Total
BELA	G1	R1	C_j
CsA	G2	R2	T_j
Total	G_j	R_j	N_j

For analysis A, i.e., with no imputation, the treatment effect (i.e., mean difference in iBox score ($\hat{\gamma}_i$) and its variance ($\hat{\sigma}^2_{\hat{\gamma}_i}$)) for each trial were calculated using a t-test for the iBox scores.

For analyses B and C, i.e., with imputation, the treatment effect ($\hat{\gamma}_i$) , was computed as the difference in medians between the iBox scores while the variance of the median difference ($\hat{\sigma}^2_{\hat{\gamma}_i}$) for each trial was calculated using 2000 bootstrap samples.

The correlation coefficient between the two treatment effects ($\hat{\rho}_{\hat{\gamma},\hat{\theta},i}$) was computed from the full and abbreviated iBox Scoring System and graft survival treatment effects generated from 2000 bootstrap samples.

5.5.3.2.2 CNI versus CNI-free subjects in the mTORi derivation subset

Observational studies: Randomization emulation:

Randomization emulation was performed to ensure that the two treatment groups were comparable in terms of baseline covariates. The method used for randomization emulation was inverse weighting based on propensity scores.

Propensity score computation:

The propensity score was computed using the logistic regression method:

Let Z denote treatment status: 0 for the reference treatment (e.g., CNI) and 1 for the other treatment (e.g., CNI-free), and e denote the propensity score.

Equation 5. Propensity score computation

$$\mathrm{logit}(\mathrm{Pr}(Z = 1)) = \beta' L$$

Where

$$\mathrm{e} = \mathrm{expit}(\beta' L)$$

The covariates L were selected based on the following criteria:

- Covariates that were predictive of both treatment and outcome were included.
- Covariates that were predictive of outcome (but unrelated to treatment) remained included in the model to help precision.
- Covariates that were predictive of treatment (but unrelated to the outcome) were not included in the propensity score model. They are detrimental for precision and increase bias due to unmeasured confounding.

<u>Computation of treatment weights:</u>

The stabilized inverse probability of treatment weights (IPTWs) was defined as

Equation 6. Computation of treatment weights

$$w = \Pr{(Z=1)}\frac{Z}{e} + \Pr{(Z=0)}\frac{1-Z}{1-e}$$

where e=propensity score (probability treatment=1 from logistic model)

<u>Treatment effects</u>

The stabilized weights were then used to compute the average treatment effects. For the iBox score, a weighted linear regression model was used while the effect on graft loss was computed using a weighted Cox PH model. The standard deviation (SD) of the estimated log-HRs and the SD of the estimated difference in mean iBox score were generated from the weighted and cox linear models. It was not possible to generate reliable bootstrap estimates of the correlation between the log HR and difference in iBox scores due to the low number of events (4) in the CNI-free arm, so a value of 0.25 was assigned. This value was estimated from the BENEFIT RCT using bootstrap samples.

5.5.3.3 Step 2: Generation of the trial-level coefficient

Coefficient of determination/STE

The treatment effects computed in step 1 above, i.e. $(\hat{\gamma}_i, \hat{\theta}_i)$ were the dependent variables in step 2. The variance covariance matrix used in Step 2 was also estimated in Step 1.

In step 2, the goal is to generate the trial-level correlation coefficient (i.e., square root of the coefficient of determination). Following Daniels 1997, a Bayesian approach was followed as described below:

Equation 7. Coefficient of determination/STE

$$\begin{pmatrix}\hat{\gamma}_i \\ \hat{\theta}_i\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma_i \\ \theta_i\end{pmatrix}, \begin{pmatrix}\hat{\sigma}^2_{\hat{\gamma}_i} & \hat{\rho}_{\hat{\gamma},\hat{\theta},i}, \hat{\sigma}_{\hat{\gamma}_i}, \hat{\sigma}_{\hat{\theta}_i} \\ \hat{\rho}_{\hat{\gamma},\hat{\theta},i}, \hat{\sigma}_{\hat{\gamma}_i}, \hat{\sigma}_{\hat{\theta}_i} & \hat{\sigma}^2_{\hat{\theta}_i}\end{pmatrix}\right)$$

Where

$$\boxed{\begin{pmatrix}\gamma_i\\ \theta_i\end{pmatrix} \sim N\left(\begin{pmatrix}\gamma\\ \theta\end{pmatrix}, \begin{pmatrix}\sigma_\gamma^2 & \rho, \sigma_\gamma, \sigma_\theta\\ \rho, \sigma_\gamma, \sigma_\theta & \sigma_\theta^2\end{pmatrix}\right)}$$

Priors:

The TTC assigned non-informative (flat) priors on all the parameters/hyperparameters except sigmas (variances) which were assigned weakly informative priors due to the small number of pseudo trials as proposed by Gelman et al.

- $\gamma, \theta \sim Normal\ (0, 1000)$
- $\rho \sim Uniform\ (-1,1)$
- $\sigma_\gamma, \sigma_\theta = |Y|\ where\ Y \sim N(0,1)$ i.e., half normal distribution.

Convergence assessment

Convergence of the posterior samples was evaluated using trace plots, autocorrelation, and the Gelman and Rubin convergence diagnostic (potential scale reduction factor).

Interpretation:

The extent of TLS was assessed via the posterior distribution of ρ. The posterior mean and SD were reported and the 95% credible interval of the distribution. TLS results would be strong if the lower 95% prediction value is greater than 0.77. (Lassere et al. 2012).

In addition, results were assessed graphically. A scatter plot of the $\begin{pmatrix}\hat{\gamma}_i\\ \hat{\theta}_i\end{pmatrix}$ was overlaid with the regression $\theta_i|\gamma_i \sim N\left(\theta + \rho\, \sigma_\theta/\sigma_\gamma\, (\gamma_i - \gamma),\ \sigma_\theta^2(1-\rho^2)\right)$.

The STE (Burzykowski, 2006) was calculated as follows. For a range of "true" values γ_i^* for the iBox Scoring System treatment effect in a new trial, the Bayesian 95% prediction interval $\theta_i^*\ |\gamma_i^* \sim N\left(\theta + \rho\, \sigma_\theta/\sigma_\gamma\, (\gamma_i^* - \gamma),\ \sigma_\theta^2(1-\rho^2)\right)$, was calculated directly from the Markov-Chain Monte Carlo samples. This considered the uncertainty in all parameters. The STE is the smallest effect size γ_i^* such that the 95% prediction interval for θ_i^* lies entirely below 0, if it exists. A narrow prediction interval is evidence of a good surrogate.

5.6 All-cause allograft loss

A one-year post-kidney transplant score to be used as a clinical trial endpoint predictive of five-year allograft survival accounting for both deaths and graft losses as events was explored. The qualification derivation and validation datasets were re-analyzed, restricting the analysis to recipients with abbreviated iBox scores (model includes eGFR, proteinuria, and presence of DSA). eGFR was calculated with the revised CKD-EPI equation without race input, described in the recent literature by Inker et al., 2021.

6 RESULTS

The execution of the modeling analysis methodologies outlined in Section 5 are reported here. The iBox Scoring System was originally developed by Loupy et al., 2019 and is rederived, validated internally and externally, and then tested to determine if a treatment effect on this surrogate corresponds to an effect on death-censored allograft survival at five-years post-transplant in the TLS analysis. The full and abbreviated iBox Scoring System was developed to predict death-censored allograft survival, and so the derivation and validation analyses also focus on death-censored allograft survival. However, in the context of clinical trials, death is also a relevant outcome, so all-cause graft survival is also considered in the TLS analysis in addition to death-censored allograft survival.

6.1 Univariate analysis

Univariate analysis of 31 candidate variables was conducted, as previously described in the Modeling analysis methodologies 5.4.1.1 (Univariate analysis). From these 31 variables, variables with P-values < 0.1 were considered eligible for backward elimination, summarized in Table 39. The four variables excluded were: (1) recipient age (P-value = 0.46), (2) recipient sex (P-value = 0.97), (3) number of HLA-A/B/DR mismatches (P-value = 0.29), and (4) donor sex (P-value = 0.83).

There were 27 candidate variables remaining that were eligible for backward elimination. These 27 candidate variables included: donor age, donor type, donor hypertension history, donor diabetes mellitus history, donor creatinine concentration, ECD, previous kidney transplant, CIT, Thymoglobulin™ induction immunosuppression, DGF, pre-existing DSA, time of risk evaluation, eGFR, proteinuria, IFTA, arteriosclerosis, hyalinosis, interstitial inflammation and tubulitis (i and t score), transplant glomerulopathy (cg score), endarteritis, C4d graft deposition, microcirculation inflammation (g score and ptc score), PVAN, nephropathy recurrence, aAMR, aTCMR, and DSA MFI.

Of the 27 candidate variables, nine variables were ruled out due to clinical considerations, previously described in Data 4.3.3.4 (Clinical considerations). These nine excluded variables included: donor age (1), donor hypertension (2), donor creatinine concentration (3), previous kidney transplant (4), DGF (5), and four Banff histopathological diagnoses [i.e., PVAN (6), nephropathy recurrence (7), aAMR (8), and aTCMR (9)].

Table 39. Univariate analysis of 31 candidate variables

	No. of subjects	No. of events*	HR (95% C.I.)	P-value
1 Time from transplant to evaluation (per 1 year increment)	4,000	549	1.26 (1.20 to 1.33)	< 0.001
Recipient Characteristics				
2 Recipient age (per 1-year increment)	4,000	549	1 (1.00 to 1.01)	0.46
3 Recipient sex:				
Female	1,550	214	1	

Male	2,450	335	1 (0.85 to 1.19)	0.97
Transplant Characteristics				
4 Donor age (per 1-year increment)	4,000	549	1.02 (1.01 to 1.02)	< 0.001
5 Donor sex:				
Female	1,849	254	1	
Male	2,151	295	0.98 (0.83 to 1.16)	0.83
6 Donor type:				
Deceased	3327	498		
Living	673	51	0.49 (0.36 to 0.65)	< 0.001
7 Donor history of hypertension:				
No	2,898	340	1	
Yes	1,005	195	1.84 (1.54 to 2.2)	< 0.001
8 Donor history of diabetes:				
No	3,630	491	1	
Yes	231	39	1.39 (1.00 to 1.93)	0.05
9 Donor creatinine concentration:				
< 1.5 mg/dL	3,540	467		
≥ 1.5 mg/dL	422	75	1.43 (1.12 to 1.82)	0.004
10 ECD:				
No	2,586	285	1	
Yes	1409	263	1.9 (1.60 to 2.24)	< 0.001
11 Previous kidney transplant:				
No	3,395	421	1	
Yes	605	128	1.86 (1.53 to 2.27)	< 0.001
12 CIT:				
< 12 hrs	1,120	106	1	
≥ 24 hours	757	121	1.73 (1.33 to 2.25)	< 0.001

12-24 hours	2,099	319	1.61 (1.30 to 2.01)	< 0.001
13 Thymoglobulin™ induction immunosuppression:				
No	1,643	209	1	
Yes	2,104	316	1.25 (1.05 to 1.49)	0.01
14 Number of HLA-A/B/DR mismatches	4,000	549	1.03 (0.97 to 1.1)	0.29
15 DGF:				
No	2,851	319	1	
Yes	1,046	215	1.94 (1.63 to 2.3)	< 0.001
16 Pre-existing DSA:				
No	3,278	425		
Yes	722	124	1.51 (1.23 to 1.84)	< 0.001
Functional Variables				
17 eGFR (mL/min/1.73m^2)	4,000	549	0.94 (0.93 to 0.95)	< 0.001
18 Log transformed UPCR proteinuria (g/g)*	4,000	549	1.99 (1.86 to 2.13)	< 0.001
Post-Transplantation Structural Histopathology Variables				
19 IFTA:				
0-1	3,099	331	1	
2	555	116	2.15 (1.74 to 2.65)	< 0.001
3	321	95	3.36 (2.67 to 4.22)	< 0.001
20 Vascular fibrous intimal thickening (Arteriosclerosis) (cv score):				
0	1,365	137	1	
≥ 1	2,446	386	1.62 (1.33 to 1.97)	< 0.001
21 Arteriolar hyalinosis (ah score):				
0	1,567	149	1	
≥ 1	2,360	381	1.74 (1.44 to 2.1)	< 0.001
22 Interstitial inflammation and tubulitis (i score and t score):				

0-2	3,610	456	1	
≥ 3	390	93	1.97 (1.58 to 2.46)	< 0.001
23 Transplant glomerulopathy (cg score):				
0	3,702	449	1	
≥ 1	260	94	3.7 (2.96 to 4.62)	< 0.001
24 Intimal arteritis (endarteritis) (v score):				
0	3,794	506	1	
≥1	96	27	2.26 (1.54 to 3.33)	< 0.001
25 C4d graft deposition:				
No	3,452	416	1	
Yes	548	133	2.45 (2.01 to 2.98)	< 0.001
26 Microcirculation inflammation (g score and ptc score):				
0-2	3,616	422	1	
3-4	308	92	3.07 (2.45 to 3.85)	< 0.001
5-6	76	35	4.99 (3.53 to 7.04)	< 0.001
27 PVAN:				
No	3,902	518	1	< 0.001
Yes	97	31	2.82 (1.96 to 4.05)	
28 Nephropathy recurrence:				
No	3,868	510	1	
Yes	130	38	2.55 (1.83 to 3.55)	< 0.001
29 aAMR:				
No	3,380	405	1	
Yes	620	144	2.09 (1.73 to 2.53)	< 0.001
30 aTCMR:				
No	3,784	497	1	
Yes	216	52	1.91 (1.44 to 2.54)	< 0.001
Post-Transplantation Immunological Variables				

31 DSA MFI:				
< 1400	3,659	444	1	
≥ 1400	341	105	3.23 (2.61 to 4.00)	< 0.001

*Proteinuria values of 0 will have a small positive value added to prevent undefined values

6.2 Multivariate analysis

6.2.1 Full iBox Scoring System

The 18 candidate variables identified in the univariate analysis were considered eligible for backward elimination. These variables included: donor type, donor diabetes mellitus history, ECD, CIT, Thymoglobulin™ induction immunosuppression, pre-existing DSA, time of risk evaluation, eGFR, proteinuria, IFTA, arteriosclerosis, hyalinosis, interstitial inflammation and tubulitis, transplant glomerulopathy, endarteritis, C4d graft deposition, microcirculation inflammation, and DSA MFI.

Variables were dropped sequentially until the maximum P-value was observed to be < 0.05. The variables dropped were: (1) donor type, (2) donor diabetes mellitus history, (3) ECD, (4) CIT, (5) Thymoglobulin™ induction immunosuppression, (6) pre-existing DSA, (7) arteriosclerosis, (8) hyalinosis, (9) endarteritis, and (10) C4d graft deposition.

There were eight variables retained in the full iBox Scoring System multivariate analysis. These variables included: (1) time of risk evaluation, (2) eGFR, (3) proteinuria, (4) interstitial fibrosis/tubular atrophy, (5) microcirculation inflammation (g and ptc), (6) interstitial inflammation and tubulitis, (7) transplant glomerulopathy, and (8) DSA MFI and were combined to generate the full iBox Scoring System, summarized in Table 40.

Table 40. Final eight variables retained in the full iBox Scoring System multivariate analysis

Factor	No. of subjects	No. of events*	HR (exp $[\hat{\beta}_j]$) (95% C.I.)*	P-value
Time from transplant to evaluation (years)	3,941	538	1.08 (1.03 to 1.14)	0.0032
eGFR (mL/min/1.73 m2)	3,941	538	0.96 (0.95 to 0.96)	<0.0001
Log transformed UPCR Proteinuria (g/g)	3,941	538	1.5 (1.39 to 1.62)	<0.0001
Interstitial fibrosis/tubular atrophy (IFTA score):				
0-1	3,074	330	1	
2	550	115	1.14 (0.92 to 1.43)	0.2256
3	317	93	1.41 (1.1 to 1.8)	0.0059

Microcirculation inflammation (g score and ptc score):				
0-2	3,568	414	1	
3-4	299	90	1.43 (1.11 to 1.85)	0.0057
5-6	74	34	1.84 (1.25 to 2.7)	0.0019
Interstitial inflammation and tubulitis (i score and t score):				
0-2	3,559	447	1	
≥ 3	382	91	1.33 (1.06 to 1.68)	0.0141
Transplant glomerulopathy (cg score)				
0	3,684	445	1	
≥ 1	257	93	1.47 (1.14 to 1.9)	0.0033
DSA MFI				
< 1400	3,607	435	1	
≥ 1400	334	103	1.84 (1.44 to 2.34)	<0.001

* $\hat{\beta}_j$ = the log of the HR values

6.2.2 Abbreviated iBox Scoring System

The rationale to support two iBox Scoring System models (with and without biopsy) can be found in Methods 4.1.1 (Rationale to support two iBox Scoring System models).

The full iBox Scoring System (described above) was re-estimated by dropping the four kidney allograft biopsy histopathology variables in Table 38. The HRs for these variables in the abbreviated iBox Scoring System are described in Table 41.

Table 41. Final four variables retained in the abbreviated iBox Scoring System multivariate analysis

Factor	No. of subjects	No. of events*	HR (exp [$\hat{\beta}_j$]) (95% C.I.)*	P-value
1. Time from transplant to evaluation (years)	4,000	549	1.12 (1.07 to 1.18)	<0.0001
2. eGFR (mL/min/1.73 m^2)	4,000	549	0.95 (0.95 to 0.96)	<0.0001
3. Log transformed UPCR proteinuria (g/g)	4,000	549	1.59 (1.48 to 1.71)	<0.0001
4. DSA MFI				

< 1400	3659	444	1	
≥ 1400	341	105	2.26 (1.82 to 2.82)	<0.001

* $\hat{\beta}_j$ = the log of the HR values

6.3 Model diagnostics

Model diagnostics were performed as described previously in Modeling analysis methodologies 5.4.1.3 (Model diagnostics). The martingale residuals were plotted for the eight components in the full iBox Scoring System core to test the linearity assumption of covariate relationships as shown in Figure 10. The loess line did not differ from the linear regression line suggesting that a linear relationship was sufficient.

An error bar of the mean martingale residuals for the categorical variables was also included (Figure 11). Figure 11 shows that including the categorical variables in additive form was reasonable as the 95% interval crossed zero for all levels of each categorical variable. The error bar on each side is 2 SDs from the mean martingale values to generate the 95% CI; SDs were generated using 5000 bootstrap samples for each categorical variable. Additionally, the martingale residuals were used to examine its robustness at one-year post-transplant, consistent with the proposed COU of the iBox Scoring System (Figure 12). To do this, the time of risk evaluation was stratified to 0.9-1.1 years and those outside this range. The martingale residuals were then plotted as shown below in Figure 12. The mean martingale residuals did not differ between the two-time groups since the 95% interval included 0.

The graphical plots were used to assess the viability of the PH assumption, as shown in Appendix (Revised-Supporting results [Model diagnostics]). The log of the cumulative hazard was plotted against the categorical covariates. The plots were approximately parallel indicating that the PH assumption was reasonable as shown below.

The dfbeta and Schoenfeld residuals were used to identify influential observations and outliers. The results are shown in Appendix (Revised-Supporting results [Model diagnostics]). There was no evidence that some observations were highly influential or outliers.

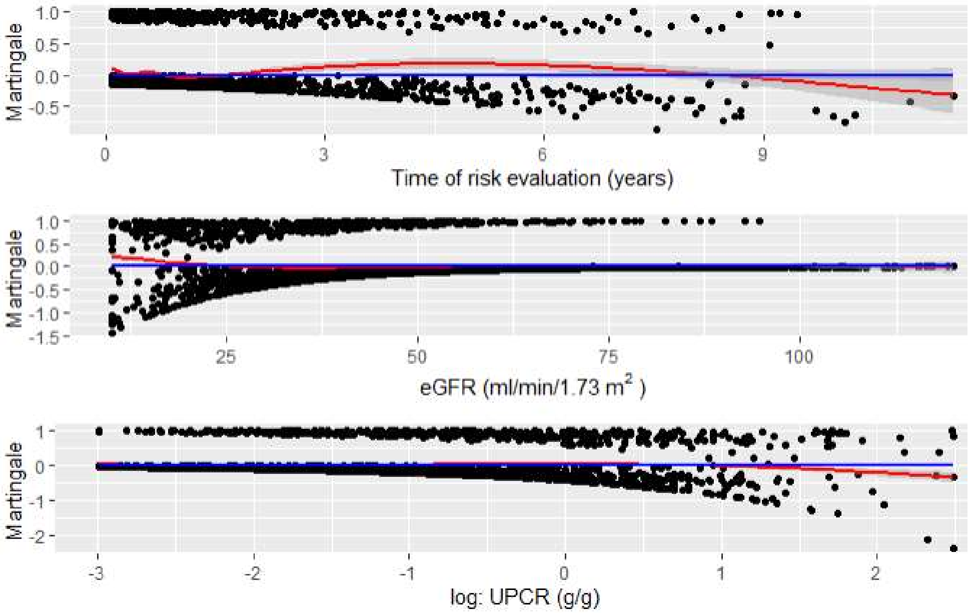


Figure 10. Plot of martingale residuals of continuous covariates for the null model i.e., model with only one covariate. Red is loess curve and blue is zero line.

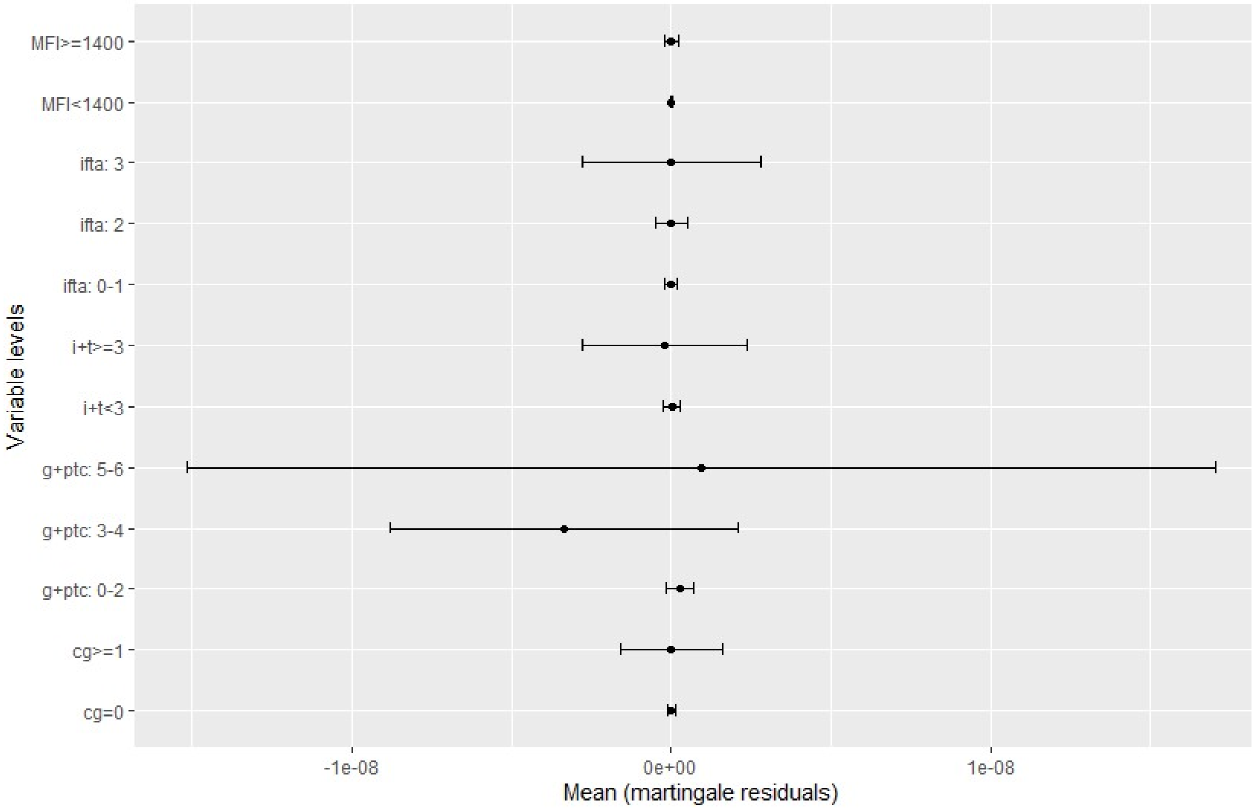


Figure 11. Error bar plot (Mean ± 2*SD) of the martingale residuals for the categorical variables for the null model (i.e., Cox PH model with one categorical variable only).

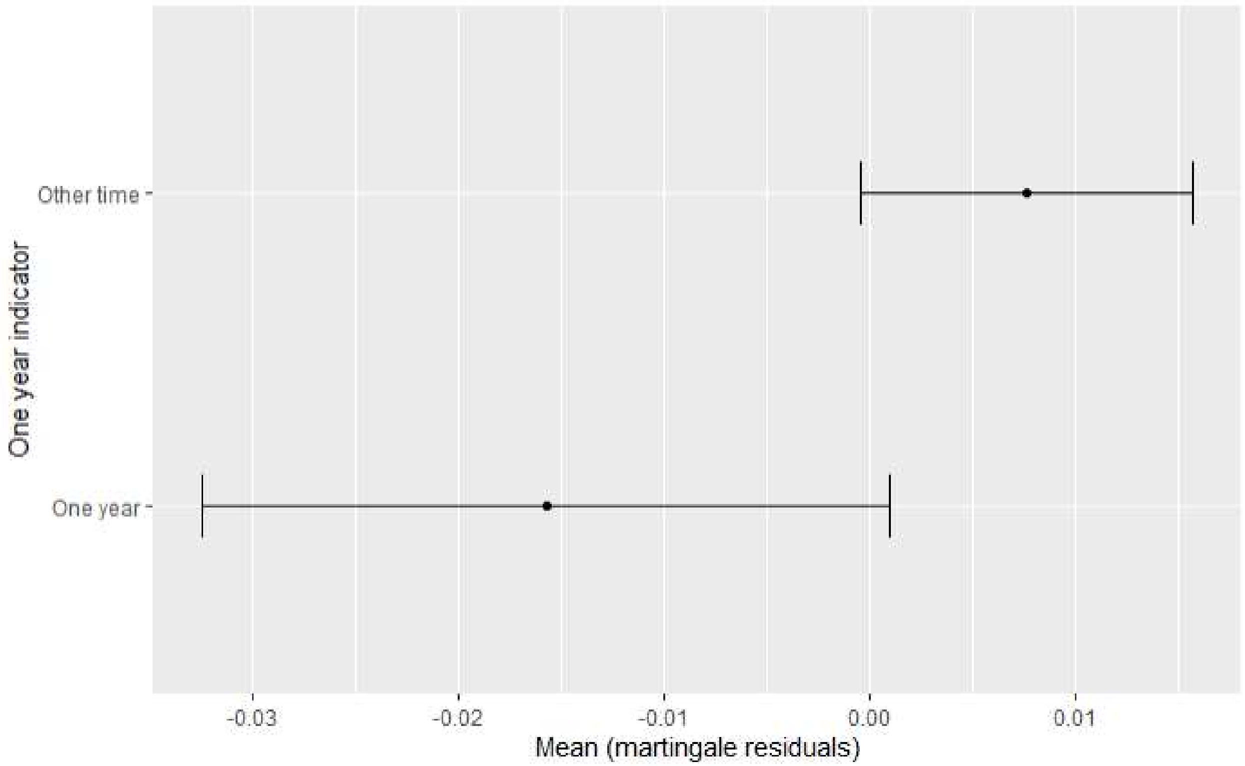


Figure 12. Error bar plot of the martingale residuals for the categorical covariates at one-year post-transplant versus other time from transplant.

6.4 iBox Scoring System compared with single components of the iBox Scoring System

Given that the causes of late graft failure are multifactorial, predicting such failure accurately with a single marker may not be optimal. The iBox Scoring System was designed as a composite marker to fully reflect the heterogeneity of graft failure. The iBox model includes parameters that have demonstrated to be mechanistically associated with increased risk of late graft functional decline and failure (i.e., eGFR, proteinuria, DSA, and kidney allograft biopsy histopathology). Therefore, iBox Scoring System outperforms any of the single components of the iBox Scoring System (Figure 13 iBox Scoring System compared with any single components of the iBox Scoring System).

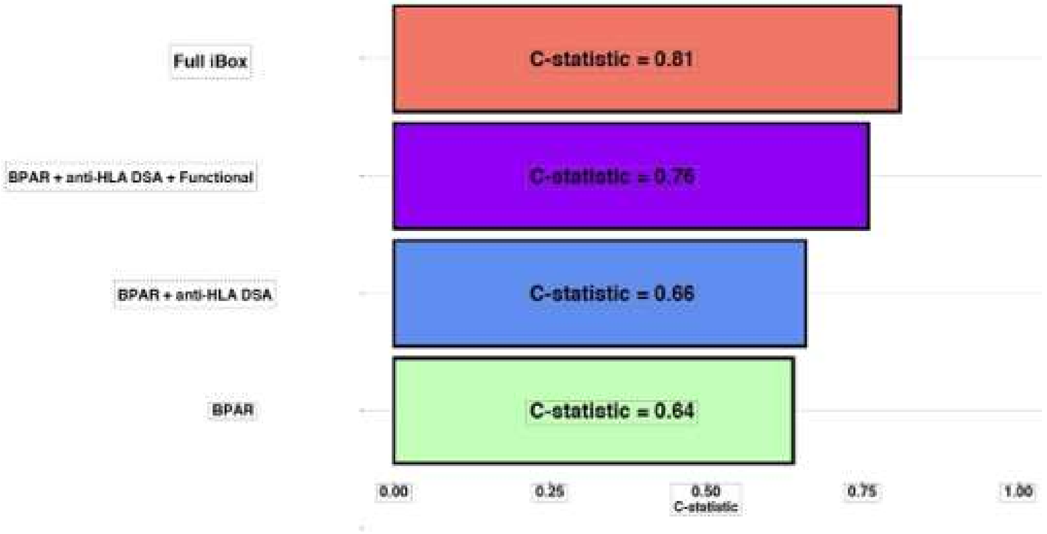


Figure 13. iBox Scoring System compared with any single components of the iBox Scoring System

6.5 Model validation

The following two sections explore model validation. The first section (6.5.1 Internal validation) focuses on the internal validation of the full iBox Scoring System to verify performance on the data the model was trained on (i.e., the qualification derivation dataset) and identify contexts in which the model may lose predictive power. Because the iBox Scoring System was trained on data out to seven years post-evaluation, internal validation includes all data out to seven years unless otherwise noted. The abbreviated iBox Scoring System is not internally validated as it is treated as a modification of the full iBox Scoring System and not as a new model separate from the original. The second section (6.5.2 External validation) focuses on the external validation of the full and abbreviated iBox Scoring System models by assessing their discrimination and calibration on external datasets using the proposed COU. External validation here both evaluates the efficacy of the full iBox Scoring System and investigates the loss, if any, in predictive power that comes from removing potentially useful model coefficients. An abbreviated iBox Scoring System without biopsy is investigated here, while the additional loss of DSA and proteinuria information is investigated in the Appendix (Revised-Supporting results).

6.5.1 Internal validation

The c-statistics for the derivation dataset were 0.809 and 0.803 for the full and abbreviated iBox Scoring Systems, respectively (Table 42). The c-statistic for the abbreviated iBox Scoring System is shown here to demonstrate that it is not significantly different than the c-statistics for the full iBox Scoring System (0.809 [CI 0.791 to 0.827]). The full iBox Scoring System will be focused on for the remainder of internal validation for the reasons stated in section 6.5.

Table 42. C-statistics for the full qualification derivation dataset (censored at seven years post-evaluation)

Dataset	C-statistics for full iBox Scoring System (SE)	C-statistics for abbreviated iBox Scoring System (SE)
	Qualification derivation	
Loupy et al., 2019	0.809 (0.01)	0.803 (0.01)

To test whether the baseline hazard function is different between treatment centers, a Cox PH model was built and stratified by treatment center to assess whether a model stratified by treatment center had a significantly different c-statistic than the non-stratified model. The non-stratified model had a c-statistic of 0.809 (S.E. = 0.009), and the stratified model had a c-statistic of 0.817 (S.E. = 0.009). The two c-statistics are within one SE of each other, suggesting the two are not significantly different at discriminating between higher and lower risk subjects and suggesting that the baseline hazard function is not different between centers. The eight predictors in the full iBox Scoring System remained independently predictive of death-censored allograft survival (p-values given in Table 43), further suggesting that including the center effect did not appreciably change the full iBox Scoring System, and the model without center effect is sufficient. Comparison between parameter estimates for the non-stratified and stratified models shows that stratifying by treatment center has little effect on model parameters, as shown in the Table 43.

Table 43. Comparison between non-stratified and stratified by treatment center - Full iBox Scoring System model shows that center effect has little effect

	Full iBox Scoring System					
Variable	**Non-stratified**			**Stratified by treatment center**		
	HR	**95% CI**	**P-value**	**HR**	**95% CI**	**P-value**
Time from transplant to evaluation	1.082	(1.027 to 1.141)	0.003	1.077	(1.021 to 1.137)	0.007
eGFR (ml/min/1.73m²)	0.955	(0.949 to 0.961)	<0.001	0.954	(0.948 to 0.96)	<0.001
Log transformed UPCR proteinuria (g/g)	1.502	(1.393 to 1.62)	<0.001	1.517	(1.406 to 1.637)	<0.001

IFTA score	0-1	1	NA	NA	1	NA	NA
	2	1.145	(0.92 to 1.425)	0.226	1.292	(1.033 to 1.615)	0.025
	3	1.409	(1.104 to 1.8)	0.006	1.737	(1.342 to 2.248)	<0.001
Microcirculation Inflammation **(g score and ptc score)**	0-2	1	NA	NA	1	NA	NA
	3-4	1.433	(1.11 to 1.851)	0.006	1.488	(1.148 to 1.929)	0.003
	5-6	1.837	(1.25 to 2.698)	0.002	2.048	(1.388 to 3.024)	<0.001
Interstitial inflammation and tubulitis **(i score and t score)**	0-2	1	NA	NA	1	NA	NA
	≥ 3	1.335	(1.06 to 1.68)	0.014	1.347	(1.068 to 1.698)	0.012
Transplant glomerulopathy **(cg score)**	0	1	NA	NA	1	NA	NA
	≥ 1	1.469	(1.137 to 1.9)	0.003	1.483	(1.142 to 1.924)	0.003
DSA MFI	< 1400	1	NA	NA	1	NA	NA
	≥ 1400	1.837	(1.445 to 2.335)	<0.001	1.884	(1.476 to 2.406)	<0.001

To confirm the full iBox Scoring System performs well in different clinically relevant situations, various scenarios and subpopulations were examined for their c-statistic using the iBox Scoring System (Table 44). The full iBox Scoring System showed a good ability to discriminate between higher and lower risk subjects for all scenarios and subpopulations, with c-statistic values ranging from 0.76 to 0.87.

Three subsets showed significantly different c-statistic values from the qualification derivation dataset (i.e., all 3,941 subjects for the full iBox Scoring System) c-statistic of 0.809 (Table 44). Subjects evaluated after the first-year post-transplant were significantly improved (c-statistic of 0.843, 95% CI from 0.817 to 0.869). Comparatively, the following subjects had lower but still good c-statistic values: elderly (aged 60 or older) donors (c-statistic of 0.777, 95% CI from 0.746 to 0.808) and hypertensive donors (c-statistic of 0.771, 95% CI from 0.737 to 0.805). Since the full iBox Scoring System has good c-statistics for all clinically relevant subpopulations, no subpopulations are excluded from the proposed COU.

Importantly, the subset for the proposed COU for the iBox Scoring System (i.e., evaluation at one-year post-transplant ± 28 days and censored at five-years and 28 days post-transplant) shows a good c-statistic value of 0.849 (95% CI from 0.804 to 0.893), suggesting the iBox Scoring System discriminates appropriately among subjects who meet the proposed COU (red box in Table 44). Furthermore, the high c-statistic value for subjects on an mTORi (0.872, 95% CI from 0.808 to 0.936) suggests that the full iBox Scoring System can discriminate accurately even in CNI-free subjects. Table 44 shows the derivation subsets' c-statistics and suggests that the full iBox Scoring System performs well in various clinically relevant scenarios and subpopulations. Confidence intervals calculated as c-statistic ± SE × 1.96.

Table 44. mTORi derivation subset c-statistic for the full iBox Scoring System in various clinically relevant scenarios and subpopulations

Subset	C-statistic	SE	95% CI	Subjects	Events
COU: One-year post-transplant censored at 5 years post-transplant and 28 days	0.849	0.023	0.804 to 0.893	1,174	67
Stable subjects (protocol biopsy)	0.812	0.024	0.765 to 0.858	1,160	85
Unstable subjects (for-cause biopsy)	0.796	0.011	0.776 to 0.817	2,781	453
One-year post-transplant ± 28 days	0.827	0.020	0.788 to 0.867	1,174	105
After first-year post-transplant	0.843	0.013	0.817 to 0.869	1,641	247
Living donors	0.812	0.034	0.746 to 0.877	662	51
Deceased donors	0.803	0.010	0.783 to 0.823	3,279	487
Sensitized recipients	0.798	0.020	0.759 to 0.836	715	121
Non-sensitized recipients	0.809	0.011	0.788 to 0.831	3,226	417
Anti-IL2 receptor induction subjects	0.787	0.016	0.755 to 0.818	1,621	206

Anti-thymocyte globulin induction subjects	0.826	0.012	0.803 to 0.850	2,069	308
Male recipients	0.818	0.011	0.796 to 0.841	2,416	329
Female recipients	0.796	0.016	0.764 to 0.828	1,525	209
Male donors	0.799	0.013	0.774 to 0.825	2,123	293
Female donors	0.824	0.014	0.797 to 0.852	1,818	245
Elderly (60 or older) donors	0.777	0.016	0.746 to 0.808	1,291	224
Non-elderly (age < 60) donors	0.815	0.013	0.790 to 0.840	2,650	314
ECD donors	0.781	0.015	0.752 to 0.810	1,387	258
Non-ECD donors	0.810	0.014	0.783 to 0.837	2,549	279
Obese subjects	0.760	0.033	0.694 to 0.825	327	55
Non-obese subjects	0.813	0.010	0.793 to 0.833	3,432	453
Obese donors	0.792	0.024	0.745 to 0.840	519	86
Non-obese donors	0.811	0.010	0.790 to 0.831	3,414	451
DSA at day 0	0.798	0.020	0.759 to 0.836	715	121
No DSA at day 0	0.809	0.011	0.788 to 0.831	3,226	417
Diabetic donors	0.769	0.041	0.688 to 0.850	228	38
Non-diabetic donors	0.810	0.010	0.790 to 0.830	3,578	481

Hypertensive donors	0.771	0.017	0.737 to 0.805	990	190
Non-hypertensive donors	0.812	0.012	0.787 to 0.836	2,857	334
Prior kidney transplant subjects	0.823	0.018	0.788 to 0.857	596	125
No prior kidney transplant subjects	0.803	0.011	0.781 to 0.825	3,345	413
mTORi subjects (includes subjects on both mTORi and CNI therapies)	0.872	0.033	0.808 to 0.936	239	33
mTORi-only subjects	0.858	0.039	0.781 to 0.935	171	23

6.5.2 External validation

6.5.2.1 External validation on the qualification datasets

External validation was performed using the four external qualification datasets: Mayo Clinic Rochester and Helsinki University Hospital observational transplant centers, and BMS' BENEFIT and BENEFIT-EXT RCTs. Analysis for these qualification validation datasets was restricted to the proposed COU, so only patients with full and abbreviated iBox Scoring System evaluations at one-year ± 28 days were retained for analysis, and data were censored at five-years and 28 days post-transplant. The discrimination ability of the full and abbreviated iBox Scoring System models on each dataset was evaluated using Harrell's c-statistic (Harrell, Lee, and Mark 1996) censored at five-years plus 28 days post-transplant. A c-statistic value of 0.7 indicates good ability to discriminate between higher and lower-risk recipients, a value of 0.5 indicates no discriminatory ability, and a value of 1.0 indicates perfect discriminatory ability (Collett 2015).

All c-statistic values in Table 45 are 0.70 or greater for each qualification validation dataset, indicating good discriminatory ability. C-statistic values were found using the concordance function from the survival R package (Therneau 2020).

Table 45. Five-year full and abbreviated iBox Scoring System c-statistic values for the qualification validation datasets

Dataset	C-statistic for full iBox Scoring System (SE)	C-statistic for abbreviated iBox Scoring System (SE)
Mayo Clinic Rochester	0.93 (0.03)	0.84 (0.05)
Helsinki University Hospital	0.78 (0.06)	0.77 (0.06)

BENEFIT RCT	0.70 (0.09)	0.70 (0.08)
BENEFIT-EXT RCT	0.81 (0.07)	0.78 (0.06)

The BENEFIT RCT in particular appears to have a lower c-statistic value than the other datasets. A brief exploration of the data showed that BENEFIT has the highest average eGFR of all datasets (Table 11), with two patients that experience a graft loss having a one-year eGFR value of over 100 ml/min/1.73m^2. Specifically, these two patients have eGFR values of 100.1 and 111.3 ml/min/1.73m^2, UPCR of 0.13 and 0.13 g/g, presence and absence of DSA, and experienced a graft loss at about 4.4 and 2.1 years, respectively. Given that there are only twelve graft loss events in BENEFIT for the full and fifteen for the abbreviated iBox Scoring Systems, these two extreme values in graft loss patients result in a notably lower c-statistic value; removal of these two patients changes the c-statistic to 0.82. In contrast, Helsinki also has two graft loss patients with eGFR values over 100 (specifically, eGFR of 102.8 and 109.8 ml/min/1.73m^2, UPCR of 0.13 and 0.36 g/g, absence of DSA for both, and graft losses at 1.4 and 2.2 years, respectfully), but Helsinki has 21 total events for both iBox Scoring System versions, so these two patients have a comparatively smaller impact on the c-statistic value. Also notable is that the Mayo Clinic Rochester c-statistic drops from 0.93 with the full iBox Scoring system to 0.84 with the abbreviated version (Table 45). This is likely because there are two added graft loss events with the abbreviated version in the Mayo Clinic Rochester dataset and both have values indicating lower graft loss risk (eGFR of 66.0 and 84.0 ml/min/1.73m^2, UPCR of 0.053 and 0.21, presence and absence of DSA, and graft losses at 4.2 and 4.1 years, respectively). Moreover, the non-event patients with only an abbreviated iBox Scoring System evaluation at one-year in Mayo Clinic Rochester have somewhat low eGFR values (mean of 52 ml/min/1.73m^2), which may also contribute to the lower c-statistic with the abbreviated version. These findings highlight the importance of eGFR for accurate predictions with the iBox Scoring System.

The datasets from the BENEFIT and BENEFIT-EXT RCT studies included a substantial proportion of CNI-free (BELA based) regimens. The results suggest that the full and abbreviated iBox Scoring System models are capable of accurately discriminating between high and low risk transplant recipients not only for CNI transplant recipient, but also for datasets including CNI-free transplant recipients.

To assess the consistency of the discriminatory ability of the full and abbreviated iBox Scoring System models across treatments using higher sample sizes, c-statistics were calculated for a CNI recipient pool and a CNI-free recipient pool. The CNI pool comprised the Mayo Clinic Rochester subjects receiving CNI regimens, all the Helsinki University Hospital subjects, and the CsA subjects from the two BMS RCTs, while the CNI-free pool comprised the mTORi subjects from Mayo Clinic Rochester and the BELA subjects from the two BMS RCTs. Subjects receiving both CNI and mTORi or neither in the Mayo Clinic Rochester dataset were not included. For a further breakdown of drug regimens, see Data 4.3.3 (Qualification derivation dataset) and Data 4.3.8.1 (Therapeutics across qualification datasets).

The c-statistics for these combined datasets are greater than 0.7 for CNI and CNI-free subjects for both the full and abbreviated iBox models (Table 46), suggesting overall good discriminatory ability regardless of treatment type. The full and abbreviated iBox Scoring System models also showed good performance for both CNI therapies, TAC and CsA (Table 46). The CNI-free subjects could not be split into mTORi and BELA because there were so few mTORi subjects (n = 38) that no events occurred in that group, making c-statistic calculation impossible, and all mTORi subjects were missing both biopsy and DSA MFI data. As a result, the c-statistics for the BELA subjects are essentially the c-statistics for the CNI-free subjects.

Table 46. Five-year full and abbreviated iBox Scoring System models c-statistic values for CNI and CNI-free subjects

Subject Regimen	Full iBox Scoring System (SE)	Abbreviated iBox Scoring System (SE)
CNI (TAC, CsA)	0.82 (0.04) [TAC 0.86 (0.05), CsA 0.77 (0.05)]	0.79 (0.04) [TAC 0.81 (0.05), CsA 0.77 (0.05)]
CNI-free	0.75 (0.08)	0.73 (0.07)

Model calibration was evaluated for the Mayo Clinic Rochester and Helsinki University Hospital observational datasets, and the BENEFIT and BENEFIT-EXT RCTs. Calibration was assessed using the Poisson "fit1" method for calibration (Crowson, Atkinson, and Therneau 2016), which evaluates whether observed events in the data match the number of events predicted from the iBox Scoring System model using Poisson regression (for more details, see Methods 5.4.1.4.2). Tables for calibration for the full and abbreviated iBox Scoring System models are shown below (Table 47 and Table 48, respectively), and indicate whether observed events in the data and predicted events from the iBox Scoring System match within a reasonable margin of error. These results suggest both the full and abbreviated iBox Scoring System model performs well on the examined external datasets; however, due to low numbers of events in individual datasets, error bars are relatively wide (Figure 14 and Figure 15 for full and abbreviated iBox Scoring System models, respectively).

Table 47. Poisson calibration results for the full iBox Scoring System. Z-scores and p-values were calculated from a Poisson regression model

Dataset	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	Observed /Predicted	z score for Observed /Predicted	P-value
Combined observational	827	39	38.74	1.01	0.04	0.97
Helsinki University Hospital	344	21	14.40	1.46	1.73	0.08
Mayo Clinic Rochester	483	18	24.34	0.74	-1.28	0.20
Combined RCTs	676	24	29.49	0.81	-1.01	0.31
BENEFIT RCT	416	12	14.52	0.83	-0.66	0.51
BENEFIT-EXT RCT	260	12	14.97	0.80	-0.77	0.44

Table 48. Poisson calibration results for the abbreviated iBox Scoring System. Z-scores and p-values were calculated from a Poisson regression model

Dataset	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	Observed /Predicted	z score for Observed /Predicted	P-value
Combined observational	841	41	40.61	1.01	0.06	0.95
Helsinki University Hospital	344	21	16.19	1.30	1.19	0.23
Mayo Clinic Rochester	497	20	24.41	0.82	-0.89	0.37
Combined RCTs	872	38	41.74	0.91	-0.58	0.56
BENEFIT RCT	515	15	18.77	0.80	-0.87	0.39
BENEFIT-EXT RCT	357	23	22.97	1.00	0.01	1.00

Ordinarily, calibration is visualized by splitting the data into different risk groups and evaluating observed versus predicted events in each risk group. Due to the low numbers of events and wide error bars, splitting the data into different risk groups would have depleted event numbers too greatly for some groups and resulted in error bars too wide for reliable inference. Instead, the TTC visualized calibration by plotting the survival curve from an idealized perfect model where predicted events match observed exactly (identity line) against the predicted survival from the Poisson model fit (red line, Figure 14 and Figure 15). These visualizations showed that the full and abbreviated iBox Scoring System models predicted survival falls within a reasonable margin of error of the idealized perfect model.

CNI and CNI-free subjects, as described previously, were also evaluated for their calibration. Results suggest that the full and abbreviated iBox Scoring System models have reasonable calibration (i.e., predicted numbers of events matched to observed events within a margin of error) for both CNI and CNI-free subjects, as described in Table 49).

Table 49. Poisson calibration results for CNI and CNI-free subjects

Subject regimen	Full iBox Scoring System				Abbreviated iBox Scoring System			
	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	P-value	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	P-value
CNI	1045	50	51.6	0.82	1124	61	58.9	0.78
CNI-free	456	13	16.6	0.38	587	18	23.4	0.26

Model validation suggests that the full and abbreviated iBox Scoring System models are effective in diverse populations. Importantly, the models have reasonable discrimination and calibration in subjects on CNI-free regimens despite being trained primarily on subjects receiving CNI therapies. To investigate this further in RCTs specifically, the two RCTs

(BENEFIT and BENEFIT-EXT) were broken down into the CsA (CNI) and BELA (CNI-free) subjects. Similar results were found suggesting good discrimination and calibration in each treatment type (Table 50 and Table 51 for discrimination and calibration, respectively). Additionally, results suggest that removing the biopsy variables has little impact on model discrimination and calibration.

Table 50. Five-year iBox c-statistics for CNI and CNI-free subjects in combined BENEFIT and BENEFIT-EXT RCTs

RCT Subject Regimen	Full iBox Scoring System (SE)	Abbreviated iBox Scoring System (SE)
CNI (CsA)	0.75 (0.08)	0.75 (0.07)
CNI-free (BELA)	0.75 (0.08)	0.73 (0.07)

Table 51. Poisson calibration results for CNI and CNI-free subjects in combined RCTs

Subject regimen	Full iBox Scoring System				Abbreviated iBox Scoring System			
	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	P-value	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	P-value
CNI (CsA)	220	11	12.9	0.59	285	20	18.3	0.69
CNI-free (BELA)	456	13	16.6	0.38	587	18	23.4	0.26

These validation analyses indicate that the full and abbreviated iBox Scoring System models can accurately predict the number of events and discriminate between higher and lower risk subjects in diverse datasets. Importantly, despite the full and abbreviated iBox Scoring System models being trained on primarily CNI subjects, model discrimination and calibration is reasonably good in both CNI and CNI-free populations.

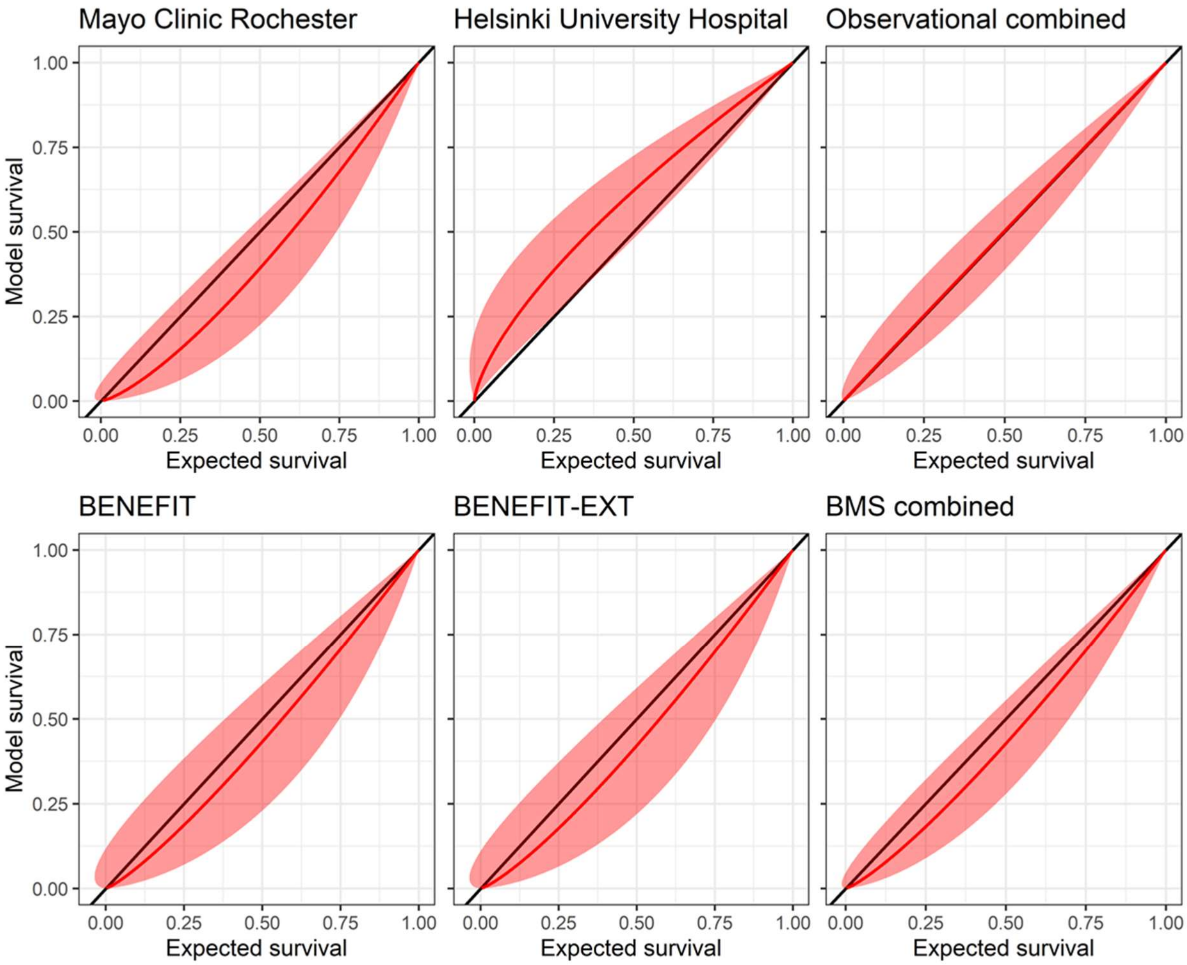


Figure 14. Five-year calibration plot for full iBox Scoring System.

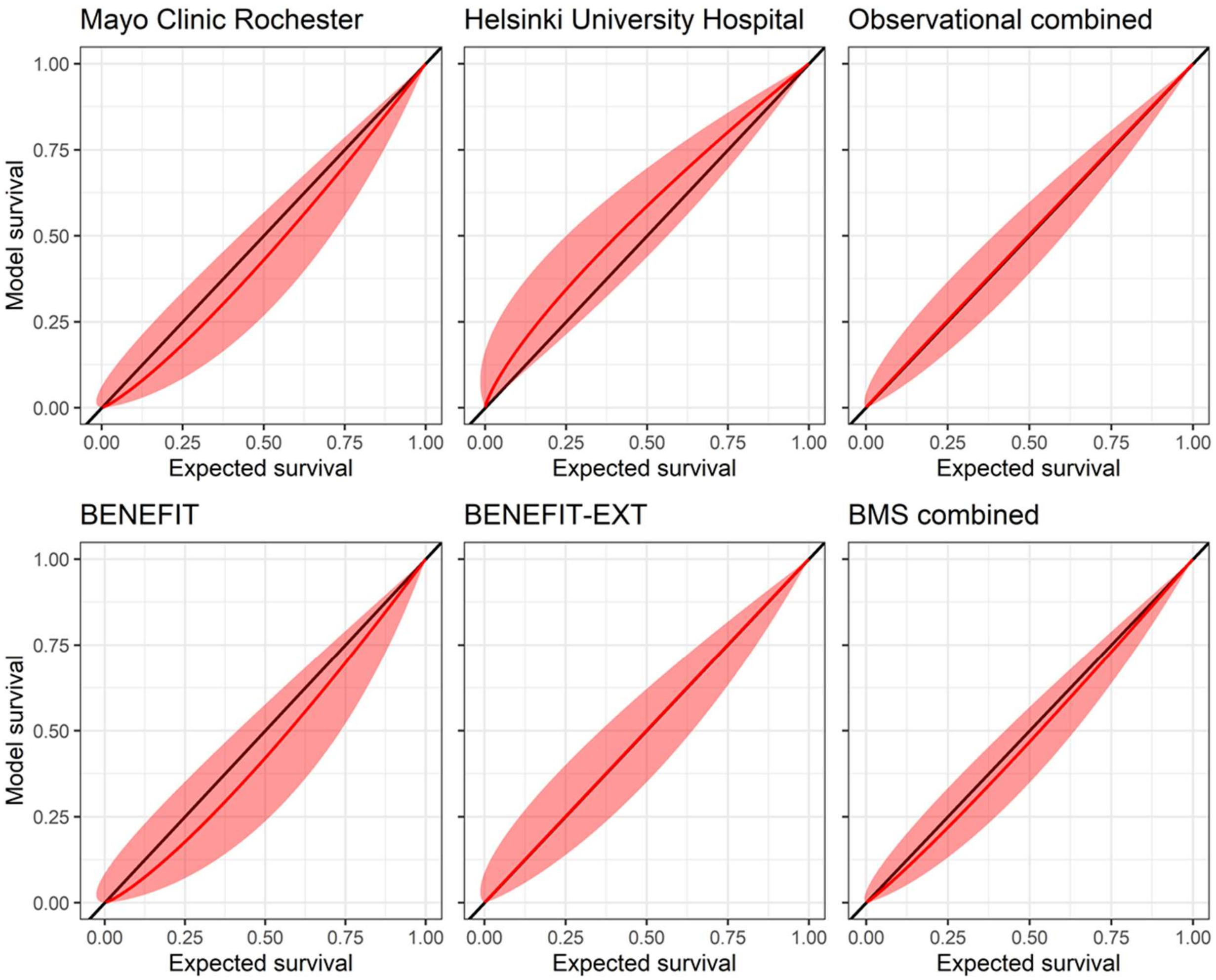


Figure 15. Five-year calibration plot for abbreviated iBox Scoring System.

Five-year calibration plots show alignment between predicted and observed survival curves, as seen in Figure 14 and Figure 15 for both full and abbreviated iBox Scoring System models. Model survival is shown (red line) compared to the null survival expectation if model predictions and observed events match exactly (black identity line). Confidence regions (red shaded area) were generated from the Poisson model fit by solving for the survival curves and applying propagation of error to get the SE. Results show that the black identity line is within the confidence band around the model estimate (red line), suggesting model predicted survival falls within a reasonable margin of error of observed survival.

6.5.2.2 External validation on the European cohort and three randomized controlled trials from Loupy et al., 2019

External validation was previously performed using the three European centers in Loupy et al., 2019 that were part of the European validation cohort: Hôpital Hôtel Dieu, Nantes, France; Hospices Civils, Lyon, France; and the University Hospitals, Leuven, Belgium. Additionally, external validation was previously performed using the three RCTs, CERTITEM by Rostaing et al., 2015, RITUX ERAH by Sautenet et al., 2016, and BORTEJECT by Eskandary et al., 2017, also part of Loupy et al., 2019. Of particular interest is the c-statistic value of CERTITEM RCT, a *de novo* phase III IST minimization trial.

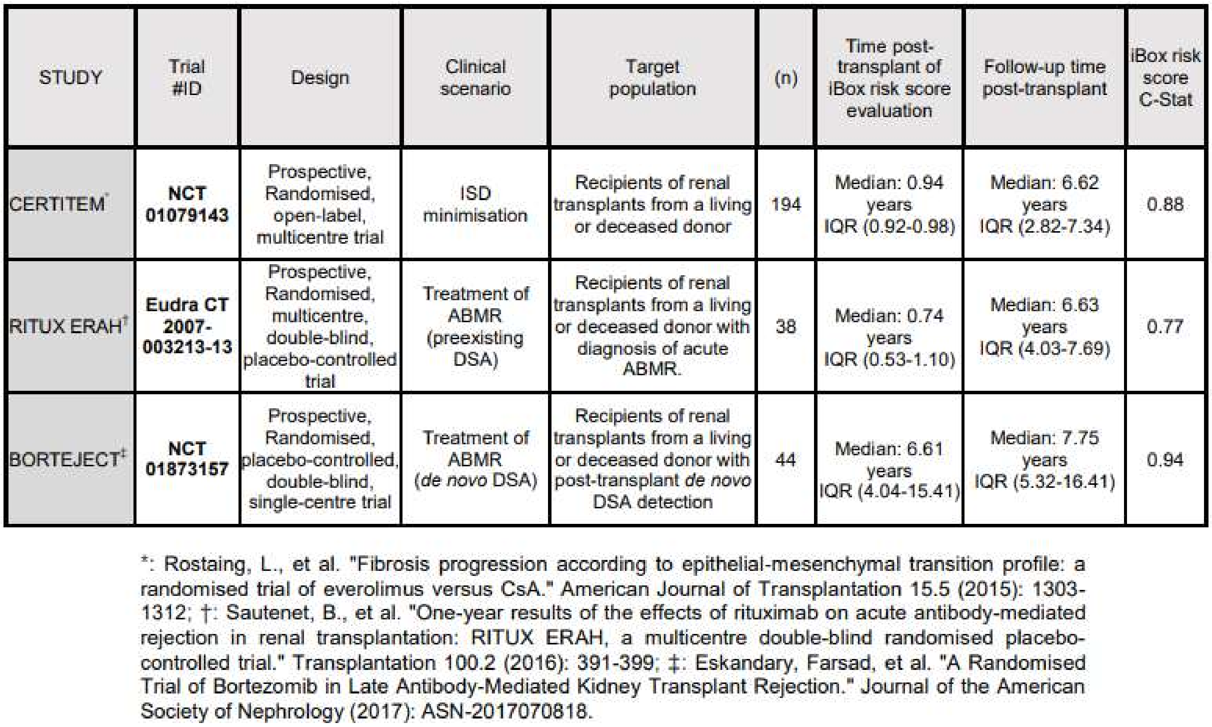
All c-statistic values in Table 52 and Table 53 are 0.70 or greater for each Loupy et al., 2019 external validation datasets, indicating good discriminatory ability.

Table 52. Seven-year iBox c-statistic value for the European validation cohort from Loupy et al., 2019

Dataset	C-statistic (CI)*
European validation cohort	0.81 (0.78 to 0.84)

* CI used instead of SE consistent with Loupy et al., 2019

Table 53. iBox Scoring System c-statistic values for the three RCTs in Loupy et al., 2019.



STUDY	Trial #ID	Design	Clinical scenario	Target population	(n)	Time post-transplant of iBox risk score evaluation	Follow-up time post-transplant	iBox risk score C-Stat
CERTITEM[*]	**NCT 01079143**	Prospective, Randomised, open-label, multicentre trial	ISD minimisation	Recipients of renal transplants from a living or deceased donor	194	Median: 0.94 years IQR (0.92-0.98)	Median: 6.62 years IQR (2.82-7.34)	0.88
RITUX ERAH[†]	**Eudra CT 2007-003213-13**	Prospective, Randomised, multicentre, double-blind, placebo-controlled trial	Treatment of ABMR (preexisting DSA)	Recipients of renal transplants from a living or deceased donor with diagnosis of acute ABMR.	38	Median: 0.74 years IQR (0.53-1.10)	Median: 6.63 years IQR (4.03-7.69)	0.77
BORTEJECT[‡]	**NCT 01873157**	Prospective, Randomised, placebo-controlled, double-blind, single-centre trial	Treatment of ABMR (*de novo* DSA)	Recipients of renal transplants from a living or deceased donor with post-transplant *de novo* DSA detection	44	Median: 6.61 years IQR (4.04-15.41)	Median: 7.75 years IQR (5.32-16.41)	0.94

*: Rostaing, L., et al. "Fibrosis progression according to epithelial-mesenchymal transition profile: a randomised trial of everolimus versus CsA." American Journal of Transplantation 15.5 (2015): 1303-1312; †: Sautenet, B., et al. "One-year results of the effects of rituximab on acute antibody-mediated rejection in renal transplantation: RITUX ERAH, a multicentre double-blind randomised placebo-controlled trial." Transplantation 100.2 (2016): 391-399; ‡: Eskandary, Farsad, et al. "A Randomised Trial of Bortezomib in Late Antibody-Mediated Kidney Transplant Rejection." Journal of the American Society of Nephrology (2017): ASN-2017070818.

6.6 Supplementary analyses

Proteinuria conversions (section 6.6.1), competing risk (section 6.6.2), and TLS and treatment effect analyses (Section 6.6.3) supplementary analyses were conducted to support the qualification effort, as discussed in Modeling analysis methodologies (Supplementary analyses) of the Briefing Dossier.

6.6.1 Proteinuria conversions

The results for the three proteinuria measurements (i.e., UACR, dipstick proteinuria, and 24-hour proteinuria) that required additional supporting evidence for use in the full and abbreviated iBox Scoring System models are shown in the following sections:

- UACR to UPCR: A full and detailed description of the conversion can be found in 6.6.1.1.
- Dipstick proteinuria to UPCR: A full and detailed description of the conversion can be found in 6.6.1.2.
- 24-hour proteinuria to UPCR: A full and detailed description of the rationale to support no conversion can be found in 6.6.1.3.

6.6.1.1 Converting urine albumin-to-creatinine ratio to urine protein-to-creatinine ratio

Weaver et al., 2020 (Weaver et al. 2020) developed equations to estimate median UACR from UPCR values on the log scale in a population-based cohort of 47,714 adults in Alberta, Canada, who had simultaneous assessments of UACR and UPCR. Raw UACR measurements were expressed in mg/g. The inverse of the proposed piecewise linear equation was used for the median described in Table 3 of Weaver et al., 2020. In addition, the model with log(UPCR) only as a covariate was selected due to its parsimonious nature and that models with additional covariates had similar predictive performance (Table 2 of Weaver et al., 2020). This yielded the following equation, which was used to generate estimated UPCR values based on ranges of UACR input values (bolded):

Equation 8. Formulas used to generate estimated UPCR values based on ranges for UACR input values

UACR < 4.13: Estimated UPCR = $e^{[-7.5301 and 7.9144 * ln(UACR)]}$

4.13 ≤ UACR < 5.54 Estimated UPCR = $e^{[1.7333 and 1.3791 * ln(UACR)]}$

5.54 ≤ UACR < 107.07 Estimated UPCR = $e^{[3.2691 and 0.4819 * ln(UACR)]}$

728.99 ≤ UACR < 728.99 Estimated UPCR = $e^{[2.1432 and 0.7229 * ln(UACR)]}$

728.99 ≤ UACR Estimated UPCR = $e^{[0.02496 and 1.0442 * ln(UACR)]}$

As the above formulas produce UPCR measurements with units of mg/g by default, all values are divided by 1,000 to produce units of g/g (consistent with the UPCR units in the qualification derivation dataset).

The coefficient of quantile correlation from the above equation was sqrt (0.623) = 0.79, which is moderately high. However, because the TTC did not have access to the raw data, it was not possible to model UPCR directly; this was the best (only) available formula for conversion.

6.6.1.2 Converting dipstick proteinuria to UPCR

The German cohort from Charité – Universitätsmedizin Berlin comprising 1387 subjects with 6169 dipstick and UPCR values were used to develop an algorithm for converting the dipstick proteinuria categorical results to continuous UPCR values. The median UPCR values per dipstick category, as shown in Figure 16, were used for conversion since they provided a better representation of the central location of the data points.

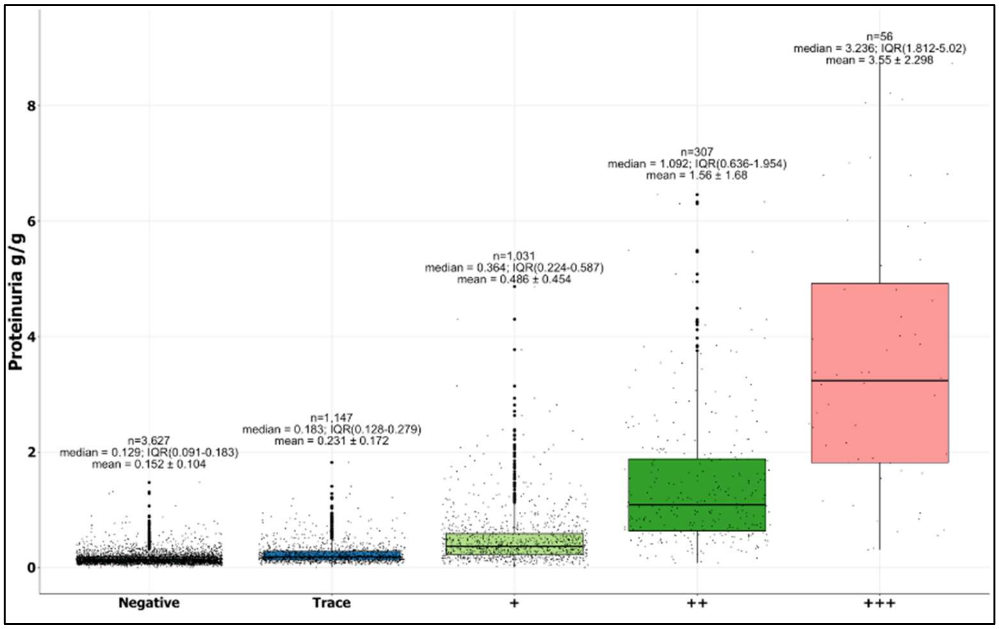


Figure 16. Association of dipstick proteinuria with UPCR in the German cohort.

The datasets used for external validation that include dipstick proteinuria values were Helsinki University Hospital, and the two RCTs, BENEFIT and BENEFIT-EXT. Trace dipstick result was present in the 1387 German cohort, as well as in the BENEFIT and BENEFIT-EXT RCTs. However, trace dipstick result was not present in the Helsinki University Hospital dataset. This algorithm for converting dipstick proteinuria to UPCR was applied in the Helsinki University Hospital cohort without consideration for trace proteinuria, consistent with the dipstick proteinuria assay capabilities. For consistency between datasets, the value of zero from Helsinki University Hospital has been equated to "negative" from BENEFIT and BENEFIT-EXT RCTs. The dipstick-imputed median UPCR values represented in Table 54 were used in the calculation of an iBox score for these qualification validation datasets. Subjects in the qualification validation datasets were assigned a UPCR value based on Table 54. A summary of the distribution of dipstick proteinuria data across external qualification validation cohorts for the full and abbreviated iBox Scoring System models is in Table 55 and Table 56, respectively.

Table 54. Dipstick proteinuria to UPCR proteinuria association

Dipstick result	Log transformed UPCR value (g/g)
Negative	0.129 IQR (0.091-0.183)
Trace	0.183 IQR (0.128-0.279)
+	0.364 IQR (0.224-0.587)
++	1.092 IQR (0.636-1.954)
+++	3.236 IQR (1.812-5.02)

Table 55. Distribution of dipstick proteinuria data across qualification validation datasets for full iBox Scoring System

	BENEFIT RCT	BENEFIT-EXT RCT	Helsinki University Hospital
	One year ± 28 days		
Negative	306 (73.56%)	160 (61.54%)	286 (83.14%)
Trace	55 (13.22%)	41 (15.77%)	Not available
+	36 (8.65%)	42 (16.15%)	48 (13.95%)
++	16 (3.85%)	11 (4.23%)	6 (1.74%)
+++	3 (0.72%)	6 (2.31%)	4 (1.16%)
Total	416	260	344

Table 56. Distribution of dipstick proteinuria data across qualification validation datasets for abbreviated iBox Scoring System

	BENEFIT RCT	BENEFIT-EXT RCT	Helsinki University Hospital
	One year ± 28 days		
Negative	374 (72.62%)	215 (60.22%)	286 (83.14%)
Trace	70 (13.59%)	56 (15.69%)	Not available
+	46 (8.93%)	59 (16.53%)	48 (13.95%)
++	20 (3.88%)	18 (5.04%)	6 (1.74%)
+++	5 (0.97%)	9 (2.52%)	4 (1.16%)
Total	515	357	344

6.6.1.3 24-hour proteinuria required no conversion to UPCR

24-hour proteinuria values (g/day) to log transformed UPCR (g/g) required no conversion as the two are approximately equal, as supported by literature precedent and clinical practice.

Ginsberg et al., 1983 (Ginsberg et al. 1983) used RCT data to have the following conclusion: "We conclude that the determination of the protein/creatinine ratio in single urine samples obtained during normal daylight activity, when properly interpreted by taking into consideration the effect of different rates of creatinine excretion, can replace the 24-hour urine collection in the clinical quantitation of proteinuria." This conclusion was confirmed in the Price et al., 2005 (Price, Newall, and Boyd 2005) publication. Price et al., 2005 conducted a systematic review of 16 studies and concluded that there were good correlations for UPCR and 24-hour proteinuria values (Table 2 of Price et al., 2005).

6.6.2 Competing risk analysis

Death is a competing risk to graft loss; failure to account for the risk of death could bias predictions of the risk of graft loss upward (Collett 2015). Such overprediction occurs when death is informative of graft loss and estimation techniques assume death is uninformative and censor deaths. If death is unrelated to graft loss (i.e., it is uninformative), censoring is reasonable because the fact that someone died gives no information as to whether a graft loss was likely to occur. But if death is related to graft loss (i.e., it is informative), then other methods must be employed that account for the risk of death directly. Death is typically assumed to be informative of graft loss over long time scales because it is a competing risk; that is, if death occurs, graft loss cannot occur. But the effect of death on graft loss may be negligible on shorter time scales. In such a case, the number of individuals at risk of graft loss will not be notably depleted by the number of deaths that occur, and Cox PH models like the full and abbreviated iBox Scoring System models that censor death should give accurate predictions of the risk of graft loss. The following section explores the relationship between graft loss and death in the full and abbreviated iBox Scoring System models to evaluate whether the models' predictions are biased by censoring deaths. Note that graft loss is the event of interest for this section and death is considered a separate competing event; all-cause graft loss is not considered.

To investigate whether the full and abbreviated iBox Scoring System models have overestimated the risk of graft loss due to failure to account for the risk of death, two approaches were taken using the derivation dataset. First, a plot of the CIFs for the probability of graft loss with and without accounting for the probability of death against the probability of death (see Figure 17) was generated from the qualification derivation cohort. This CIF plot showed that the probability of graft loss appears largely unaffected by the number of deaths that occur, as there appears to be only a slight correlation between increasing probability of death and divergence of the two graft loss CIFs, and the confidence bands remain overlapping (see Figure 17).

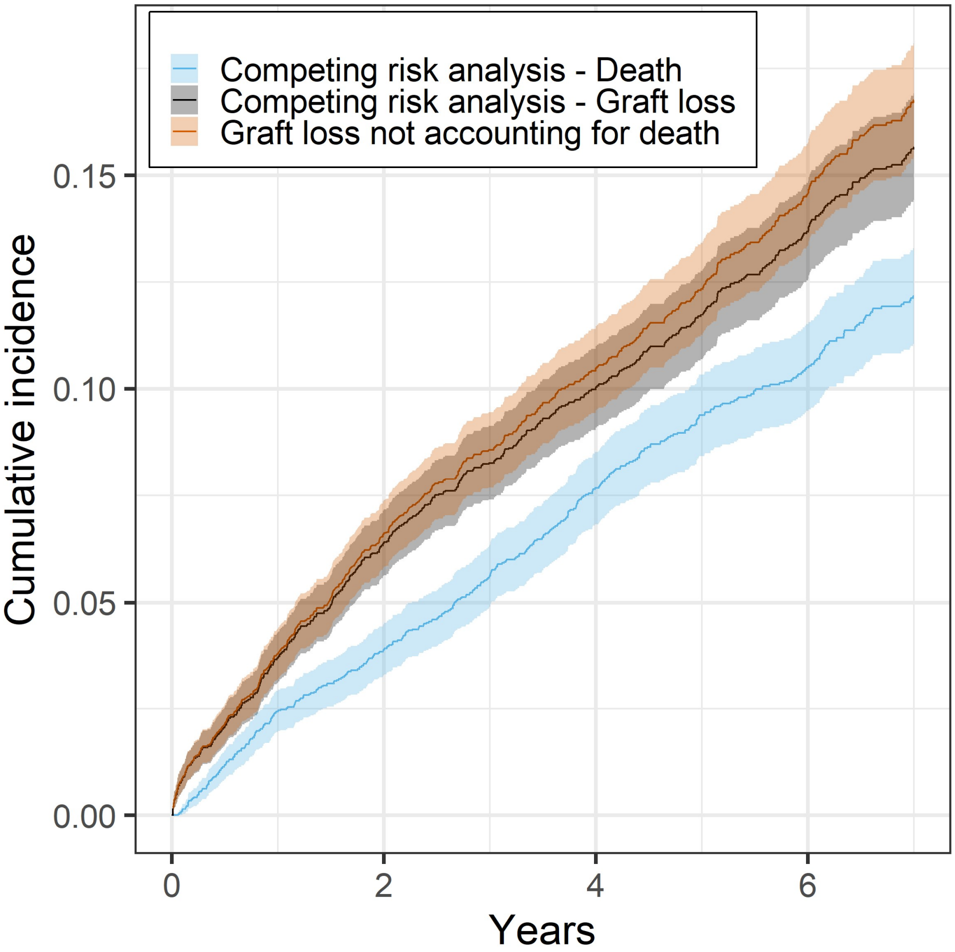


Figure 17. CIFs for the probability of graft loss with and without accounting for the probability of death against the probability of death in the qualification derivation cohort.

Figure 17. The incidence of graft loss is largely unaffected by death. Lines are the CIFs for the risk of graft loss over time with and without accounting for the competing risk of death (brown and black lines, respectively), and for the risk of death while accounting for graft loss (blue line). Shaded regions represent the CI for each CIF, calculated as $\pm 1.96 \times \mathrm{SE}$ (SE).

To verify that the calculated hazard of death-censored allograft loss from the full iBox Scoring System is largely unaffected by the number of deaths that occur, a subdistribution hazard function (Collett 2015) model of graft loss that accounts for the risk of death using the iBox covariates for the full iBox Scoring System model was built. The parameter estimates for the subdistribution model covariates all fell within the 95% CIs of the parameter estimates from the original iBox hazard model (see Figure 18), suggesting the iBox Scoring System's estimated hazard of graft loss is equivalent to the hazard from a similar model that accounts for death. Therefore, the predictions from the full iBox Scoring System do not appear to be biased by censoring deaths.

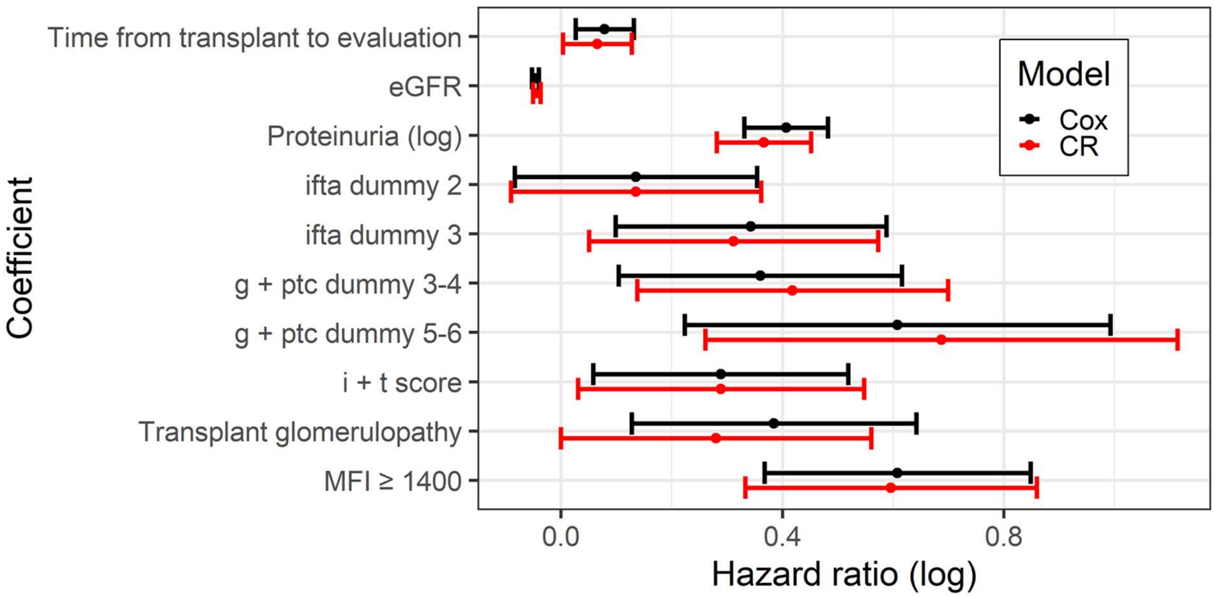


Figure 18. Parameter estimates from the competing risk model compared to the full iBox Scoring System parameter estimates.

Figure 18. Parameter estimates from the competing risk model fall within the CIs of full iBox Scoring System model estimates. Covariate names are listed on the Y-axis, while parameter estimates of the natural log of the HR are on the X-axis. Black points and error bars represent the iBox Scoring System parameter estimate and CI (calculated as $\pm 1.96 \times \mathrm{SE}$) from the Cox PH model. Red points and error bars are the competing risk model parameter estimates and CIs, respectively.

The results show that censoring deaths has little to no impact on predictions of graft loss in the derivation dataset. This is consistent with the external calibration results that suggest the full iBox Scoring System is not overpredicting the number of graft loss events that occur. These findings suggest that the iBox Scoring System model may be used to accurately assess the risk of graft loss without needing to account for the risk of death.

6.6.3 Trial-level surrogacy and treatment effect analyses

TLS analysis was performed to evaluate if a treatment effect on the full and abbreviated iBox Scoring System models is predictive of treatment effect on the true outcome, death-censored allograft survival in the context of an RCT. The TTC chose to execute TLS analysis based upon the feedback that Novartis received from EMA during the course of their sponsor-driven discussions. Novartis then shared this information with the TTC, as described in Background 3.1.1 (Regulatory history with EMA).

Previous examples of TLS analyses were examined before the execution of TTC's analyses. Two fundamental features are necessary to rigorously evaluate a surrogate endpoint's performance using historical clinical trials via TLS. First, the historical studies used in the TLS analysis are required to be of adequate power and sample size to demonstrate a statistically significant therapeutic effect on both the surrogate and the true outcome. Second, there must be several of these adequately sized and powered historical clinical trials to quantitatively describe the treatment effect relationship on the surrogate and the true outcome. Both requirements placed significant constraints on the TLS analysis executed for the full and abbreviated iBox Scoring System models. Historically, kidney transplant trials were designed for between one to three-year assessments and with appropriate power and sample size to assess the primary endpoints of death, graft loss, BPAR, and lost-to-follow-up. To the TTC's knowledge, no historical RCTs were prospectively designed to assess five-year graft survival, either death-censored or overall, and have all of the subject features necessary at one-year post-transplant (eGFR, UPCR, DSA, and/- biopsy) to evaluate the full and abbreviated iBox Scoring System models as a surrogate endpoint. This understanding placed significant constraints on data availability for the TLS analyses.

As discussed in the modeling analysis methodologies (section 5), three RCT datasets (two prospective and one retrospective) with CNI and CNI-free arms were used for three different versions of TLS analyses. The three different versions of the TLS analyses were performed for both iBox Scoring System models, as described below:

A. Analysis of five-year death-censored allograft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant. Full iBox Scoring System results (Section 6.6.3.1 Trial-level surrogacy analysis with full iBox Scoring System). Abbreviated iBox Scoring System results (Appendix-Revised Supporting Results).

B. Analysis of five-year death-censored allograft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant with the addition of subjects that died/withdrew/lost their graft within the first year of transplant Full iBox Scoring System results (Appendix: Revised-Supporting results). Abbreviated iBox Scoring System results (Appendix: Revised-Supporting results).

C. Analysis of five-year all-cause graft survival for subjects with full and abbreviated iBox Scoring System models at one-year post-transplant with the addition of subjects that died/withdrew/lost their graft within the first year of transplant. (6.8 All-cause allograft loss for iBox Scoring System).

In analyses B and C, the subjects who died/withdrew/lost their graft before the first year of transplantation have missing iBox score values. These subjects were assigned imputed iBox score values corresponding to the worst-case scenario, as previously described in Modeling analysis methodologies 5.5.3.1 (Imputation). Refer to Appendix: Revised-Supporting results for analysis and findings.

6.6.3.1 Trial-level surrogacy analysis with full iBox Scoring System

This section outlines the results of the TLS analysis for the full iBox Scoring System without imputation for death-censored allograft survival.

6.6.3.1.1 Step One: Computation of treatment effects

As a first step to perform the TLS analysis, treatment effect was estimated. Computation of the treatment effects was dependent on the data source as outlined above.

6.6.3.1.1.1 BENEFIT and BENEFIT-EXT RCTs

Due to the limited availability of RCTs, each of the two RCTs, BENEFIT and BENEFIT-EXT, were split up into pseudo trials based on their geographical regions to increase the number of pseudo trials required to apply the TLS method. There were three regions in each of the RCTs: Europe, USA and Other (this yielded a total of six pseudo trials). Table 57 shows the description of the events based on the region:

Table 57. Distribution of subjects and graft loss events per treatment arm for one-year observed iBox scores for full iBox Scoring System

		BENEFIT RCT (n = 416) CNI-free (BELA) n = 281 CNI (CsA) n = 135	BENEFIT-EXT RCT (n = 260) CNI-free (BELA) n = 85 CNI (CsA) n = 135
		Five-year follow up	**Five-year follow up**
Region/ Pseudo trial	**Treatment arm**	**No of subjects/ No of events**	**No of subjects/ No of events**
Europe	BELA	63/2	69/3
	CsA	30/2	33/2
Other	BELA	158/3	71/3
	CsA	74/4	36/0
USA	BELA	60/0	35/2
	CsA	31/1	16/2
Total events		12	12

There were 416 subjects with full iBox Scoring System evaluations at one-year post-transplant from the BENEFIT RCT. Out of the 416 subjects, 93 were from Europe, 232 from "Other" and 91 from USA. There were 12 graft losses in all regions at five years post-transplant. On average the survival rate was higher for subjects in the BELA arm compared to the CsA arm. No graft loss in the USA BELA arm was observed.

In the BENEFIT-EXT RCT, there were 260 subjects with full iBox Scoring System evaluations at one-year post-transplant. Out of the 260 subjects, 102 were from Europe, 107 from "Other" and 51 from USA. There were 12 graft losses in all regions at five years post-transplant. No graft loss was observed in the CsA arm of the "Other" region.

Figure 19 and Figure 20 show the distribution of iBox scores for the full iBox Scoring System in the BENEFIT and BENEFIT-EXT RCTs, respectively. Across all the regions within the BENEFIT RCT, the iBox scores ranged from -6.335 to -0.4173, while the scores ranged from -4.4261 to 0.3063 in the BENEFIT-EXT RCT. It was observed that, in general, subjects in the BELA arm had lower iBox scores compared to the CsA arm.

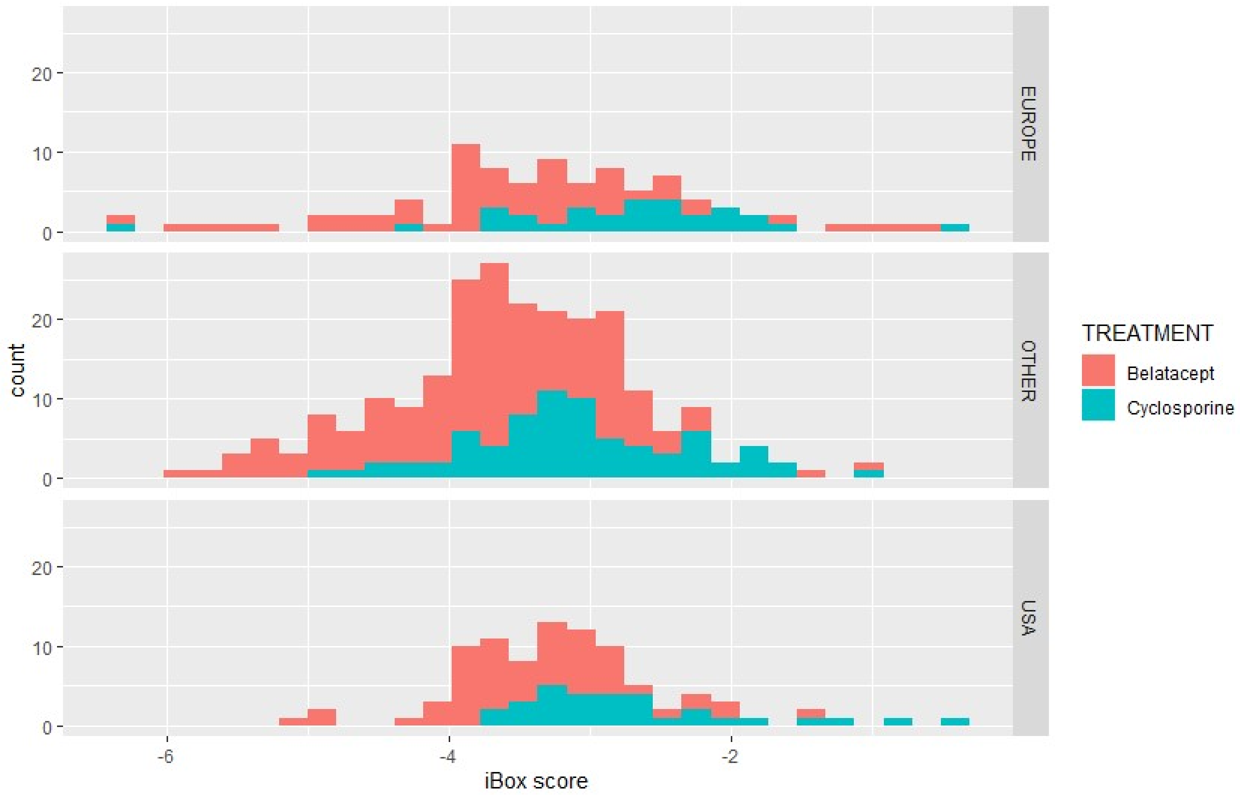


Figure 19. Distribution of observed iBox scores for the full iBox Scoring System in the BENEFIT RCT.

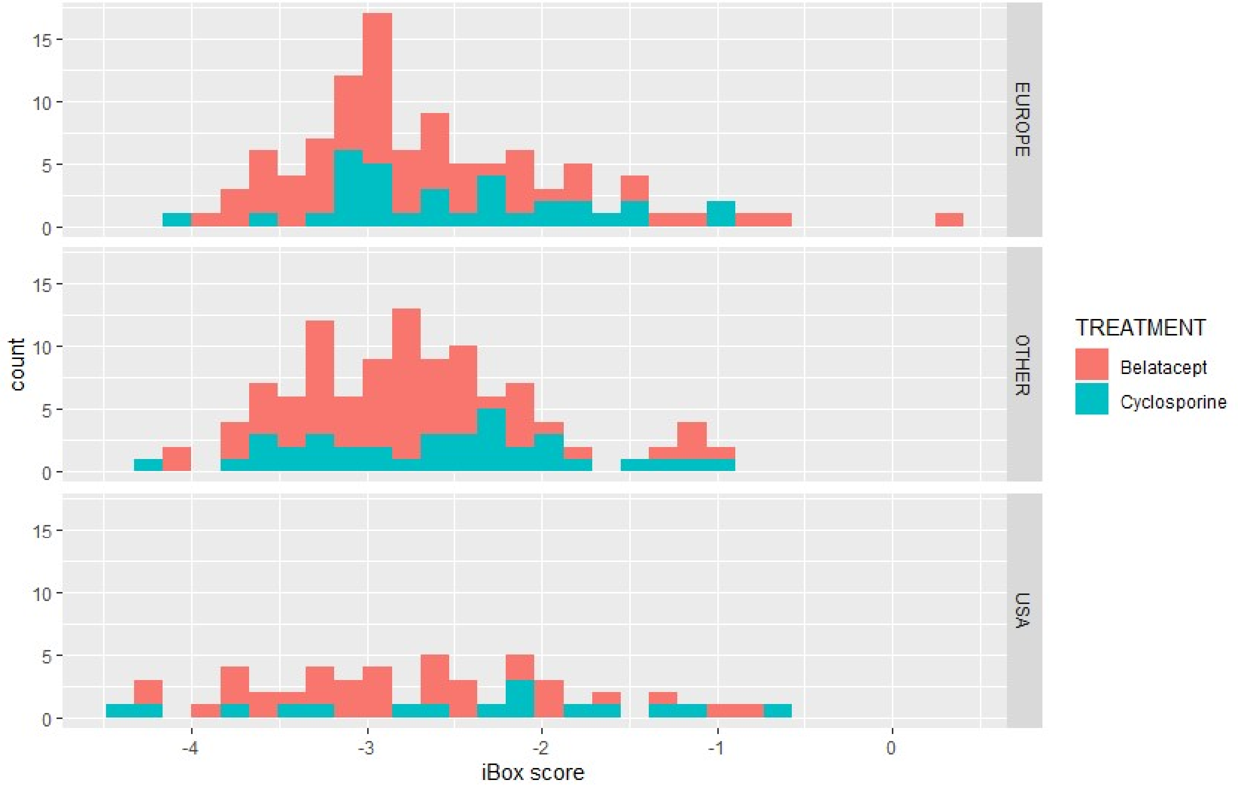


Figure 20. Distribution of observed iBox scores for the full iBox Scoring System in the BENEFIT-EXT RCT.

For each pseudo trial, the TTC computed the treatment effect and variance on the full iBox Scoring System without imputation based on the two-sample t-test, while the treatment effect (log-HR) and variance for five-year death-censored allograft survival were computed using the log-rank test. The correlation coefficient between the two treatment effects per pseudo trial was computed from 2000 bootstrap samples. Table 58 shows the step one results from the BENEFIT and BENEFIT-EXT RCTs.

Table 58. Treatment effects of the BENEFIT and BENEFIT-EXT RCTs for observed one-year iBox score for the full iBox Scoring System and five-year death-censored allograft survival

		Graft loss		iBox score		
Pseudo trial	**Treatment comparison**	**Log HR**	**SE**	**Mean difference**	**SE**	**Correlation**
			BENEFIT RCT			
Europe	BELA versus CsA	-0.9193	1.0868	-0.7375	0.2330	0.0307
USA	BELA versus CsA	-2.9355	2.1100	-0.7147	0.1692	0.2576

Other	BELA versus CsA	-1.2287	0.8180	-0.6350	0.1130	0.1161
BENEFIT-EXT RCT						
Europe	BELA versus CsA	-0.5186	0.9867	-0.2065	0.1582	0.2674
USA	BELA versus CsA	-0.8108	1.0695	-0.3275	0.3121	0.3054
Other	BELA versus CsA	1.5191	1.2174	-0.1854	0.1517	0.1921

For all the pseudo trials except the "Other" region of the BENEFIT-EXT RCT, a negative log-HR for five-year death-censored allograft survival was observed. This means the risk of graft loss was lower in the BELA arm compared to the CsA arm. However, this was not the case with the "Other" region of BENEFIT-EXT RCT, where the treatment effect was positive. In the full iBox Scoring System without imputation, a positive treatment effect (negative difference) was observed across all pseudo trials; subjects in the BELA arm had lower iBox scores than those in the CsA arm. The correlation coefficient values were also all positive but low (0.0307-0.3054). The variance of the log HR was highest in the USA arm of the BENEFIT RCT; this region had no recorded graft loss in the BELA arm.

6.6.3.1.1.2 CNI versus CNI-free subjects in the mTORi derivation subset

Since the qualification derivation dataset is not an RCT, it was necessary to perform randomization emulation to account for potential confounders at baseline. Inverse probability treatment weights based on propensity score were used for randomization emulation. This allowed computation of causal treatment effects. This method is further described in Modeling analysis methodologies 5.5.3 (Trial-level surrogacy analysis). The variables that met the criteria for inclusion in propensity score computation were recipient age, ECD, and pre-existing DSA (See Appendix: Revised-Supporting results [Trial-level surrogacy]).

CNI versus CNI-free (mTORi) subjects from the mTORi derivation subset with available iBox scores at one-year post-transplant (consistent with the proposed COU) were selected. Subjects who were on both CNI-based and mTORi-based IST regimens were excluded. This resulted in 1,143 subjects with available one-year iBox scores. Also consistent with the proposed COU, the graft survival status was censored at five years for these subjects who met inclusion criteria. As indicated in Table 59, the survival rate was higher in the CNI-free arm compared to the CNI arm in the mTORi derivation subset.

Table 59. Distribution of subjects and graft loss events per treatment arm for observed iBox scores for the full iBox Scoring System in the mTORi derivation subset

	Full iBox Scoring System	**Five-year death-censored allograft survival**
	Treatment arm	**No. of subjects /**

		No. of events
mTORi derivation subset	CNI-free (mTORi)	99 / 4
	CNI-based	1044 / 64
	Total	1143 / 68

Using the subjects included in the mTORi derivation subset described in the Table above, the propensity scores were computed using logistic regression. It was observed that including the pre-existing DSA resulted in very low propensity scores. Because low weights can make the treatment effects unstable, pre-existing DSA was excluded from the computation of propensity scores. The propensity scores were then generated based on the remaining two variables: recipient age and ECD. While this introduces some bias, the probability of being assigned to CNI-free/CNI if DSA positive is quite low (See Appendix-Revised Supporting Results), so the bias, even if large, did not significantly influence the results. Figure 21 shows the values of the propensity scores, the inverse weights, and the stabilized inverse weights.

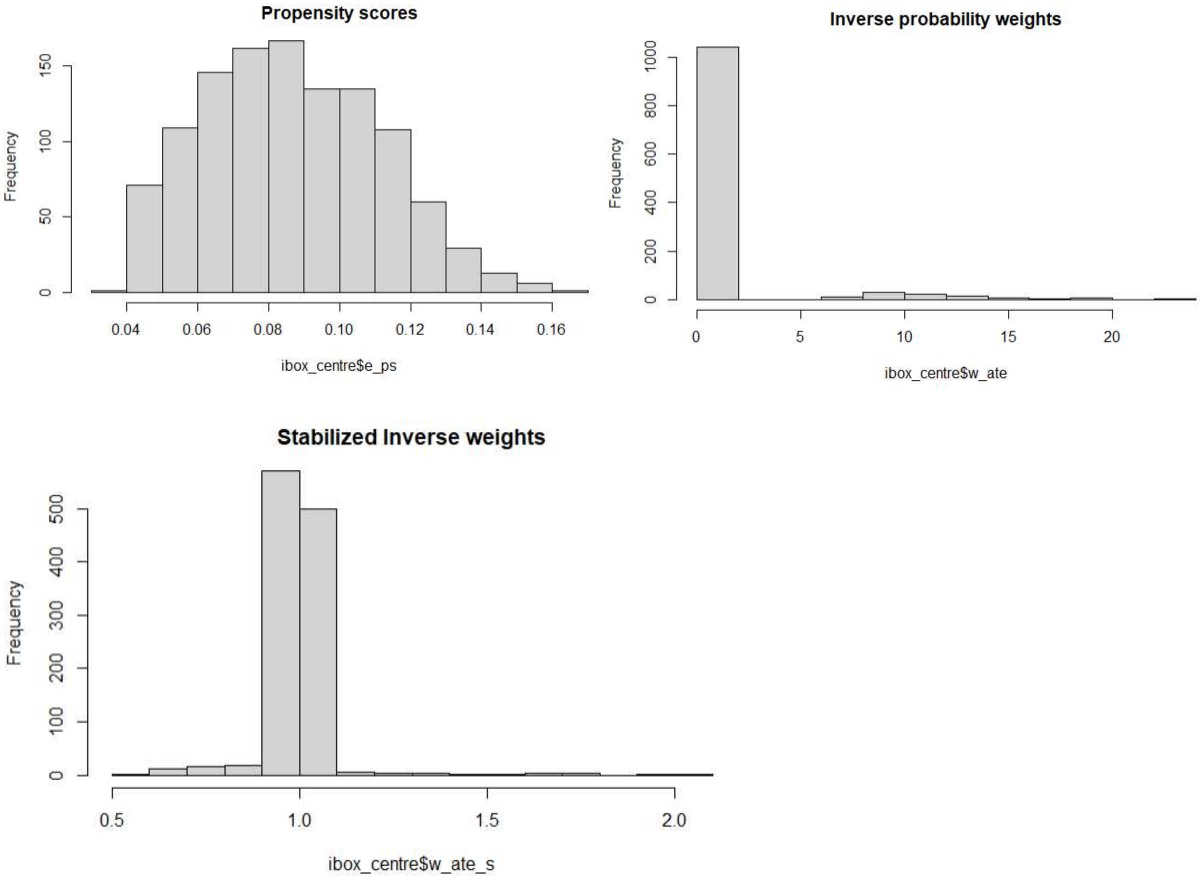


Figure 21. Distribution of propensity scores, inverse probability treatment weights and stabilized weights.

The propensity scores ranged from 0.04-0.17. The inverse probability weights mainly were less than 2, however, there were few values greater than 5. Because of these potential outliers, C-stabilized weights were computed and ranged from 0.5 to 2. These computed stabilized weights were used to generate the treatment effects using weighted linear regression for the full iBox Scoring System without imputation and weighted Cox regression for the five-year death-censored allograft survival. Because of the low number of events in the CNI-free arm, it was not possible to generate reliable bootstrap estimates of variance and correlation. A correlation coefficient value of 0.25 was proposed based on observed patterns in the overall BENEFIT RCT.

Table 60. Causal treatment effect for the mTORi derivation subset with observed iBox scores for the full iBox Scoring System for death-censored allograft survival

	Graft loss		iBox scores		
Estimation	**Log HR**	**SE**	**Mean difference**	**SE**	**Correlation**
Stabilized weights	-0.7382	0.5255	-0.2495	0.1145	0.25

Table 60 shows the treatment effect on the iBox scores and five-year death-censored allograft survival. The treatment effects were both negative, indicating that subjects in the CNI free arm had a lower risk of graft loss compared to the CNI arm. This was consistent with the results generated from the BENEFIT and BENEFIT-EXT RCTs.

Visualize treatment effects across pseudo trials

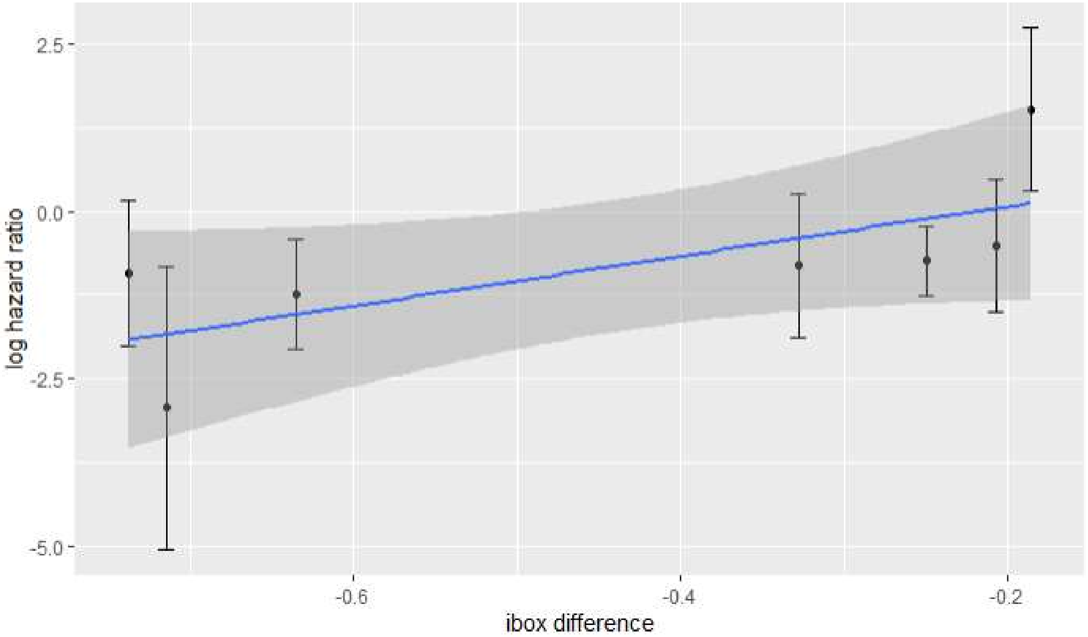


Figure 22: Distribution of treatment effects for BENEFIT RCT, BENEFIT-EXT, and historically constructed pseudo trials.

Figure 22 shows the distribution of the treatment effects of five-year death-censored allograft survival versus the treatment effect of full iBox Scoring System without imputation from the seven pseudo centers as summarized in Table 58 and Table 60 above. The blue line is the linear regression line and the 95% CI band. The results shows a positive relationship between the two treatment effects. The more negative the iBox difference the more negative the log-HR.

6.6.3.1.2 Step two: Generation of the trial-level coefficient

The trial-level correlation coefficient was generated using the results from seven pseudo trials shown in Table 58 and Table 60 above. This was done using a hierarchical Bayesian bivariate normal model described in section Modeling analysis methodologies 5.5.3 (Trial-level surrogacy). Three chains were run, each made of 2,500,000 samples, of which 1,000,000 were discarded as burn-in. The remaining 1,500,000 were used as posterior samples with a thinning rate of 20. These yielded 75,000 posterior samples. The convergence was assessed using the trace plots and auto-correlation plots.

Table 61. Posterior summaries of the parameters from Step two - Hierarchical bivariate normal model for death-censored allograft survival for subjects with observed one-year iBox scores for the full iBox Scoring System

Parameter	Bayes' estimate	SD	95% credible interval
iBox score (γ)	-0.4271	0.12	(-0.671, -0.191)
Log HR (θ)	-0.7522	0.42	(-1.591, 0.0083)
Correlation (ρ)	0.2175	0.56	(-0.901, 0.974)
iBox SD (σ_{γ})	0.2303	0.14	(0.033, 0.563)
Log HR SD (σ_{θ})	0.4696	0.38	(0.018, 1.406)

The results in Table 61 show that the trial-level correlation coefficient is not significantly different from 0 since the 95% credible interval includes 0 (-0.901, 0.974). Trial-level surrogacy was also assessed graphically as shown in Figure 23. The 95% prediction interval was wide and the trend line almost flat suggesting low correlation and precision of the estimates. The credible interval of the correlation coefficient covering almost the full range of possible values shows that the dataset used for TLS analysis does not include enough data to provide precise estimation of the trial-level correlation coefficient. This prevented a meaningful conclusion with respect to whether the full iBox Scoring System model is an adequate surrogate for five-year death-censored allograft survival. These analyses were conducted for abbreviated iBox Scoring System and overall graft survival. The findings from these two analyses were similar for the additional TLS analyses, suggesting there is insufficient data to clearly show trial-level correlation as evidenced by the wide prediction intervals.

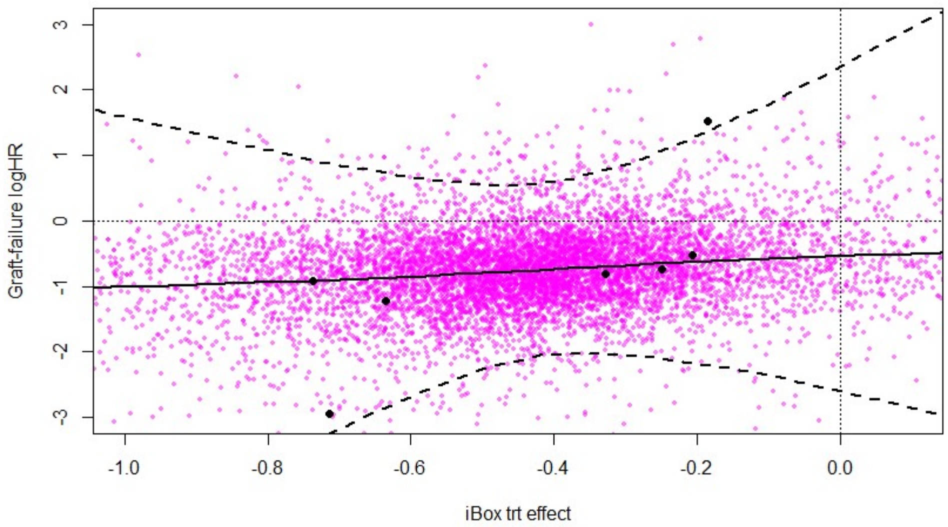


Figure 23. Trial-level surrogacy for five-year death-censored allograft loss with observed iBox scores at one-year for the full iBox Scoring System.

<u>Convergence assessment</u>

The model convergence was assessed using trace plots (

Figure 24) and autocorrelation plots (Figure 25). There was no clear pattern in the trace plots indicating that the chains had converged. The largest autocorrelation value was at most 0.25, showing that the samples had low dependence. The TTC also assessed the convergence using the Gelman Rubin diagnostic, which assesses how close the scale reduction factors were to 1.

The point estimate and 95% interval were all found to equal 1, indicating no convergence issues.

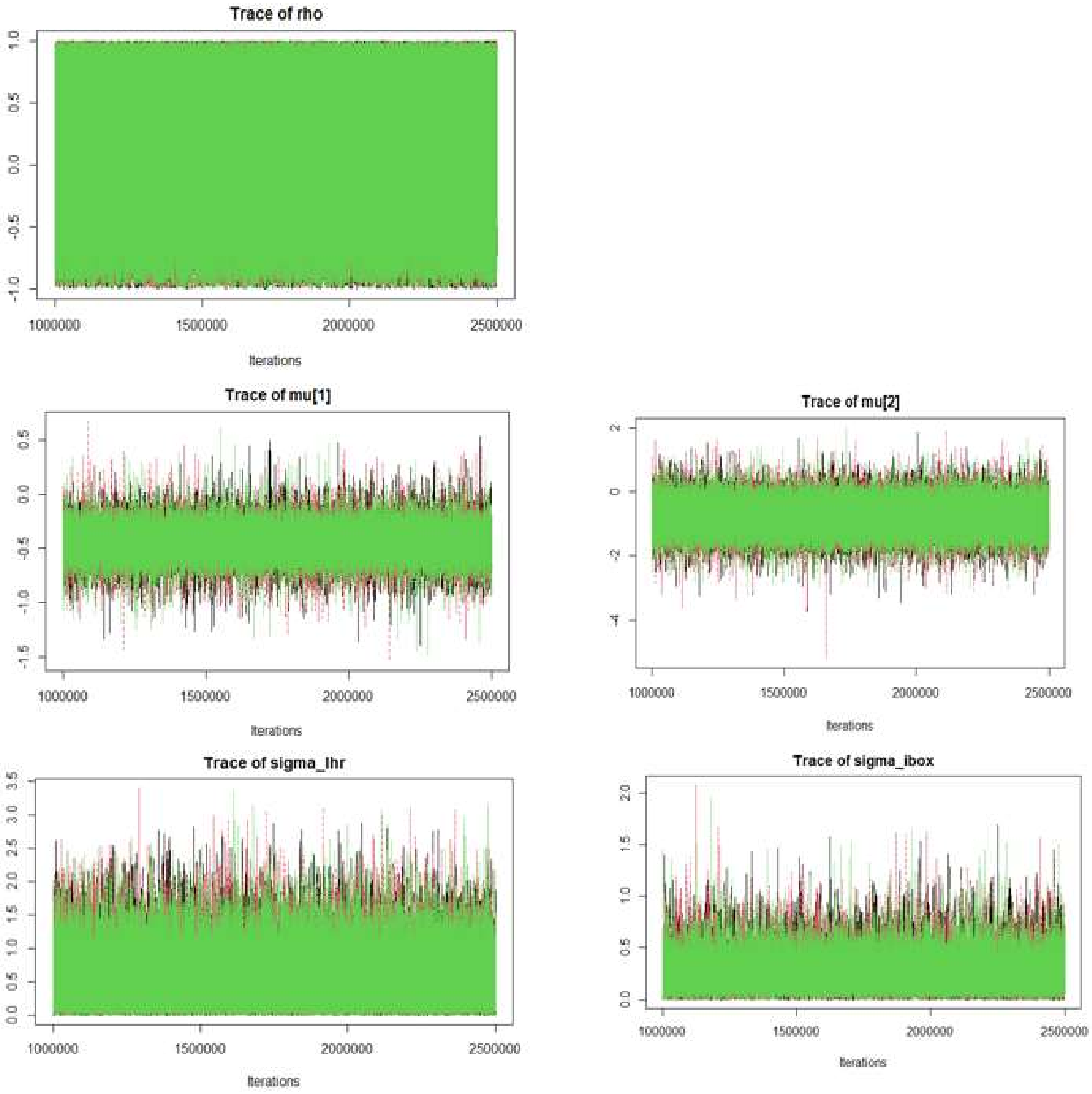


Figure 24. Model convergence assessed using trace of rho, mu[1], mu[2], sigma_lhr, sigma_iBox.

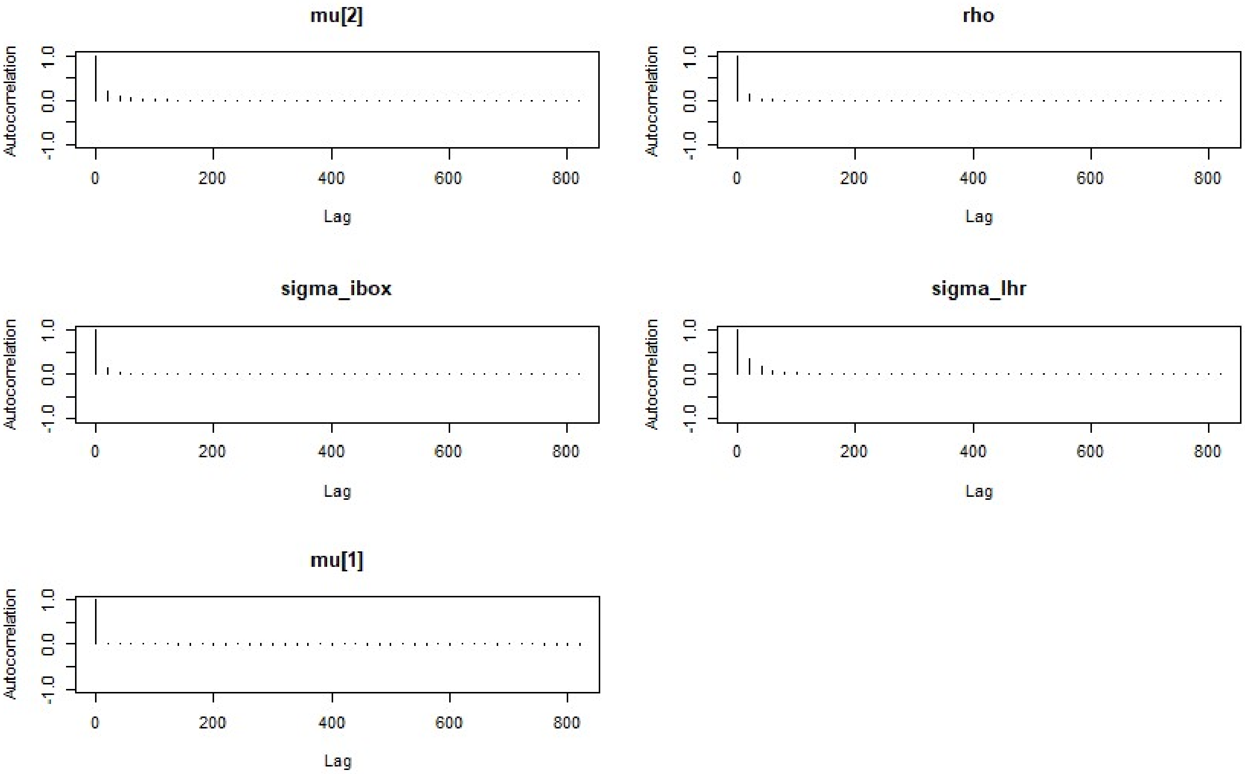


Figure 25. Autocorrelation plots of parameters.

6.6.3.1.3 Summary of overall treatment effect

In addition to the treatment effect within a pseudo trial presented above in step one, treatment effect was computed for all three RCTs. It is promising that the statistical significance of therapeutic effect could be achieved in the BENEFIT RCT (Table 62), and it is perhaps not surprising this significance was not replicated in the two smaller studies, BENEFIT-EXT RCT and mTORi derivation subset, given the lower number of subjects in the treatment and control arms. Despite the paucity of RCTs available globally, the TTC believes that the iBox Scoring System can be used as a primary endpoint in the context of CMA submissions to EMA based on its discrimination and calibration (6.5.2 External validation), and representative treatment effect demonstrated in Table 62 below.

As summarized in Table 62, in all three RCT datasets (two prospective and one retrospective), subjects in the CNI-free arms (BELA or mTORi) had numerically lower iBox scores than the CNI arms. In the BENEFIT RCT (n = 135 CNI, n = 281 CNI-free), a significant treatment effect on the iBox Scoring System corresponds to a significant treatment effect on five-year death-censored allograft survival. The BENEFIT-EXT RCT (n = 85 CNI, n = 175 CNI-free) and mTORi derivation subset (n = 1,044 CNI, n = 99 CNI-free) demonstrated a significant overall treatment effect on the iBox Scoring System with a directional effect on death-censored allograft survival which did not reach statistical significance.

Table 62. Overall treatment effects for the full iBox Scoring System without imputation for five-year death-censored allograft survival in the three RCTs

		CNI-Free	CNI	Treatment effect	P-value
BENEFIT RCT **(n = 416)** **CNI (n = 135)** **CNI-free (n = 281)**	**iBox score at 12 months: Mean (SD)**	-3.61 (0.90)	-2.93 (0.86)	-0.68	<0.0001
	KM survival probability % (SD)	98.2 (0.81)	93.8 (2.30)	-1.31	0.0400
BENEFIT-EXT RCT **(n = 260)** **CNI (n = 85)** **CNI-free (n = 175)**	**iBox score at 12 months: Mean (SD)**	-2.75(0.76)	-2.53 (0.82)	-0.22	0.0377
	KM survival probability % (SD)	95.01 (1.73)	93.98 (2.94)	-0.08	0.8948
mTORi derivation subset **(n = 1,143)** **CNI (n = 1,044)** **CNI-free (n = 99)**	**iBox score at 12 months: Mean (SD)**	-3.04 (1.10)	-2.94 (1.09)	-0.25	0.0319
	KM survival probability % (SD)	96.95 (1.52)	93.50 (0.78)	-0.74	0.1600

The treatment effect for 5-year death-censored graft survival is the log HR, while for the one-year iBox score it is the difference in means. The RCTs (BENEFIT and BENEFIT-EXT) log HRs are constructed from the log-rank test and the iBox treatment effect is the difference in the means of the CNI-free and CNI arms. The log HR and iBox treatment effect for the mTORi derivation subset is computed using the weighted cox regression and weighted linear regression using the inverse probability treatment weights based on propensity scores (see Appendix: Revised-Supporting results [Randomization emulation for TLS]).

6.6.3.2 Summary of trial-level surrogacy

Two prospective RCTs, BENEFIT and BENEFIT-EXT, and one mTORi derivation subset using mTORi versus CNI data from Loupy et al., 2019 were used in the TLS analyses. The results from step 1, i.e., computation of treatment effects for the pseudo trials, demonstrated a positive moderate association between the treatment effects of iBox Scoring System without imputation at one-year post-transplant and five-year death-censored graft survival (Figure 22). Additionally, from step 2, it was determined that there was insufficient data (three RCTs are too few trials) to provide the precise estimation of the trial-level correlation coefficient. There are too few historical clinical trials available that are adequately sized and powered to quantitatively describe the treatment effect relationship on the surrogate and the true outcome. The prediction line from Figure 23 demonstrated a positive association between the two treatment effects with a wide prediction interval (poor precision) due to the limited number of pseudo trials. Moreover, these datasets were not of an adequate sample size to demonstrate a statistically significant therapeutic effect on the true outcome. This prevented an adequate TLS analysis concerning whether the iBox Scoring System at one year detects a treatment effect that translates into differences in five-year death-censored allograft survival. Despite this, the observed positive association between the treatment effects of the two outcomes suggests that with enough trials, the iBox Scoring System may be an adequate surrogate for five-year graft survival. This is also bolstered by the overall treatment effects in the BENEFIT RCT, where a significant treatment effect on iBox score corresponded to a significant treatment effect for five-year death-censored allograft survival. These analyses were further conducted for the abbreviated iBox Scoring System and overall graft survival. The findings from these two analyses were similar for the additional TLS analyses, suggesting that these were insufficient data to support a surrogacy claim specific to these analyses.

In an ongoing effort, the TTC executed a landscape assessment of the field for RCTs that may be capable of supporting the TLS analysis for the proposed COU. Based on this assessment, the TTC believes there are insufficient completed RCTs in existence globally to execute a reasonable TLS analysis.

6.7 Conclusion of validation and supplementary analyses

The full and abbreviated iBox Scoring System measured at one-year post-transplant were investigated as a surrogate for death-censored allograft survival at five-years. The iBox Scoring System, a Cox PH model, was derived based on time of post-transplant risk evaluation, eGFR, proteinuria (measured as log-transformed UPCR), without or without kidney allograft biopsy histopathology findings (four Banff lesion scores), and the presence of DSA, For the purpose of this submission, the time to evaluation was fixed at one-year post-transplant. Their linear combination was defined as the iBox score for the full and abbreviated iBox Scoring System.

Original iBox analyses of data by Loupy et al., 2019 have been reproduced for the full iBox Scoring System for the qualification derivation dataset (n = 3,941). The full iBox Scoring System was re-estimated by dropping the four kidney allograft biopsy histopathology variables, defined as the abbreviated iBox Scoring System (n = 4,000). [6.2 Multivariate analysis].

Model performance was then validated internally using the qualification derivation dataset. For application as an endpoint in a clinical trial at one-year, the qualification derivation dataset was analyzed, restricting the analysis to those recipients with an iBox score at one-year post-transplant and follow-up to five-years for death-censored graft survival (n = 1,174). The discrimination in this group was confirmed with a c-statistic = 0.849. (6.5.1 Internal validation).

These analyses confirmed the internal validation of the full iBox Scoring System and the ability to use it at one-year post-transplant as a surrogate endpoint for the five-year risk of death-censored allograft loss. Furthermore, the high c-statistic value for subjects on an mTORi (0.872, 95% CI from 0.808 to 0.936) suggests that the iBox Scoring System can discriminate accurately even in CNI-free subjects.

Model performance was then validated externally using the qualification validation datasets, with a focus on showing that the eGFR component of the iBox Scoring System model performs well in CNI-free subjects as well as in CNI subjects. External validation was performed using discrimination (c-statistics) and calibration (observed versus predicted graft loss). The results showed that the iBox Scoring System model had a strong discriminatory ability (c-statistics of at least 0.7) across all datasets. The results also showed the full and abbreviated iBox Scoring System had good prediction accuracy based on calibration analysis. External validation was previously performed in Loupy et al., 2019 using the three European centers part of the European validation cohort and the three RCTs, CERTITEM, RITUX ERAH, and BORTEJECT as additional data supporting this qualification submission.

- In all four qualification datasets using the full and abbreviated iBox Scoring System models at one year to predict five-year death-censored allograft survival, the c-statistics ranged from 0.70-0.93, and the predicted versus observed graft losses were not significantly different. These data confirmed the external validation of the iBox Scoring System. 6.5.2 External validation).
- Poisson calibration results for CNI and CNI-free subjects for full and abbreviated iBox models are good, P-values between 0.26-0.82.(6.5.2.1)
- Discrimination (c-statistics) was also included for the European validation cohort (c-statistic = 0.81) and the three RCTs, CERTITEM (c-statistic = 0.88), RITUX ERAH (c-statistic = 0.77), and BORTEJECT (c-statistic = 0.94) described in Loupy et al., 2019 as additional data supporting this qualification submission. 6.5.2.2 (External validation on the European cohort and three RCTs from Loupy et al., 2019).

The ability of the iBox Scoring System to demonstrate a treatment effect at one-year that translates into a treatment effect on death-censored five-year graft survival was assessed in two ways. First, TLS was performed but, due to insufficient data (i.e., i.e., only two prospective RCTs and a mTORi derivation subset), it was not possible to provide the precise estimation of the trial-level correlation coefficient. Secondly, study level treatment effects in the BENEFIT RCT, BENEFIT EXT RCT, and a mTORi derivation subset using mTORi versus CNI data from the qualification derivation dataset for one-year full ad abbreviated iBox Scoring System and five-year death-censored allograft survival were also assessed. Analyses of the BENEFIT RCT included imputation of the worst-case iBox score at one-year post-transplant for recipients who died or lost their graft in the first year. This sensitivity analysis was performed to replicate the clinical trial setting where avoidance of survivor bias at one year would be necessary, and all randomized subjects would have an iBox score at one-year even if there were death or graft loss before that time.

- The iBox score at one year was consistently significantly lower in the CNI-free arm (BELA or mTORi) compared to CNI arms. The five-year death-censored allograft survival also consistently numerically favored the CNI-free arm.
- At five-years in the BENEFIT RCT, death-censored allograft survival was significantly better with BELA compared to CsA.
- The totality of these data demonstrate that the iBox Scoring System can measure treatment effects at one-year that translate into a consistent impact on the five-year death-censored allograft survival. 6.6.3.1.3 (Summary of overall treatment effect)
- The lack of statistical significance on some of the five-year death-censored allograft survival is related to limitations in power to detect differences based on sample size. 6.6.3.2 (Summary of trial-level surrogacy)

Based on these analyses, the iBox Scoring System, with or without biopsy at one-year post-transplant, is a validated surrogate for five-year death-censored allograft survival and is applicable for use in a prospective RCT with imputation for deaths and graft losses within the first year of transplant. Qualification of the iBox Scoring System as a surrogate endpoint would significantly improve upon the current standard, as it would allow drug sponsors the ability to design trials assessing the superiority of a novel agent. As a surrogate endpoint for the long-term outcome of allograft survival, the iBox Scoring System would allow drug sponsors to seek marketing authorisation of novel agents through EMA's CMA while planning and conducting studies to demonstrate longer-term therapeutic effects, significantly improving the drug development landscape by encouraging drug sponsors to engage in this therapeutic area of high unmet need. Ultimately, kidney transplant recipients will benefit from the increased drug development activity by improving access to ISTs with better short-term and long-term outcomes.

6.8 All-cause allograft loss for iBox Scoring System

The iBox Scoring System was also tested for performance when using all-cause graft survival instead of death-censored graft survival as the outcome measure.

6.8.1 External validation

The iBox Scoring System was also tested for performance when using all-cause graft survival instead of death-censored graft survival as the outcome measure. The iBox Scoring System is less performant at predicting overall graft survival, with full iBox Scoring System having reduced c-statistics (Table 63), many of which are below 0.7, and poor calibration (Table 64) compared to death-censored graft survival. These results suggest that the iBox Scoring System is less effective at predicting all-cause graft survival than predicting death-censored graft survival as would be expected given that the iBox Scoring System was trained to predict death-censored graft loss events.

Table 63. Full iBox Scoring System c-statistics for death-censored and all-cause graft survival

Dataset	C-statistic for full iBox Scoring System using death-censored graft survival	C-statistic for full iBox Scoring System using all-cause graft survival
Mayo Clinic Rochester	0.93 (0.03)	0.74 (0.05)
Helsinki University Hospital	0. 78 (0.06)	0.69 (0.04)
BENEFIT RCT	0.70 (0.09)	0.69 (0.06)
BENEFIT-EXT RCT	0.81 (0.07)	0.66 (0.05)

Table 64. Full iBox Scoring System calibration for death-censored and all-cause graft survival

Subject regimen	No. of subjects	Full iBox Scoring System using death-censored graft survival	Full iBox Scoring System

					using all-cause graft survival		
		Observed	Predicted	P-value	Observed	Predicted	P-value
Helsinki University Hospital	344	21	14.40	0.08	46	14.40	<0.01
Mayo Clinic Rochester	483	18	24.34	0.20	35	24.34	0.03
BENEFIT RCT	416	12	14.52	0.51	28	14.52	<0.01
BENEFIT-EXT RCT	260	12	14.97	0.44	41	14.97	<0.01

6.8.2 Trial-level surrogacy for all-cause allograft survival with imputations

6.8.2.1 Step One: Computation of treatment effects

6.8.2.1.1 BENEFIT and BENEFIT-EXT RCTs

TLS was also performed for the event of interest as all-cause graft survival using the BENEFIT and BENEFIT-EXT RCTs and incorporating subjects that experienced the event or were lost to follow up before one-year post-transplant. The subjects that had recorded event status before one year had imputed iBox scores (2.79) for a full iBox Scoring System evaluation. The event of interest was defined as:

1. 1 if the subject died or experienced graft loss before or after 1 year.
2. 0 otherwise.

Table 65 shows the distribution of subjects and events for the full iBox Scoring System.

Table 65. Distribution of subjects and graft loss events per treatment arm for one year with imputation for the full iBox Scoring System

		BENEFIT RCT n = 466		BENEFIT-EXT RCT n = 330	
		Five-year follow-up	**Imputed subjects**	**Five-year follow-up**	**Imputed subjects**

Pseudo trial	Treatment comparison	No. of subjects / No. of events	No. of subjects	No. of subjects / No. of events	No. of subject
Europe	BELA	64 / 3	1	87 / 26	16
	CsA	35 / 5	2	44 / 12	7
Other	BELA	176 / 19	11	85 / 25	12
	CsA	85 / 16	9	47 / 13	10
USA	BELA	69 / 8	5	43 / 15	7
	CsA	37 / 5	3	24 / 9	6
Total events		56	31	100	58

There were 466 subjects in the BENEFIT RCT with 56 deaths/graft losses and 330 subjects in the BENEFIT-EXT RCT with 100 deaths/graft losses after including the subjects for which the iBox score was imputed for a full iBox Scoring System evaluation (i.e., they were no longer in the RCT after one year). Thirty-one more subjects were added to the BENEFIT RCT, while 58 were added to the BENEFIT-EXT RCT. Out of the 31 within BENEFIT RCT, 15 died with function, 14 experienced graft loss, and two were lost to follow up before one year. From the 58 subjects in the BENEFIT-EXT RCT 14 died with function, 43 experienced graft loss, and one was lost to follow up before one year. Findings are summarized in Table 65.

The treatment effects were computed using the log-rank test method for the all-cause graft survival and difference in medians for the full iBox Scoring System as described in the Briefing Dossier (Modeling analysis methodologies [Trial-level surrogacy]). The results are shown in Table 66.

Table 66. Treatment effects for the BMS RCTs for full iBox Scoring System (observed and imputed) five-year all-cause graft survival

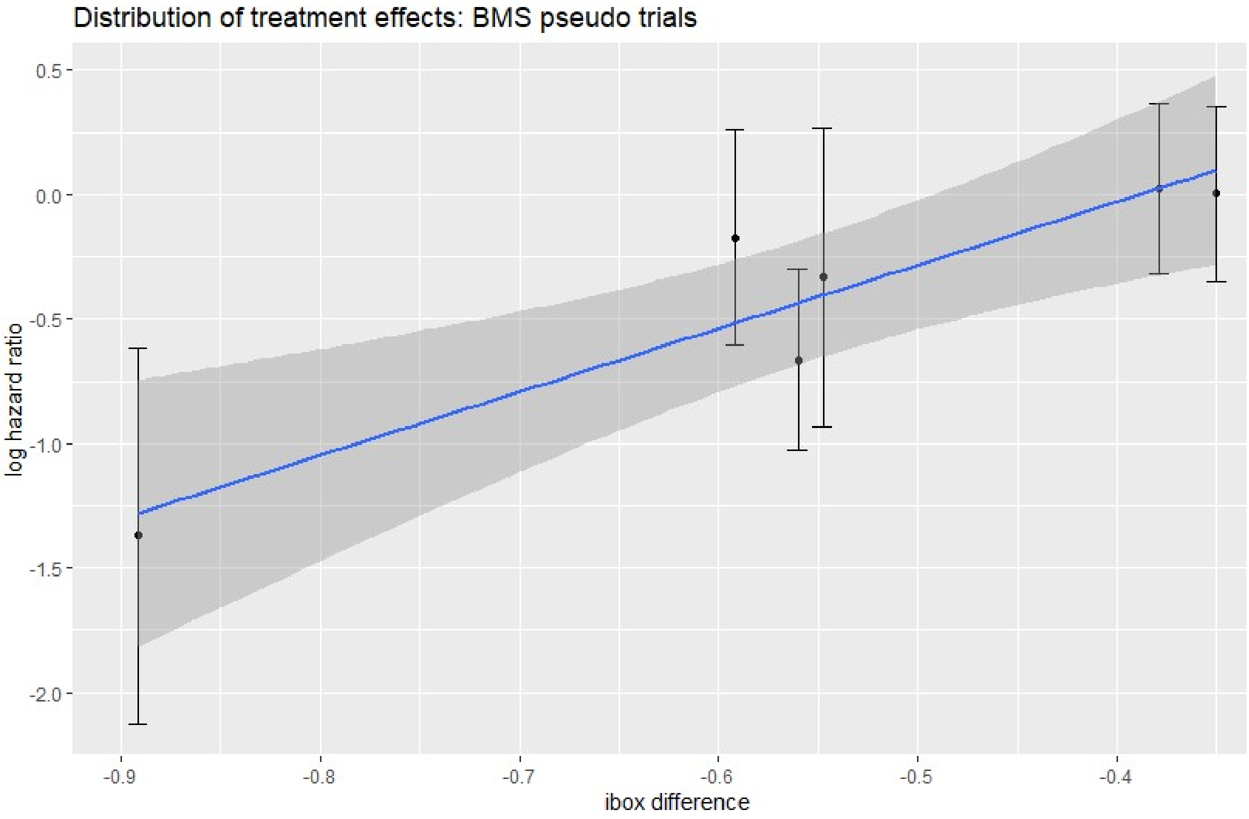
		Graft survival		iBox Scoring System		
Pseudo trial	**Treatment comparison**	**Treatment effect**	**SE**	**Treatment effect**	**SE**	**Correlation**
BENEFIT RCT						
Europe	BELA versus CsA	-1.3708	0.7562	-0.8915	0.2540	0.4250
USA	BELA versus CsA	-0.3327	0.5999	-0.5477	0.1828	0.5377

Other	BELA versus CsA	-0.6645	0.3643	-0.5601	0.1309	0.7075
BENEFIT-EXT RCT						
Europe	BELA versus CsA	0.0037	0.3504	-0.3500	0.2928	0.7847
USA	BELA versus CsA	-0.1735	0.4327	-0.5917	0.4613	0.7588
Other	BELA versus CsA	0.0251	0.3409	-0.3789	0.1678	0.7626

A negative log-HR for five-year all-cause graft survival across all the pseudo trials except Europe and "Other" regions of BENEFIT-EXT RCT were observed.

The difference in iBox scores for the full iBox Scoring System evaluation was also negative in all the pseudo trials (positive treatment effect). This implies that the risk of graft loss was lower in the BELA arm than in the CsA arm. The correlation coefficient values were all positive (0.4250 -0.7847).

Visualize treatment effects



Step 2: Generation of the trial-level coefficient

A trial-level correlation coefficient was generated using results from Table 66 above. This was done using a hierarchical Bayesian bivariate normal model as previously described. Posterior samples were generated as previously. The results of the analysis are presented in Table 67. The adequacy of TLS was also assessed by the graphical plot of treatment effects and its 95% prediction interval, as shown in Figure 26.

Table 67: Posterior summaries of the parameters from Step - Hierarchical bivariate normal model for all-cause graft survival with full iBox Scoring System (observed and imputed)

Parameter	Bayes' estimate	SD	95% credible interval
iBox score (γ)	-0.5370	0.12	(-0.769, -0.315)
Log HR (θ)	-0.3048	0.23	(-0.792, 0.136)
Correlation (ρ)	0.1638	0.58	(-0.924, 0.974)
iBox SD (σ_γ)	0.1386	0.13	(0.005, 0.475)
HR (σ_θ)	0.2893	0.25	(0.010, 0.928)

Convergence of the posterior samples was assessed using trace plots, autocorrelation plots, and Gelman diagnostic (i.e., the scale reduction factor). There was no reason to believe that the chains had not converged from the plots.

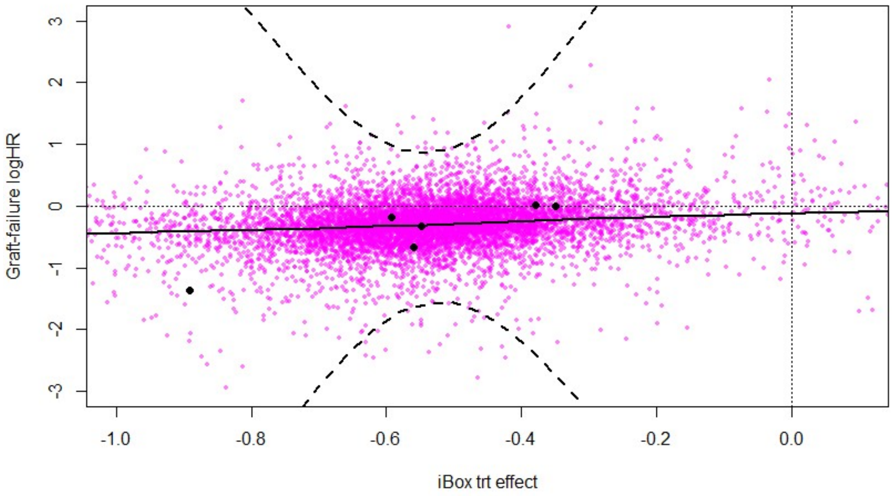


Figure 26. Trial-level surrogacy results for all-cause graft survival with full iBox Scoring System (Observed and imputed).

The trial-level correlation coefficient had a value of 0.1683 with 95% credible interval (-0.924, 0.974). The credible interval covering almost the full range of possible values shows that the dataset used for TLS analysis does not include enough data to provide a precise estimation of the trial-level correlation coefficient, which precludes an adequate assessment of TLS for a treatment effect. TLS adequacy was also examined graphically, as shown in Figure 27, and the 95% prediction interval was very wide in this analysis as well.

Abbreviated iBox Scoring System

TLS for the abbreviated iBox Scoring System for all-cause graft survival was similarly performed. The results are shown in Figure 27. The conclusion remained the same as before with the full iBox scoring system above, i.e., the dataset used for TLS analysis does not include enough data to provide a precise estimation of the trial-level correlation coefficient, which precludes an adequate assessment of TLS for a treatment effect.

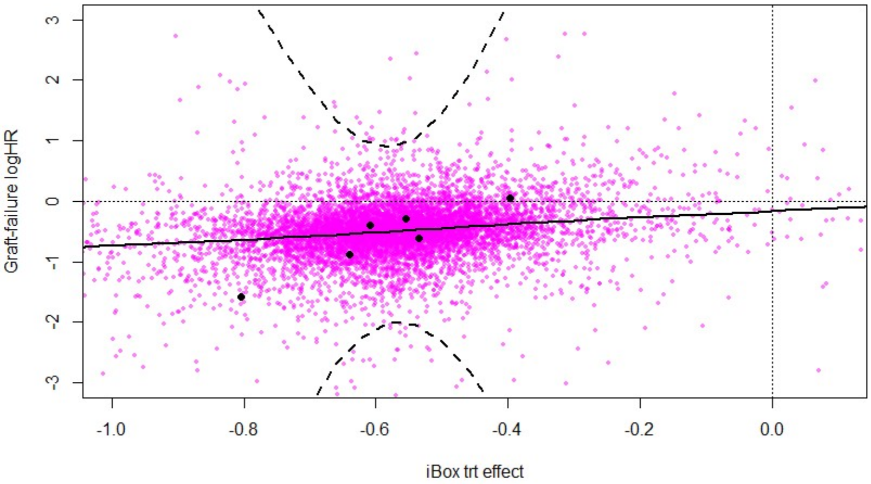


Figure 27. TLS results for all-cause graft survival with abbreviated iBox Scoring System (Observed and imputed).

Study level treatment effects

Table 68. Treatment effects for BENEFIT RCT. The treatment effect for iBox Scoring System is mean/median difference, while that of five-year graft survival, i.e., all-cause and death-censored is log HR

			BELA	CsA	Treatment effect	P-value
	Five-year death-censored graft survival					
No imputation	**Full iBox Scoring System** **(n = 416)**	iBox score at 12 months: Mean (SD)	-3.608 (0.90)	-2.927 (0.86)	-0.681	<0.0001
		KM survival probability % (SD)	98.2 (0.81)	93.8 (2.30)	-1.305	0.04
	Abbreviated iBox Scoring System	iBox score at 12 months: Mean (SD)	-3.835 (0.90)	-3.149 (0.84)	-0.686	<0.0001

	(n = 515)	KM survival probability % (SD)	98.2 (0.72)	96.6 (1.51)	-1.405	0.01
Imputation	**Full iBox Scoring System** **(n = 466)**	iBox score at 12 months: Median (SD)	-3.502 (0.07)	-2.915 (0.10)	-0.587	<0.0001
		KM survival probability % (SD)	96.0 (1.14)	89.7 (2.67)	-0.999	0.02
	Abbreviated iBox Scoring System **(n = 599)**	iBox score at 12 months: Median (SD)	-3.679 (0.05)	-3.042 (0.08)	-0.637	<0.0001
		KM survival probability % (SD)	96.3 (0.96)	89.7 (2.44)	-1.058	0.006
	Five-year all-cause graft survival					
Imputation	**Full iBox Scoring System** **(n = 466)**	iBox score at 12 months: Median (SD)	-3.502 (0.07)	-2.915(0.14)	-0.5869	< 0.0001
		KM survival probability % (SD)	89.64 (1.77)	80.95 (3.36)	-0.6809	0.02
	Abbreviated iBox Scoring System **(n = 599)**	iBox score at 12 months: Median (SD)	-3.679 (0.05)	-3.042 (0.08)	-0.6375	<0.0001
		KM survival probability % (SD)	91.29 (1.46)	79.52 (3.12)	-0.9506	0.0002

Table 69. Treatment effects for BENEFIT-EXT RCT. The treatment effect for iBox Scoring System is mean/median difference, while that of all-cause graft survival is log HR

			BELA	**CsA**	**Treatment effect**	**P-value**
	Five-year death-censored graft survival					

No imputation	**Full iBox Scoring System** **(n = 260)**	iBox score at 12 months: Mean (SD)	-2.7537 (0.76)	-2.5340 (0.82)	-0.2197	0.0377
		KM survival probability % (SD)	95.01 (1.73)	93.98 (2.94)	-0.0822	0.8948
	Abbreviated iBox Scoring System **(n = 358)**	iBox score at 12 months: Mean (SD)	-3.0068 (0.79)	-2.6448 (0.88)	-0.362	0.0002
Imputation	**Full iBox Scoring System** **(n = 330)**	KM survival probability % (SD)	94.50 (1.55)	88.08 (3.43)	-0.8163	0.071
		iBox score at 12 months: Median (SD)	-2.6804 (0.065)	-2.1848 (0.12)	-0.4957	0.0005
		KM survival probability % (SD)	82.92 (2.64)	79.77 (4.04)	-0.1599	0.6
	Abbreviated iBox Scoring System **(n = 455)**	iBox score at 12 months: Median (SD)	-2.9057 (0.07)	-2.4255 (0.12)	-0.4803	0.0007
		KM survival probability % (SD)	85.05 (2.15)	78.54 (3.75)	-0.3292	0.2
	Five-year all-cause graft survival					
Imputation	**Full iBox Scoring System** **(n = 330)**	iBox score at 12 months: Median (SD)	-2.6804 (0.065)	-2.1848 (0.12)	-0.4957	0.0005
		KM survival probability % (SD)	68.53(3.22)	67.46 (4.71)	-0.0259	0.9
	Abbreviated iBox Scoring System **(n = 455)**	iBox score at 12 months: Median (SD)	-2.9057 (0.07)	-2.4255 (0.12)	-0.4803	0.0007
		KM survival probability % (SD)	72.56 (2.67)	66.42 (4.24)	-0.1977	0.3

6.8.2.1.2 CNI versus CNI-free subjects in the mTORi qualification derivation subset

Table 70. Treatment effects for the mTORi qualification derivation subset. The treatment effect for iBox Scoring System is the weighted mean difference, while that of five-year graft survival, i.e., all-cause and death-censored is log HR from the weighted cox model using inverse weights based on propensity scores

		CNI-free (mTORi)	CNI	Treatment effect	P-value
Five-year death-censored graft survival					
Full iBox Scoring System **(n = 1143)**	iBox score at 12 months: Mean (SD)	-3.0355 (1.10)	-2.9448 (1.09)	-0.2469	0.0319
	KM survival probability % (SD)	96.95 (1.52)	93.50 (0.78)	-0.7382	0.16
Abbreviated iBox Scoring System **(n = 1159)**	iBox score at 12 months: Mean (SD)	-3.3256 (1.10)	-3.2426 (1.08)	-0.2262	0.0417
	KM survival probability % (SD)	95.86 (2.03)	93.40 (0.78)	-0.7452	0.156
Five-year all-cause graft survival					
Full iBox Scoring System **(n = 1143)**	iBox score at 12 months: Mean (SD)	-3.0355 (1.10)	-2.9448 (1.09)	-0.2469	0.0319
	KM survival probability % (SD)	88.72 (3.21)	80.59 (1.01)	-0.3173	0.33
Abbreviated iBox Scoring System **(n = 1159)**	iBox score at 12 months: Mean (SD)	-3.3256 (1.10)	-3.2426 (1.08)	-0.2262	0.0417
	KM survival probability % (SD)	91.43 (2.55)	88.28 (1.01)	-0.3283	0.312

6.9 All-cause endpoint score for predicting deaths and graft losses

Recognizing that Regulatory Authorities have historically relied on all-cause graft survival (including death and graft loss as events) to assess long-term kidney transplant outcomes, C-Path explored the performance of various models for predicting all-cause graft loss based on assessments at one-year post-transplant in the qualification datasets. Initially, the iBox score, as derived for predicting death-censored graft loss, was tested for predicting all-cause graft survival (6.8 All-cause allograft loss for iBox Scoring System). As expected, the iBox score underpredicted events with c-statistics ranging from 0.66-0.74 (6.8.1 External validation).

Subsequently, C-Path reviewed the transplant literature to assess factors that have been described to be associated with all-cause graft loss and are post-transplant modifiable parameters. In addition to the components of the abbreviated iBox score, DGF and rejection in the first year were identified as potential additional predictors. Examining the qualification datasets, rejection within the first year was not available in the PTG derivation data. Therefore, this factor was not considered for inclusion in the model. Subsequently, a model including DGF, eGFR, proteinuria, and DSA at one-year post-transplant was derived, and the performance was compared to a new model including the components of the abbreviated iBox Scoring System (eGFR, proteinuria, and DSA) without DGF. Since there was no substantial improvement in the model with the addition of DGF, it was decided to move forward with a new one-year post-kidney transplant ACE score based on eGFR, proteinuria, and DSA to be used as a clinical trial endpoint predictive of five-year graft survival accounting for both deaths and graft losses.

For application as an endpoint in a clinical trial at one year, the derivation dataset from the PTG and the qualification validation datasets were re-analyzed, restricting the analysis to those recipients with measurements of eGFR, proteinuria, and DSA at one-year post-transplant and follow-up to five years. Since the COU and the application of the surrogate as a clinical trial endpoint is fixed at one-year post-transplant, the ACE score did not include time post-transplant in the model, unlike the original derivation of the iBox as described by Alexandre Loupy et al. 2019.

Table 71. Qualification derivation dataset to support all-cause allograft survival model

Dataset	All-cause allograft survival
Loupy et al., 2019 derivation n = 4,000	**Number of subjects**
	n = 1,180

Table 72. Qualification validation datasets to support all-cause allograft survival model

Dataset	All-cause allograft survival
	Number of subjects
Mayo Clinic Rochester	n = 497
Helsinki University Hospital	n = 344
BENEFIT RCT	n = 515
BENEFIT-EXT RCT	n = 357

6.9.1 Model variables

There were three candidate variables considered for inclusion in the all-cause allograft survival model described in Table 15. These candidate variables are commonly and routinely collected in kidney transplant centers worldwide (Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group 2009). These variables were chosen because of their importance to graft loss and their usefulness for observing a treatment effect between a new treatment and control in a *de novo* kidney transplant randomized trial. Backwards elimination using the 31 candidate covariates described previously was performed, but additional predictors were found to have little impact on the estimates of the candidate variables (for more information, see Appendix: Revised-Supporting results) and were therefore excluded.

The final covariates included in the all-cause allograft survival model are described in Table 73.

Table 73. Final covariates in all-cause allograft survival model

	Description of Co-variate at Baseline	Type
1	eGFR (in ml/min/1.73m^2) at 12-months post-transplant	Continuous
2	Log transformed UPCR (g/g)* at 12-months post-transplant	Continuous
3	Donor Specific Antibody (DSA) MFI at 12-months post-transplant	Ordinal (binary)

*Proteinuria values of 0 will have a small positive value added to prevent undefined values

6.9.2 Events of interest for modeling analyses

The primary event of interest was all-cause allograft loss (including death). Validation analyses assessed how well the model predicted all-cause allograft loss specifically; lost to follow up were right-censored. Treatment effect analysis investigated whether there was a significant treatment effect on the surrogate at one year and a corresponding treatment effect on the five-year all-cause graft survival.

All-cause allograft loss model

The semiparametric Cox PH model relates the graft loss events with covariates, as described in 5.4 Cox proportional hazard (PH) model.

The component measures (eGFR, proteinuria, and DSA) were assessed at 12 months post-transplantation. The determined weighting for each component was a coefficient in the multivariate Cox PH model.

Table 74. Calculation of the all-cause allograft loss survival model System

$\boldsymbol{All - cause\ risk\ scorex_i = \Sigma_{j=1}^{4} \widehat{b}_j X_{i,j}}$ for subject i where	
$X_{i,1}$	eGFR, where eGFR is measured in ml/min/1.73m^2
$X_{i,2}$	Log transformed (UPCR value), where UPCR is measured in g/g

$X_{i,3}$	DSA MFI: Categorical variable with 2 levels • MFI < 1400 (reference group) • MFI ≥ 1400

Proteinuria values below 0.05 g/g are replaced by 0.05 g/g before log-transformation.

6.9.3 Multivariate analysis

A multivariate analysis was performed by estimating a Cox PH model for the covariates listed in Methods 4.3.3.3 (Model variables). The 'coxph' function in the 'survival' R package was used for Cox PH analysis (Therneau 2020).

Three variables were explored in the all-cause allograft loss model with and without high-risk donors. These variables included: (1) eGFR, (2) proteinuria, and (3) DSA MFI were combined to generate the all-cause allograft loss model with and without high-risk donors, summarized in Table 75 and Table 80, respectively. There were 1148 subjects with the three variables included in the all-cause allograft loss model, including high-risk donors.

Table 75. Variables explored in the all-cause allograft loss model

Factor	No. of subjects	No. of events*	HR (exp $\widehat{\beta}_j$]) (95% C.I.)*	P-value
eGFR (mL/min/1.73 m^2)	1148	130	0.96 (0.95 to 0.97)	<0.0001
Log transformed UPCR Proteinuria (g/g)	1148	130	1.39 (1.16 to 1.66)	0.0003
DSA MFI				
< 1400	1072	108		
≥ 1400	76	22	2.51 (1.58 to 4.00)	0.0001

6.9.4 Model validation

The following two sections explore model validation. The first section (6.9.4.1 Internal validation) focuses on the internal validation of the all-cause allograft loss model to verify performance on the data the model was trained on (i.e., the qualification derivation dataset restricting the analysis to those recipients with an abbreviated iBox Scoring System evaluation at one-year post-transplant and follow-up to five-years) and identify contexts in which the model may lose predictive power. The second section (6.9.4.2 External validation on the qualification datasets) focuses on the external validation of the all-cause allograft loss model by assessing its discrimination and calibration on external datasets.

6.9.4.1 Internal validation

The c-statistic for the derivation dataset was 0.75 (Table 76).

Table 76. C-statistics for all-cause allograft loss

Dataset	C-statistics for all-cause allograft loss (SE)
Qualification derivation	
Loupy et al., 2019	0.75 (0.02)

6.9.4.2 External validation on the qualification datasets

External validation was performed on the Mayo Clinic Rochester and Helsinki University Hospital observational datasets, and the BENEFIT and BENEFIT-EXT RCTs. C-statistic values were found using the concordance function from the survival R package (Therneau 2020). C-statistics across the qualification validation datasets suggest inconsistent performance on their discriminatory ability for all-cause graft loss, as summarized in Table 77. Good discrimination (meaning a c-statistic of at least 0.7) was observed in the Mayo Clinic Rochester and BENEFIT RCT datasets, while c-statistics for the Helsinki University Hospital and BENEFIT-EXT RCT datasets were lower.

Table 77. C-statistic values at five-years for the qualification validation datasets

Dataset	C-statistic for all-cause allograft loss (SE)
Mayo Clinic Rochester	0.70 (0.06)
Helsinki University Hospital	0.67 (0.05)
BENEFIT RCT	0.78 (0.05)
BENEFIT-EXT RCT	0.67 (0.05)

Calibration was also tested and is shown in Table 78. Once again, Helsinki University Hospital and the BENEFIT-EXT RCT display poorer performance than the other datasets.

Table 78. Poisson calibration results for the all-cause allograft loss model at five-years for the qualification validation datasets

Dataset	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	Observed /Predicted	z score for Observed /Predicted	P-value
Helsinki University Hospital	344	46	27.90	1.65	3.39	<0.01
Mayo Clinic Rochester	497	37	43.95	0.84	-1.05	0.29

BENEFIT RCT	515	35	28.49	1.23	1.22	0.22
BENEFIT-EXT RCT	358	59	39.45	1.50	3.09	<0.01

BENEFIT-EXT RCT, where the model performs poorly, is comprised of subjects receiving extended criteria donors (ECD) (donors ≥60 years old; or donors ≥50 years old and who had at least two other risk factors (cerebrovascular accident, hypertension or serum creatinine >1.5 mg/dL); or an anticipated cold ischemia time of ≥24 h; or donation after cardiac death). Comparatively, the BENEFIT RCT excludes subjects who received a kidney from a donor >60 years old or donors with an anticipated CIT ≥ 24 hours. This practice of excluding high-risk donors is consistent with standard risk *de novo* kidney clinical trials that typically exclude high-risk donors based on age and/or CIT criteria. To investigate how well the model performs in high-risk, defined as donor age ≥ 60 or CIT ≥ 24 hours, compared to standard risk donors, the high-risk donors in Mayo Clinic Rochester and Helsinki University were separated out and evaluated separately. Both discrimination and calibration were evaluated and shown in Table 79. In Helsinki University Hospital, the model showed poor discrimination and calibration on patients with high-risk donors, while the model performed reasonably well in comparison when high-risk donors were excluded (the c-statistic improved but was still slightly below 0.7, and the calibration was reasonable). Mayo Clinic Rochester had too few patients with high-risk donors for meaningful inference, although the c-statistic still appeared poor; however, the model performed reasonably well when high-risk donors were excluded. These results suggest that a model trained on all-cause graft loss has most consistent performance when high-risk donors are excluded.

Table 79. External validation results for high-risk and excluding high-risk donor subjects in Helsinki University Hospital and Mayo Clinic Rochester

Dataset	No. of subjects	C-statistic (SE)	Observed # of all-cause events	Predicted # of all-cause events	z score for Observed /Predicted	P-value
Helsinki University Hospital **High-risk donors***	182	0.65 (0.07)	33	17.82	1.85	<0.01
Helsinki University Hospital **Excluded high-risk donors**	162	0.69 (0.10)	13	10.08	1.29	0.36
Mayo Clinic Rochester **High-risk donors***	64	0.64 (0.18)	5	8.47	0.59	0.24

Mayo Clinic Rochester **Excluded high-risk donors**	422	0.71 (0.06)	30	34.45	0.87	0.45

* Definition for high-risk donors: donor age ≥ 60 or CIT ≥ 24 hours.

The model was refit excluding high-risk donors (results shown in Table 80) and re-evaluated internally and externally for its performance. Out of the 1148 subjects with an abbreviated iBox score and excluding high-risk donors, there were 642 subjects with the three variables included in the model. The internal validation c-statistic increased from 0.75 to 0.77 (Table 81).

Table 80. Variables explored in the all-cause allograft loss model when high-risk donors were excluded

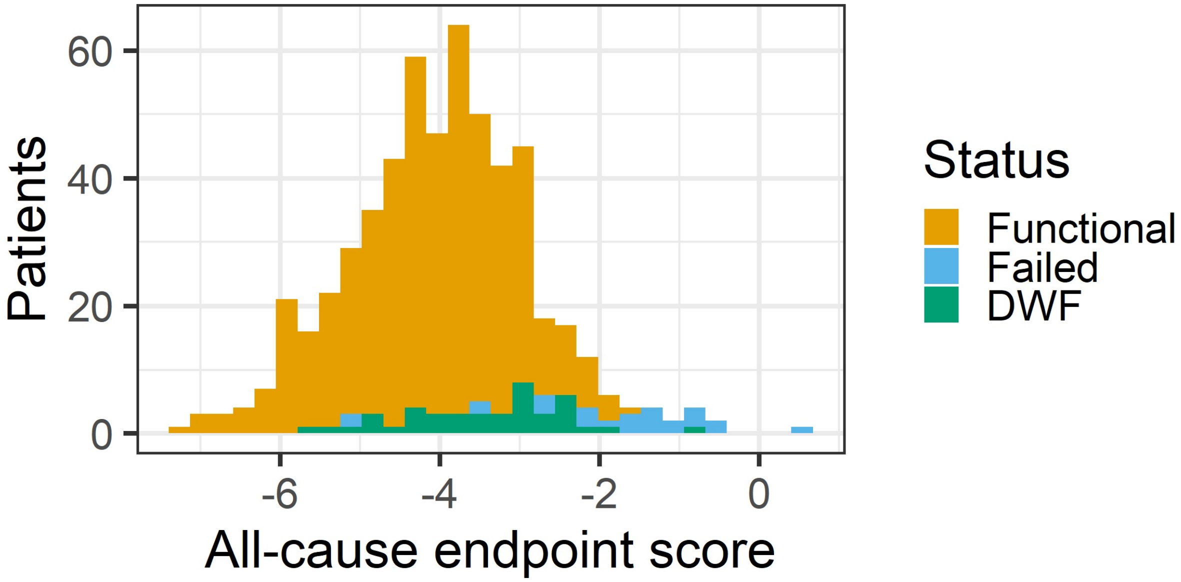
Factor	No. of subjects	No. of events*	HR (exp $\widehat{\beta}_j$]) (95% C.I.)*	P-value
eGFR (mL/min/1.73 m2)	642	60	0.95 (0.94 to 0.97)	<0.0001
Log transformed UPCR Proteinuria (g/g)	642	60	1.63 (1.25 to 2.12)	0.0003
DSA MFI				
< 1400	602	49		
≥ 1400	40	11	3.56 (1.85 to 6.86)	0.0001

Table 81. C-statistics for all-cause allograft loss, with and without high-risk donors

Dataset	C-statistics for all-cause allograft loss, including high-risk donors (SE)	C-statistics for all-cause allograft loss excluding high-risk donors (SE)
Qualification derivation		
Loupy et al., 2019	0.75 (0.02) n = 1148	0.77 (0.03) n = 642

To understand how the all-cause predictor's risk scores are distributed internally, Figure 6 was recreated using COU patients from the qualification derivation dataset without high-risk donors. Graft loss and death with function (DWF) were graphed independently. From Figure 28, graft losses can be seen to have a right-shifted distribution while the distribution of risk scores for patients who died is closer to those who have functional grafts. However, the distribution for DWF patients still appears right-shifted compared to the functional grafts, suggesting the model may have some modest ability to predict DWF when high-risk patients are excluded.

Figure 28. Distribution of ACE scores in the qualification derivation dataset for COU patients with high-risk donors excluded.



External validation was repeated on the qualification derivation datasets with high-risk patients excluded. The rederived model had c-statistics that indicated improved performance in all qualification validation datasets with high-risk donors excluded except BENEFIT-EXT RCT (Table 82, right column). The BENEFIT-EXT RCT consisted entirely of extended criteria donors. While 110 of these did not meet the definition used here to define high-risk (donor age ≥ 60 or CIT ≥ 24), the current definition was chosen out of practicality for applicability across the qualification datasets and in future clinical trials. The remaining 110 patients in the BENEFIT-EXT RCT included donors ≥ 50-60 years old who had at least two other risk factors (cerebrovascular accident, hypertension or serum creatinine > 1.5 mg/dL); or donation after cardiac death. These are still high-risk, and hence the lack of improved c-statistic is not surprising.

Table 82. C-statistic values at five-years for the qualification validation datasets

Dataset	C-statistic for all-cause allograft loss, including high-risk donors (SE)	C-statistic for all-cause allograft loss excluding high-risk donors (SE)
Mayo Clinic Rochester	0.70 (0.06) n = 497	0.71 (0.06) n = 422
Helsinki University Hospital	0.67 (0.07) n = 344	0.69 (0.10) n = 162

BENEFIT RCT	0.78 (0.05) n = 515	0.80 (0.05) n = 487
BENEFIT-EXT RCT	0.67 (0.05) n = 358	0.67 (0.09) n = 110

Model calibration, using the Poisson method as described previously, showed generally good performance once high-risk donors were excluded (Table 83) on all datasets, including the BENEFIT-EXT RCT. This consistent performance was also shown in the calibration plots (Figure 29).

Table 83. Poisson calibration results for the all-cause allograft loss, excluding high-risk donors. Z-scores and p-values were calculated from a Poisson regression model

Dataset	No. of subjects	Observed # of graft loss events	Predicted # of graft loss events	Observed /Predicted	z score for Observed /Predicted	P-value
Helsinki University Hospital	162	13	10.19	1.28	0.88	0.38
Mayo Clinic Rochester	422	30	37.30	0.80	-1.19	0.23
BENEFIT RCT	487	33	26.79	1.23	1.20	0.23
BENEFIT-EXT RCT	110	17	12.14	1.40	1.39	0.16

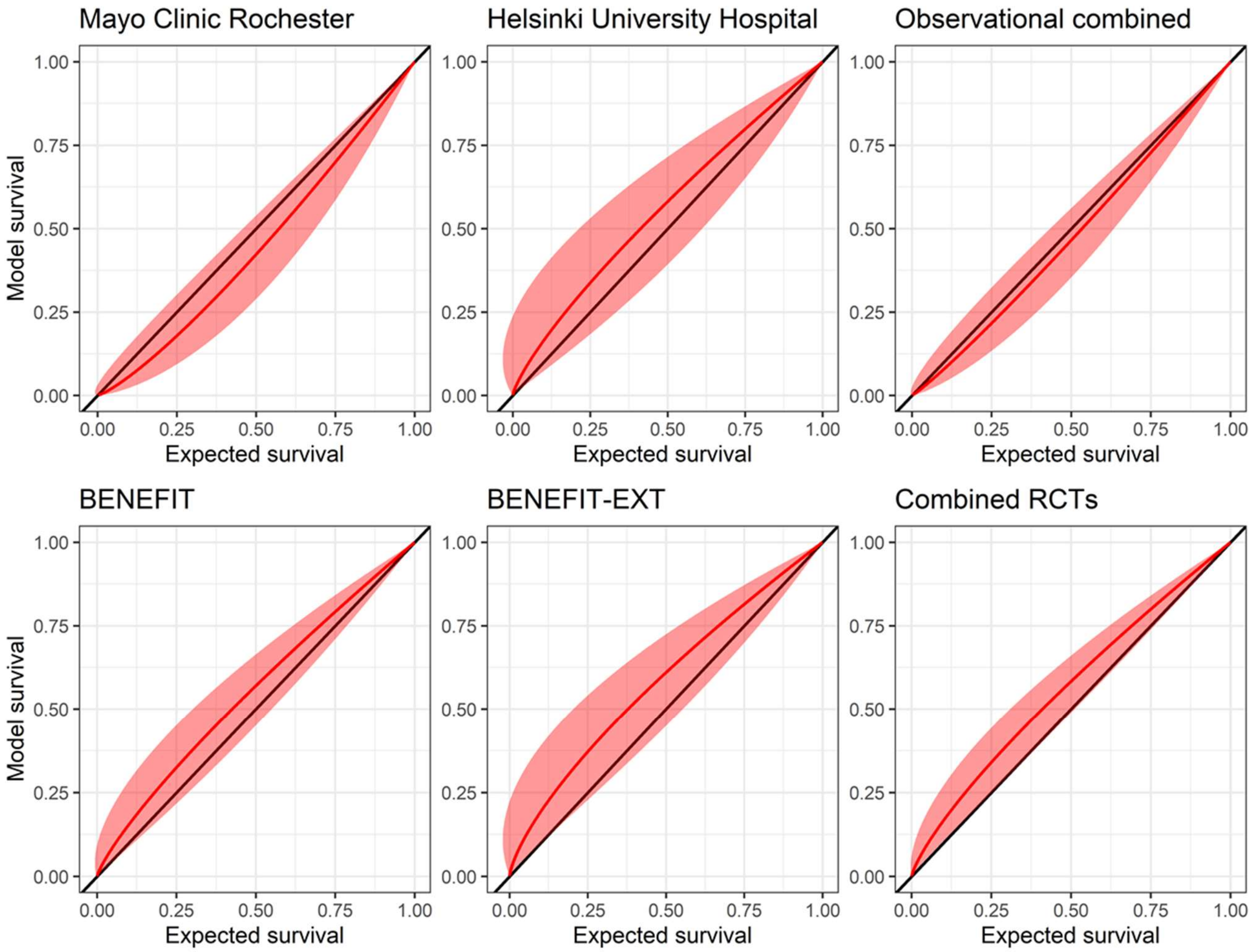


Figure 29. Five-year calibration plot for all-cause allograft loss model with high-risk donors excluded.

6.9.5 Treatment effects

Treatment effects analyses were performed to investigate whether treatment effect was significant on both the surrogate (ACE score) and the five-year all-cause graft survival using 2 RCTs and the mTORi derivation subset. Methodologies for treatment effect computation are previously described in 5.5.3 Trial-level surrogacy analysis.

Since the mTORI derivation subset is not an RCT, randomization emulation is necessary for computation of causal treatment effects. Randomization emulation was performed on the mTORi derivation subset to ensure that the two treatment groups were comparable in terms of baseline covariates. The method used for randomization emulation was inverse weighting based on propensity scores, as previously described in 5.5.3.2.2 CNI versus CNI-free subjects in the mTORi derivation subset.

6.9.5.1 Treatment effects, including high-risk donors

Analyses of five-year all-cause allograft survival for subjects with all-cause endpoint (ACE) score at one-year post-transplant with and without the addition of subjects that died/withdrew/lost their graft within the first year of transplant was conducted.

In all three datasets (two prospective RCTs and one retrospective subset analysis), subjects in the CNI-free arms (BELA or mTORi) had significantly lower ACE scores than the CNI arms. In the BENEFIT RCT (n = 169 CNI, n = 346 CNI-free), a significant treatment effect on the ACE score corresponds to a significant treatment effect on five-year all-cause allograft survival. The BENEFIT-EXT RCT (n = 116 CNI, n = 242 CNI-free) and the mTORi derivation subset (n = 1026 CNI, n = 99 CNI-free) demonstrated a significant overall treatment effect on the ACE score with a directional effect on the five-year all-cause allograft survival that did not reach statistical significance. Findings are summarized in Table 84 below.

Table 84. Overall treatment effects for the all-cause allograft loss without imputation for five-year all-cause allograft survival in the three RCTs (including high-risk donors)

		CNI-Free	CNI	Treatment effect	P-value
BENEFIT RCT (n = 515) CNI (n = 169) CNI-free (n =346)	**ACE score at 12 months: Mean (SD)**	-3.88 (0.89)	-3.16 (0.83)	-0.72	<0.0001
	KM survival probability % (SD)	95.10 (1.20)	86.72 (2.87)	-1.15	0.0018
BENEFIT-EXT RCT (n = 358) CNI (n = 116) CNI-free (n = 242)	**ACE score at 12 months: Mean (SD)**	-2.96 (0.80)	-2.58 (0.91)	-0.38	0.0001
	KM survival probability % (SD)	82.82 (2.51)	78.87 (4.25)	-0.17	0.5546
mTORi derivation subset (n = 1125) CNI (n =1026) CNI-free (n =99)	**ACE score at 12 months: Mean (SD)**	-0.09 (1.08)	0.01 (1.05)	-0.27	0.0139
	KM survival probability % (SD)	88.72 (3.21)	88.21 (1.03)	-0.35	0.2798

The treatment effect for 5-year all-cause graft survival is the log HR, while for the one-year ACE score it is the difference in means. The RCTs (BENEFIT and BENEFIT-EXT) log HRs are constructed from the log-rank test and the ACE score treatment effect is the difference in the means of the CNI-free and CNI arms. The log HR and ACE score treatment effect for the mTORi derivation subset is computed using the weighted cox regression and weighted linear regression using the inverse probability treatment weights based on propensity scores (see Appendix Supporting results - Randomization emulation for TLS).

In the two RCTs, BENEFIT and BENEFIT-EXT, subjects in the CNI-free arm (BELA) had significantly lower ACE scores than the CNI arms. In the BENEFIT RCT (n = 184 CNI, n = 365 CNI-free), a significant treatment effect on both the ACE score and on five-year all-cause allograft survival was found. The BENEFIT-EXT RCT (n = 142 CNI, n = 284 CNI-free) demonstrated a significant overall treatment effect on the all-cause allograft loss risk score with a directional effect on all-cause allograft loss, suggesting improved performance with CNI-free drugs, but did not achieve statistical significance. Findings are summarized in Table 85 below.

Table 85. Overall treatment effects for the all-cause allograft loss with imputation for five-year all-cause allograft survival in the BENEFIT and BENEFIT-EXT RCTs (including high-risk donors)

		CNI-Free	CNI	Treatment effect	P-value
BENEFIT RCT (n = 549) CNI (n = 184) CNI-free (n = 365)	**All-cause event risk score at 12 months: Median (SD)**	-3.79 (0.13)	-3.10 (0.13)	-0.69	<0.0001
	KM survival probability % (SD)	90.40 (1.57)	80.11 (3.15)	-0.82	0.0018
BENEFIT-EXT RCT (n = 426) CNI (n = 142) CNI-free (n = 284)	**All-cause event risk at 12 months: Median (SD)**	-2.86 (0.17)	-2.32 (0.17)	-0.54	0.0019
	KM survival probability % (SD)	70.57 (2.76)	64.43 (4.32)	-0.21	0.2743

The treatment effect for 5-year all-cause graft survival is the log HR, while for the one-year ACE score it is the difference in medians. The RCTs (BENEFIT and BENEFIT-EXT) log HRs are constructed from the log-rank test and the ACE score treatment effect is the difference in the means of the CNI-free and CNI arms.

6.9.5.2 Treatment effects, excluding high-risk donors

The analysis in section 6.9.5 was repeated, this time excluding high-risk donors. As before, subjects in the CNI-free arm (BELA) had numerically lower all-cause allograft loss risk scores than the CNI arms. In the BENEFIT RCT (n = 158 CNI, n = 329 CNI-free), a significant treatment effect was again found on both the ACE score and on five-year all-cause allograft survival. With high-risk donors excluded, the BENEFIT-EXT RCT (n = 30 CNI, n = 80 CNI-free) demonstrated significant overall treatment effect on the ACE score with a directional effect on five-year all-cause allograft survival, suggesting improved performance with CNI-free drugs, but did not achieve statistical significance. The mTORi derivation subset (n = 573 CNI, n = 59 CNI-free) demonstrated a directional effect on both ACE score and the five-year all-cause allograft survival but neither was statistically significant with a trend seen in the ACE score. Of note, the statistical analysis of this subset is limited by the number of CNI-free (mTORi treated) subjects (n=59). Findings are summarized in Table 86 below.

Table 86. Overall treatment effects for the all-cause allograft loss without imputation for five-year all-cause allograft survival in the three RCTs (excluding high-risk donors)

		CNI-Free	CNI	Treatment effect	P-value
BENEFIT RCT **(n = 487)** **CNI (n = 158)** **CNI-free (n = 329)**	**ACE score at 12 months: Mean (SD)**	-4.46 (0.99)	-3.70 (0.94)	-0.76	<0.0001
	KM survival probability % (SD)	95.17 (1.22)	86.66 (2.95)	-1.18	0.0019
BENEFIT-EXT RCT **(n = 110)** **CNI (n = 30)** **CNI-free (n = 80)**	**ACE score at 12 months: Mean (SD)**	-3.64 (0.81)	-2.91 (1.05)	-0.73	0.0006
	KM survival probability % (SD)	83.80 (4.30)	79.25 (8.34)	-0.22	0.6921
mTORi derivation subset **(n = 632)** **CNI (n = 573)** **CNI-free (n =59)**	**ACE score at 12 months: Mean (SD)**	-0.21 (1.23)	0.02 (1.17)	-0.29	0.0995
	KM survival probability % (SD)	94.03 (3.34)	90.04 (1.27)	-0.63	0.2954

The treatment effect for 5-year all-cause graft survival is the log HR, while for the one-year ACE score it is the difference in means. The RCTs (BENEFIT and BENEFIT-EXT) log HRs are constructed from the log-rank test and the ACE score treatment effect is the difference in the means of the CNI-free and CNI arms. The log HR and ACE score treatment effect for the mTORi derivation subset is computed using the weighted cox regression and weighted linear regression using the inverse probability treatment weights based on propensity scores (see Appendix: Revised-Supporting results [Randomization emulation for TLS]).

When deaths or graft losses in the first year post-transplant were imputed as a worst case 12 month ACE score, in the two RCTs, BENEFIT and BENEFIT-EXT, subjects in the CNI-free arm (BELA) had significantly lower ACE scores than the CNI arms. Also in both studies, the BENEFIT RCT (n = 172 CNI, n = 347 CNI-free) and the BENEFIT-EXT RCT (n = 142 CNI, n = 284 CNI-free), a significant treatment effect on the ACE score at 12-months and on the five-year all-cause allograft survival was found. Findings are summarized in Table 87 below.

Table 87. Overall treatment effects for the all-cause allograft loss with imputation for five-year all-cause allograft survival in the BENEFIT and BENEFIT-EXT RCTs (excluding high-risk donors)

		CNI-Free	CNI	Treatment effect	P-value
BENEFIT RCT (n = 519) CNI (n = 172) CNI-free (n = 347)	**ACE score at 12 months: Median (SD)**	-4.38 (0.16)	-3.65 (0.17)	-0.73	<0.0001
	KM survival probability % (SD)	90.50 (1.60)	80.10 (3.24)	-0.84	<0.0001
BENEFIT-EXT RCT (n = 124) CNI (n = 39) CNI-free (n = 85)	**ACE risk at 12 months: Median (SD)**	-3.59 (0.39)	-2.74 (0.39)	-0.85	<0.0001
	KM survival probability % (SD)	78.83 (4.58)	60.97 (8.35)	-0.90	<0.0001

The treatment effect for 5-year all-cause graft survival is the log HR, while for the one-year ACE score it is the difference in medians. The RCTs (BENEFIT and BENEFIT-EXT) log HRs are constructed from the log-rank test and the ACE score treatment effect is the difference in the means of the CNI-free and CNI arms.

6.9.6 Conclusion of validation of the All-cause endpoint score (ACE score)

The ACE score measured at one-year post-transplant was investigated as a surrogate for all-cause allograft survival (including both deaths and graft losses) at five-years. The all-cause allograft loss model, also a Cox PH model, was derived based on the same components as the abbreviated iBox Scoring System (i.e., eGFR, proteinuria, and DSA).

Model performance was then validated internally using the qualification derivation dataset, restricting the analysis to those recipients with measurement of the components of the abbreviated iBox score at one-year post-transplant and follow-up to five-years (n = 1148). The discrimination in this group was confirmed with a high c-statistic = 0.75. (6.9.4.1 Internal validation).

- Model performance was then validated externally using the qualification validation datasets, also restricting the analyses to those recipients with measurements of the components of an abbreviated iBox score at one-year post-transplant and follow-up to five years. External validation was performed using discrimination (c-statistics) and calibration (observed versus predicted graft loss). In the four qualification validation datasets using the ACE model at one year to predict five-year all-cause allograft survival, the c-statistics ranged from 0.67-0.78, where Helsinki University Hospital and the BENEFIT-EXT RCT exhibited c-statistics less than 0.7. Likewise, both Helsinki University Hospital and BENEFIT-EXT RCT also displayed poorer calibration compared to the other datasets. (6.9.4.2 External validation on the qualification datasets).

- Further exploration was performed to understand the differences between datasets and the varying external validation findings. Both Helsinki University Hospital and

BENEFIT-EXT RCT had high proportions of high-risk patients (defined as donor age ≥ 60 or CIT ≥ 24 hours). One possibility was that high-risk patients have generally higher risk of death from non-immune mediated causes resulting in noisier data. C-path investigated this possibility by splitting the Helsinki University Hospital and Mayo Clinic Rochester datasets into high-risk donors and standard-risk donors with standard risk similar to patients in the BENEFIT study. Both c-statistics and calibration indicated better model performance in the population that excluded high-risk donors.

- The model was refit excluding high-risk donors and re-evaluated internally and externally for its performance. The internal validation c-statistic increased from 0.75 to 0.77. External validation was then performed showing improved performance in all qualification validation datasets with high-risk donors excluded except the BENEFIT-EXT RCT, which is comprised entirely of ECD and is relatively high-risk even after the exclusions of donors ≥ 60 or CIT ≥ 24 hours. C-statistics were still somewhat low in the Helsinki University Hospital dataset (0.69), suggesting even without high-risk patients the model may have difficulty discriminating between higher and lower risk patients. However, calibration appeared reasonable in all datasets, including the BENEFIT-EXT RCT, suggesting the model was accurately predicting the total number of events. (6.9.4.2 External validation on the qualification datasets).

- Study level treatment effects in the BENEFIT RCT, BENEFIT EXT RCT, and a mTORi derivation subset using mTORi versus CNI data from the qualification derivation dataset for the one year ACE score and five-year all-cause allograft survival were also assessed. Analyses of the BENEFIT and BENEFIT EXT RCTs included imputation of the worst-case iBox score at one-year post-transplant for recipients who died or lost their graft in the first year. This sensitivity analysis was performed to replicate the clinical trial setting where avoidance of survivor bias at one year would be necessary, and all randomized subjects would have an iBox score at one-year even if there were death or graft loss before that time. (6.9.5 Treatment effects).

- The ACE score at one year was consistently significantly lower in the CNI-free arm (BELA or mTORi) compared to CNI arms. The five-year all-cause allograft survival also consistently numerically favored the CNI-free arm.

- At five-years in the BENEFIT RCT and the BENEFIT EXT, all-cause allograft survival was significantly better with BELA compared to CsA when deaths and graft losses in the first year were imputed.

- The totality of these data demonstrate that the ACE score can measure treatment effects at one-year that translate into a consistent impact on the five-year all-cause allograft survival. (6.9.5.2 Treatment effects, excluding high-risk donors).

- The lack of statistical significance on some of the five-year all-cause allograft survival is related to limitations in power to detect differences based on sample size.

Based on these analyses, the ACE score could be considered an alternative to the iBox Scoring System if all-cause graft loss is the preferred confirmatory long-term endpoint. ACE is a validated surrogate for five-year all-cause allograft survival and is applicable for use in a prospective RCT with imputation for deaths and graft losses within the first year of transplant. If the ACE score is qualified, the associated COU would be: The all-cause endpoint score used at one-year post-transplant is a surrogate endpoint for the five-year risk of all-cause allograft loss (allograft failure) when excluding high-risk donor (i.e., donor age ≥ 60 or CIT ≥ 24 hours) kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA.

7 SUMMARY AND CONCLUSIONS

Qualification of the iBox Scoring System as a surrogate endpoint would significantly improve upon the current standard, by allowing drug sponsors the ability to design trials assessing the superiority of a novel agent. Demonstrating improved long-term outcomes currently is challenging and requires trials of long duration (i.e., five years or more) and with a large number of subjects. As a result, one-to-two-year non-inferiority studies are more likely to be initiated, despite not adequately addressing the challenges of improving long-term graft survival. Surrogate endpoints such as the proposed iBox Scoring System can enable sponsors to seek CMA for novel agents based on clinical trials of reasonable duration (i.e., one year) that predict long-term outcomes (i.e., five years or greater), while sponsors plan and conduct studies to demonstrate longer-term therapeutic effects. The availability of a surrogate endpoint is vital to stimulate innovation in immunosuppressive drug development that will serve transplant recipients by improving short- and long-term outcomes. The ultimate goal is to improve the long-term outcomes in kidney transplant recipients, and a short-term surrogate endpoint is key to reaching that goal.

In 2019, the PTG, together with 29 key opinion leaders of the transplant community from 10 referral centers across Europe and the USA, published a seminal paper on the iBox Scoring System (Alexandre Loupy et al. 2019). Each individual component of the iBox Scoring System is biologically linked to key aspects of kidney health and kidney allograft function. However, the composite gives broader biological insight into the current health of the kidney and the pathologies that lead to death-censored allograft loss than do the individual components in isolation. As such, the iBox Scoring System was derived on eGFR calculated by the 4-variable MDRD-186 Study equation, proteinuria (measured as log-transformed UPCR), with or without kidney allograft biopsy histopathology findings (four Banff lesion scores), presence of DSA, and time of post-transplant risk evaluation. For the purpose of this submission, the time to evaluation was fixed at one-year post-transplant. Their linear combination was defined as the iBox score. The additional information from biopsies needs to be weighed against the challenges of obtaining protocol/surveillance biopsies in all subjects within multinational, multicenter clinical trials. With the choice of two iBox Scoring System models, with and without biopsy data input, a sponsor can assess the ability to perform surveillance biopsies and, if impractical or not feasible, design a simpler, less burdensome clinical trial, knowing both models perform well.

Datasets from relevant clinical trials of ISTs, including the data published in Loupy et al., 2019, and real-world data from international clinical transplant centers were prioritized. Of these 31 datasets, five contained all of the necessary variables (i.e., eGFR, proteinuria, kidney allograft biopsy histopathology, and DSA), long-term death and graft loss follow-up of at least five years, immunosuppressive regimen information (i.e., induction and maintenance IST) to test the performance of the surrogate with all three MOA, and the documentation required to support the description of the analytical considerations for each dataset (see Appendix: Revised-Transplant Therapeutics Consortium's Kidney Transplant Database for more information). The five datasets that had requisite patient-level data to conduct the internal and external validation analyses for this Qualification Opinion submission included those from clinical transplant centers (i.e., Loupy et al., 2019 derivation, Mayo Clinic Rochester, and Helsinki University Hospital) and clinical trials (i.e. [BENEFIT] Vincenti et al., 2012 and [BENEFIT-EXT] Medina-Pestana., 2012) representing over 5,500 *de novo* kidney transplant recipients. The qualification derivation and validation datasets were aligned and curated to support the regulatory endorsement of the iBox Scoring System.

Original iBox analyses of data by Loupy et al., 2019 have been reproduced for the full iBox Scoring System (n = 3,941). Analyses using the abbreviated iBox Scoring System have been performed with the data from the PTG (n = 4,000 for abbreviated iBox Scoring System). [4.3.1 Introduction to data] To qualify the iBox Scoring System models for application as an endpoint in a clinical trial at one-year, the qualification derivation dataset was analyzed, restricting the analysis to those recipients with an iBox score at one-year post-transplant and follow-up to five-years for graft loss (n = 1,174). The discrimination in this group was confirmed with a c-statistic = 0.849. [6.5.1 Internal validation]. Subsequently, external validation was performed in the four qualification datasets (i.e., two observational datasets from Helsinki University Hospital and Mayo Clinic Rochester and two RCTs from BMS, BENEFIT and BENEFIT-EXT) [6.5.2 External Validation]. In all four of the qualification datasets using the full and abbreviated iBox Scoring System models at one year to predict five-year death-censored allograft survival, the c-statistics ranged from 0.70-0.93, and the predicted versus observed graft losses were not significantly different. These data confirmed the external validation of the iBox Scoring System. Discrimination (c-statistics) was also included for the European validation cohort (c-statistic = 0.81, 95% CI 0.78 to 0.84) and the three RCTs, CERTITEM (c-statistic = 0.88), RITUX ERAH (c-statistic = 0.77), and BORTEJECT (c-statistic = 0.94) described in Loupy et al., 2019 as additional data supporting this qualification submission.

The ability of the iBox Scoring System to demonstrate a treatment effect at one-year that translates into a treatment effect on death-censored five-year graft survival was assessed in two ways. First, TLS was performed but, due to insufficient data (i.e., only two prospective RCTs and a mTORi derivation subset), it was not possible to provide the precise estimation of the trial-level correlation coefficient. Secondly, study level treatment effects in the BENEFIT RCT, BENEFIT EXT RCT, and a mTORi derivation subset using mTORi versus CNI data from Loupy et al., 2019 derivation data for one-year iBox Scoring System (full and abbreviated) and five-year death-censored allograft survival were also assessed. The iBox score at one year was consistently significantly lower in the CNI-free arm (BELA or mTORi) compared to the CNI arms. The five-year death-censored allograft survival also consistently numerically favored the CNI-free arm. At five-years in the BENEFIT RCT, death-censored allograft survival was significantly better in subjects treated with BELA compared to those treated with CsA. Analyses of the BENEFIT RCT included imputation of the worst-case iBox Scoring System at one-year post-transplant for recipients who died or experienced graft loss in the first year after transplant. This sensitivity analysis was performed to replicate the clinical trial setting where avoidance of survivor bias at one year would be necessary, and all randomized subjects would have an iBox score at one-year even if death or graft loss occurred before that time point. The totality of these data demonstrates that the iBox Scoring System can measure treatment effects at one-year that translate into a consistent impact on the five-year death-censored allograft survival. The lack of statistical significance on some of the five-year death-censored allograft survival rates is related to limitations in power to detect differences based on sample size.

Additionally, exploratory analyses were performed to assess a surrogate endpoint at one-year post-transplant that would be predictive of all-cause five-year graft loss. The ACE score was found to have sufficient discrimination, calibration, and predictive ability of a treatment effect in *de novo* kidney transplant recipients when high-risk donors were excluded.

The decision to design the iBox Scoring System for predicting death-censored graft failure rather than all-cause graft failure (including death with a functioning graft) was made because recipient death and loss of graft function have different causes. In sensitivity analyses of the iBox Scoring System using competing risk regression models, allograft survival analyses performed in the presented iBox Scoring System model were not affected by competition with patient death.

Based on these analyses, the iBox Scoring System, with or without biopsy, at one-year post-transplant, is a validated surrogate for the five-year death-censored allograft survival and is applicable for use in a prospective RCT with imputation for deaths and graft losses within the first year of transplant. The TTC presents this Briefing Dossier to request a Qualification Opinion from the Agency on the proposed COU for the iBox Scoring System at one-year post-transplant as a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA. The TTC believes a Qualification Opinion is critical for accelerating the development of ISTs in kidney transplantation clinical trials.

8 INTENDED APPLICATION OF PROPOSED TOOL

The iBox Scoring System (Composite Biomarker Panel) and the ACE score are intended to be used at one-year post-transplant as a surrogate endpoint for the five-year risk of allograft loss (allograft failure) in kidney transplant recipients for use in clinical trials to support evaluation of novel IST applications via CMA.

8.1 Methodology of tool

This proposed tool is intended to support the drug development in kidney transplantation by providing a surrogate endpoint capable of predicting the five-year risk of death-censored allograft loss in kidney transplant recipients, significantly improving upon the current standard as it would allow drug sponsors the ability to design trials assessing the superiority, of a novel agent without the need for a cumbersome, lengthy, and prohibitively expensive clinical trial. Additionally, drug sponsors may seek marketing authorisation of novel agents through EMA's CMA while planning and conducting studies to demonstrate longer-term therapeutic effects.

A. The sponsor will collect the proposed iBox Scoring System or ACE score components at one-year post-transplant (i.e., eGFR, UPCR, presence of DSA, with or without kidney allograft biopsy histopathology) to calculate an iBox or ACE score. This would not add an additional burden on the sponsor because all these components are routinely collected in transplant centers and clinical trials worldwide.

B. A comparison between the iBox or ACE scores for the intervention and control arm would allow the sponsor to assess the efficacy of the drug at the end of one year.

Based on 'B' the sponsor could then approach the regulatory authorities for CMA of the drug.

8.2 Key deliverables

The following key deliverables are included in this submission.

- REVISED-User Guide – Master of Table of Contents for Briefing Package.
- The complete modeling analysis report (this document).
- A complete set of modeling scripts and data files to reproduce the results in the report.

9 QUESTIONS FOR EMA FOLLOWED BY TTC's POSITION

1. **Does EMA agree with the COU?**

   TTC's position: The proposed COU provides a quantitative basis to support the use of the iBox Scoring System (Composite Biomarker Panel) at one-year post-transplant is a surrogate endpoint for the five-year risk of death-censored allograft loss (allograft failure) in kidney transplant recipients for use in a clinical trial endpoint at a fixed landmark. Qualifying two iBox Scoring System models, with and without biopsy input, will provide sponsors and investigators flexibility in clinical trial design, with or without a surveillance biopsy at one-year post-transplant.

   As this surrogate endpoint is proposed to be used in the context of CMA with EMA, where full approval of a product will not be authorized until the clinically meaningful outcome (five-year death-censored allograft survival) has been met, the TTC feels that sufficient evidence is provided in this dossier to support qualification of the iBox Scoring System.

2. **Does EMA agree that the data sources are adequate to support the proposed COU?**

   TTC's position: The TTC led an extensive data collaboration effort across the field of kidney transplantation. Datasets from relevant clinical trials of ISTs, including the data in Loupy et al., 2019 publication and real-world data from international clinical transplant centers, were prioritized. There were five datasets that contained all of the necessary clinical variables collected at one-year post-transplant (i.e., eGFR, proteinuria, kidney allograft biopsy histopathology, and presence of DSA), long-term death and graft loss follow-up of at least five years, immunosuppressive regimen information (i.e., induction and maintenance IST) to test the performance of the surrogate with all three MOA, and the documentation required to support the description of the analytical considerations for each dataset in this qualification submission. C-Path has reviewed the documentation and deemed that the analytical methods were robust, reliable, and fit-for-purpose.

   The available data sources, and their alignment through experienced and quality data management, represent a unique opportunity to transform these data into valuable knowledge to provide the necessary evidence to support the qualification of the iBox Scoring System (Composite Biomarker Panel) for the proposed COU. The population captured in the data sources represents the population likely to be considered as candidates to participate in clinical trials of therapies intended to improve long-term graft survival.

3. **Does EMA agree that the iBox Scoring System (Composite Biomarker Panel) or the all-cause endpoint (ACE) score have been validated as a surrogate endpoint for use in CMA submissions per their respective COU?**

   TTC's position: The iBox Scoring System has been internally validated by the PTG and externally validated based on data from two transplant centers (one in Europe and one in the USA) and two Phase III multicenter, multinational RCTs. This external validation demonstrated both calibration and discrimination across the four qualification datasets. The presented analyses show that the iBox Scoring System can discriminate between higher and lower risk subjects in diverse datasets, including CNI and CNI-free populations. The results also showed the full and abbreviated iBox Scoring System had good prediction accuracy based on calibration analysis, including CNI and CNI-free populations in both transplant centers and RCTs.

   The presented results demonstrate that the full and abbreviated iBox Scoring System models at one-year post-transplant are validated surrogates for the five-year death-censored graft survival and are applicable for use in a prospective RCT with imputation for deaths and graft losses within the first year of transplant.

   The iBox Scoring System was designed to assess the long-term risk of allograft failure. Graft failure is defined as return to dialysis or pre-emptive re-transplantation. Death of the recipient with a functioning graft is typically a primary safety endpoint, with a wide variety of underlying causes of death observed (e.g., malignancy, infection, cardiovascular disease) and different risk factors compared with those for graft failure.

The ACE score has been internally validated in the qualification derivation dataset and externally validated in the qualification validation datasets. The ACE score was found to have modest discrimination, calibration, and predictive ability of a treatment effect in *de novo* kidney transplant recipients when high-risk donors were excluded and reduced discrimination as compared to the iBox Scoring System for predicting allograft loss.

10 REFERENCES

Ahmad, Iftikhar. 2004. "Biopsy of the Transplanted Kidney." *Seminars in Interventional Radiology* 21 (4): 275–81. https://doi.org/10.1055/s-2004-861562.

Aubert, Olivier, Nassim Kamar, Dewi Vernerey, Denis Viglietti, Frank Martinez, Jean-Paul Duong-Van-Huyen, Dominique Eladari, et al. 2015. "Long Term Outcomes of Transplantation Using Kidneys from Expanded Criteria Donors: Prospective, Population Based Cohort Study." *BMJ*, July, h3557. https://doi.org/10.1136/bmj.h3557.

Austin, Peter C., Douglas S. Lee, and Jason P. Fine. 2016. "Introduction to the Analysis of Survival Data in the Presence of Competing Risks." *Circulation* 133 (6): 601–9. https://doi.org/10.1161/CIRCULATIONAHA.115.017719.

Bentall, Andrew, Byron H. Smith, Manuel Moreno Gonzales, Keisha Bonner, Walter D. Park, Lynn D. Cornell, Patrick G. Dean, et al. 2019. "Modeling Graft Loss in Patients with Donor-specific Antibody at Baseline Using the Birmingham-Mayo (BirMay) Predictor: Implications for Clinical Trials." *American Journal of Transplantation* 19 (8): 2274–83. https://doi.org/10.1111/ajt.15312.

Campistol, Josep M., Johan W. de Fijter, Björn Nashan, Hallvard Holdaas, Stefan Vítko, and Christophe Legendre. 2011. "Everolimus and Long-Term Outcomes in Renal Transplantation." *Transplantation* 92 (3 Suppl): S3-26. https://doi.org/10.1097/TP.0b013e3182230900.

Collett, David. 2015. *Modelling Survival Data in Medical Research*. 3rd edition. Chapman and Hall/CRC.

Crowson, Cynthia S., Elizabeth J. Atkinson, and Terry M. Therneau. 2016. "Assessing Calibration of Prognostic Risk Scores." *Statistical Methods in Medical Research* 25 (4): 1692–1706. https://doi.org/10.1177/0962280213497434.

Debout, Agnes, Yohann Foucher, Katy Trébern-Launay, Christophe Legendre, Henri Kreis, Georges Mourad, Valérie Garrigue, et al. 2015. "Each Additional Hour of Cold Ischemia Time Significantly Increases the Risk of Graft Failure and Mortality Following Renal Transplantation." *Kidney International* 87 (2): 343–49. https://doi.org/10.1038/ki.2014.304.

Drachenberg, C.B., and J.C. Papadimitriou. 2006. "Polyomavirus-Associated Nephropathy: Update in Diagnosis." *Transplant Infectious Disease* 8 (2): 68–75. https://doi.org/10.1111/j.1399-3062.2006.00154.x.

Dunn, TB, H Noreen, K Gillingham, D Maurer, O. Goruroglu Ozturk, TL Pruett, RA Bray, HM Gebel, and AJ Matas. 2011. "Revisiting Traditional Risk Factors for Rejection and Graft Loss after Kidney Transplantation." *American Journal of Transplantation : Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 11 (10): 2132–43. https://doi.org/10.1111/j.1600-6143.2011.03640.x.

Durrbach, A., J. M. Pestana, T. Pearson, F. Vincenti, V. D. Garcia, J. Campistol, M. del Carmen Rial, et al. 2010. "A Phase III Study of Belatacept Versus Cyclosporine in Kidney Transplants from Extended Criteria Donors (BENEFIT-EXT Study)." *American Journal* *of Transplantation* 10 (3): 547–57. https://doi.org/10.1111/j.1600-6143.2010.03016.x.

Einecke, G., J. Reeve, and P. F. Halloran. 2017. "Hyalinosis Lesions in Renal Transplant Biopsies: Time-Dependent Complexity of Interpretation." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 17 (5): 1346–57. https://doi.org/10.1111/ajt.14136.

Foucher, Yohann, Pascal Daguin, Ahmed Akl, Michèle Kessler, Marc Ladrière, Christophe Legendre, Henri Kreis, et al. 2010. "A Clinical Scoring System Highly Predictive of Long-Term Kidney Graft Survival." *Kidney International* 78 (12): 1288–94. https://doi.org/10.1038/ki.2010.232.

Gerds, Thomas Alexander, Paul Blanche, Rikke Mortensen, Marvin Wright, Nikolaj Tollenaar, John Muschelli, Ulla Brasch Mogensen, and Brice Ozenne. 2020. *RiskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks* (version 2020.12.08). https://CRAN.R-project.org/package=riskRegression.

Ginsberg, J. M., B. S. Chang, R. A. Matarese, and S. Garella. 1983. "Use of Single Voided Urine Samples to Estimate Quantitative Proteinuria." *The New England Journal of Medicine* 309 (25): 1543–46. https://doi.org/10.1056/NEJM198312223092503.

Gondos, Adam, Bernd Döhler, Hermann Brenner, and Gerhard Opelz. 2013. "Kidney Graft Survival in Europe and the United States: Strikingly Different Long-Term Outcomes." *Transplantation* 95 (2): 267–74. https://doi.org/10.1097/TP.0b013e3182708ea8.

Gonzales, Manuel Moreno, Andrew Bentall, Walter K. Kremers, Mark D. Stegall, and Richard Borrows. 2016. "Predicting Individual Renal Allograft Outcomes Using Risk Models with 1-Year Surveillance Biopsy and Alloantibody Data." *Journal of the American Society of Nephrology* 27 (10): 3165–74. https://doi.org/10.1681/ASN.2015070811.

Gray, Bob. 2020. *Cmprsk: Subdistribution Analysis of Competing Risks* (version 2.2-10). https://CRAN.R-project.org/package=cmprsk.

Harrell, Frank E., Kerry L. Lee, and Daniel B. Mark. 1996. "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." *Statistics in Medicine* 15 (4): 361–87. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Hernández, Domingo, Margarita Rufino, Sergio Bartolomei, Víctor Lorenzo, Ana González-Rinne, and Armando Torres. 2005. "A Novel Prognostic Index for Mortality in Renal Transplant Recipients after Hospitalization." *Transplantation* 79 (3): 337–43. https://doi.org/10.1097/01.tp.0000151003.30089.31.

Ho, Julie, Chris Wiebe, David N. Rush, Claudio Rigatto, Leroy Storsley, Martin Karpinski, Ang Gao, Ian W. Gibson, and Peter W. Nickerson. 2013. "Increased Urinary CCL2: Cr Ratio at 6 Months Is Associated with Late Renal Allograft Loss." *Transplantation* 95 (4): 595–602. https://doi.org/10.1097/TP.0b013e31826690fd.

Kaboré, Rémi, Maria C. Haller, Jérôme Harambat, Georg Heinze, and Karen Leffondré. 2017. "Risk Prediction Models for Graft Failure in Kidney Transplantation: A Systematic Review." *Nephrology Dialysis Transplantation* 32 (suppl_2): ii68–76. https://doi.org/10.1093/ndt/gfw405.

Kaplan, Bruce, Jesse Schold, and Herwig-Ulf Meier-Kriesche. 2003. "Poor Predictive Value of Serum Creatinine for Renal Allograft Loss." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 3 (12): 1560–65. https://doi.org/10.1046/j.1600-6135.2003.00275.x.

Kidney Disease: Improving Global Outcomes (KDIGO) Transplant Work Group. 2009. "KDIGO Clinical Practice Guideline for the Care of Kidney Transplant Recipients." *American Journal of Transplantation: Official Journal of the American Society of Transplantation* *and the American Society of Transplant Surgeons* 9 Suppl 3 (November): S1-155. https://doi.org/10.1111/j.1600-6143.2009.02834.x.

Lachmann, Nils, Kremena Todorova, Harald Schulze, and Constanze Schönemann. 2013. "Luminex® and Its Applications for Solid Organ Transplantation, Hematopoietic Stem Cell Transplantation, and Transfusion." *Transfusion Medicine and Hemotherapy* 40 (3): 182–89. https://doi.org/10.1159/000351459.

Lassere, Marissa N., Kent R. Johnson, Michal Schiff, and David Rees. 2012. "Is Blood Pressure Reduction a Valid Surrogate Endpoint for Stroke Prevention? An Analysis Incorporating a Systematic Review of Randomised Controlled Trials, a by-Trial Weighted Errors-in-Variables Regression, the Surrogate Threshold Effect (STE) and the Biomarker-Surrogacy (BioSurrogate) Evaluation Schema (BSES)." *BMC Medical Research Methodology* 12 (March): 27. https://doi.org/10.1186/1471-2288-12-27.

Lefaucheur, Carmen, Alexandre Loupy, Gary S. Hill, Joao Andrade, Dominique Nochy, Corinne Antoine, Chantal Gautreau, Dominique Charron, Denis Glotz, and Caroline Suberbielle-Boissel. 2010. "Preexisting Donor-Specific HLA Antibodies Predict Outcome in Kidney Transplantation." *Journal of the American Society of Nephrology: JASN* 21 (8): 1398–1406. https://doi.org/10.1681/ASN.2009101065.

Lefaucheur, Carmen, Alexandre Loupy, Dewi Vernerey, Jean-Paul Duong-Van-Huyen, Caroline Suberbielle, Dany Anglicheau, Jérôme Vérine, et al. 2013. "Antibody-Mediated Vascular Rejection of Kidney Allografts: A Population-Based Study." *Lancet (London, England)* 381 (9863): 313–19. https://doi.org/10.1016/S0140-6736(12)61265-3.

Levey, Andrew S., Josef Coresh, Tom Greene, Lesley A. Stevens, Yaping (Lucy) Zhang, Stephen Hendriksen, John W. Kusek, Frederick Van Lente, and for the Chronic Kidney Disease Epidemiology Collaboration*. 2006. "Using Standardized Serum Creatinine Values in the Modification of Diet in Renal Disease Study Equation for Estimating Glomerular Filtration Rate." *Annals of Internal Medicine* 145 (4): 247. https://doi.org/10.7326/0003-4819-145-4-200608150-00004.

Levin, Adeera, Rajiv Agarwal, William G. Herrington, Hiddo L. Heerspink, Johannes F.E. Mann, Shahnaz Shahinfar, Katherine R. Tuttle, et al. 2020. "International Consensus Definitions of Clinical Trial Outcomes for Kidney Failure: 2020." *Kidney International* 98 (4): 849–59. https://doi.org/10.1016/j.kint.2020.07.013.

Lim, Mary Ann, Jatinder Kohli, and Roy D. Bloom. 2017. "Immunosuppression for Kidney Transplantation: Where Are We Now and Where Are We Going?" *Transplantation Reviews* 31 (1): 10–17. https://doi.org/10.1016/j.trre.2016.10.006.

Lim, Wai H., Meena Shingde, and Germaine Wong. 2019. "Recurrent and de Novo Glomerulonephritis After Kidney Transplantation." *Frontiers in Immunology* 10: 1944. https://doi.org/10.3389/fimmu.2019.01944.

Loupy, A., M. Haas, K. Solez, L. Racusen, D. Glotz, D. Seron, B. J. Nankivell, et al. 2017. "The Banff 2015 Kidney Meeting Report: Current Challenges in Rejection Classification and Prospects for Adopting Molecular Pathology." *American Journal of Transplantation* 17 (1): 28–41. https://doi.org/10.1111/ajt.14107.

Loupy, Alexandre, Olivier Aubert, Babak J. Orandi, Maarten Naesens, Yassine Bouatou, Marc Raynaud, Gillian Divard, et al. 2019. "Prediction System for Risk of Allograft Loss in Patients Receiving Kidney Transplants: International Derivation and Validation Study." *BMJ (Clinical Research Ed.)* 366: l4923. https://doi.org/10.1136/bmj.l4923.

Loupy, Alexandre, Carmen Lefaucheur, Dewi Vernerey, Christof Prugger, Jean-Paul Duong van Huyen, Nuala Mooney, Caroline Suberbielle, et al. 2013. "Complement-Binding Anti-HLA Antibodies and Kidney-Allograft Survival." *New England Journal of Medicine* 369 (13): 1215–26. https://doi.org/10.1056/NEJMoa1302506.

Matas, A. J., J. M. Smith, M. A. Skeans, B. Thompson, S. K. Gustafson, D. E. Stewart, W. S. Cherikh, et al. 2015. "OPTN/SRTR 2013 Annual Data Report: Kidney." *American Journal of Transplantation: Official Journal of the American Society of Transplantation* *and the American Society of Transplant Surgeons* 15 Suppl 2 (January): 1–34. https://doi.org/10.1111/ajt.13195.

Matos, Ana Cristina, Niels O. Câmara, Lúcio R. REQUIãO-Moura, Eduardo J. Tonato, Thiago C. Filiponi, Marcelino Souza-DURãO, Denise M. Malheiros, Maurício Fregonesi, Milton Borrelli, and Alvaro Pacheco-Silva. 2016. "Presence of Arteriolar Hyalinosis in Post-Reperfusion Biopsies Represents an Additional Risk to Ischaemic Injury in Renal Transplant." *Nephrology (Carlton, Vic.)* 21 (11): 923–29. https://doi.org/10.1111/nep.12699.

Moore, Jason, Xiang He, Shazia Shabir, Rajesh Hanvesakul, David Benavente, Paul Cockwell, Mark A. Little, et al. 2011. "Development and Evaluation of a Composite Risk Score to Predict Kidney Transplant Failure." *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 57 (5): 744–51. https://doi.org/10.1053/j.ajkd.2010.12.017.

Naesens, Maarten, Evelyne Lerut, Marie-Paule Emonds, Albert Herelixka, Pieter Evenepoel, Kathleen Claes, Bert Bammens, et al. 2016. "Proteinuria as a Noninvasive Marker for Renal Allograft Histology and Failure: An Observational Cohort Study." *Journal of the American Society of Nephrology: JASN* 27 (1): 281–92. https://doi.org/10.1681/ASN.2015010062.

Peters-Sengers, Hessel, Julia H.E. Houtzager, Mirza M. Idu, Martin B.A. Heemskerk, Ernst L.W. van Heurn, Jaap J. Homan van der Heide, Jesper Kers, Stefan P. Berger, Thomas M. van Gulik, and Frederike J. Bemelman. 2019. "Impact of Cold Ischemia Time on Outcomes of Deceased Donor Kidney Transplantation: An Analysis of a National Registry." *Transplantation Direct* 5 (5). https://doi.org/10.1097/TXD.0000000000000888.

Port, Friedrich K., Jennifer L. Bragg-Gresham, Robert A. Metzger, Dawn M. Dykstra, Brenda W. Gillespie, Eric W. Young, Francis L. Delmonico, et al. 2002. "Donor Characteristics Associated with Reduced Graft Survival: An Approach to Expanding the Pool of Kidney Donors1:" *Transplantation* 74 (9): 1281–86. https://doi.org/10.1097/00007890-200211150-00014.

Price, Christopher P., Ronald G. Newall, and James C. Boyd. 2005. "Use of Protein:Creatinine Ratio Measurements on Random Urine Samples for Prediction of Significant Proteinuria: A Systematic Review." *Clinical Chemistry* 51 (9): 1577–86. https://doi.org/10.1373/clinchem.2005.049742.

Rostaing, L., A. Hertig, L. Albano, D. Anglicheau, A. Durrbach, V. Vuiblet, B. Moulin, et al. 2015. "Fibrosis Progression According to Epithelial-Mesenchymal Transition Profile: A Randomized Trial of Everolimus Versus CsA." *American Journal of Transplantation* 15 (5): 1303–12. https://doi.org/10.1111/ajt.13132.

Roufosse, Candice, Naomi Simmonds, Marian Clahsen-van Groningen, Mark Haas, Kammi J. Henriksen, Catherine Horsfield, Alexandre Loupy, et al. 2018. "A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology." *Transplantation* 102 (11): 1795–1814. https://doi.org/10.1097/TP.0000000000002366.

Schinstock, Carrie A., Roslyn B. Mannon, Klemens Budde, Anita S. Chong, Mark Haas, Stuart Knechtle, Carmen Lefaucheur, et al. 2020. "Recommended Treatment for Antibody-Mediated Rejection After Kidney Transplantation: The 2019 Expert Consensus From the Transplantion Society Working Group." *Transplantation* 104 (5): 911–22. https://doi.org/10.1097/TP.0000000000003095.

Shabir, Shazia, Jean-Michel Halimi, Aravind Cherukuri, Simon Ball, Charles Ferro, Graham Lipkin, David Benavente, et al. 2014. "Predicting 5-Year Risk of Kidney Transplant Failure: A Prediction Instrument Using Data Available at 1 Year Posttransplantation." *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 63 (4): 643–51. https://doi.org/10.1053/j.ajkd.2013.10.059.

Sis, B., G. S. Jhangri, J. Riopel, J. Chang, D. G. de Freitas, L. Hidalgo, M. Mengel, A. Matas, and P. F. Halloran. 2012. "A New Diagnostic Algorithm for Antibody-Mediated Microcirculation Inflammation in Kidney Transplants." *American Journal of Transplantation* 12 (5): 1168–79. https://doi.org/10.1111/j.1600-6143.2011.03931.x.

Sis, B., M. Mengel, M. Haas, R. B. Colvin, P. F. Halloran, L. C. Racusen, K. Solez, et al. 2010. "Banff '09 Meeting Report: Antibody Mediated Graft Deterioration and Implementation of Banff Working Groups." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 10 (3): 464–71. https://doi.org/10.1111/j.1600-6143.2009.02987.x.

Solez, K., R. B. Colvin, L. C. Racusen, B. Sis, P. F. Halloran, P. E. Birk, P. M. Campbell, et al. 2007. "Banff '05 Meeting Report: Differential Diagnosis of Chronic Allograft Injury and Elimination of Chronic Allograft Nephropathy ('CAN')." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 7 (3): 518–26. https://doi.org/10.1111/j.1600-6143.2006.01688.x.

Stegall, M. D., R. E. Morris, R. R. Alloway, and R. B. Mannon. 2016. "Developing New Immunosuppression for the Next Generation of Transplant Recipients: The Path Forward: Improving Transplant Immunosuppression." *American Journal of Transplantation* 16 (4): 1094–1101. https://doi.org/10.1111/ajt.13582.

Summers, Dominic M, Rachel J Johnson, Alex Hudson, David Collett, Christopher J Watson, and J Andrew Bradley. 2013. "Effect of Donor Age and Cold Storage Time on Outcome in Recipients of Kidneys Donated after Circulatory Death in the UK: A Cohort Study." *The Lancet* 381 (9868): 727–34. https://doi.org/10.1016/S0140-6736(12)61685-7.

Tait, Brian D., Caner Süsal, Howard M. Gebel, Peter W. Nickerson, Andrea A. Zachary, Frans H. J. Claas, Elaine F. Reed, et al. 2013. "Consensus Guidelines on the Testing and Clinical Management Issues Associated With HLA and Non-HLA Antibodies in Transplantation." *Transplantation* 95 (1): 19–47. https://doi.org/10.1097/TP.0b013e31827a19cc.

Therneau, Terry. 2020. "A Package for Survival Analysis in R," April, 90.

Vincenti, F., B. Charpentier, Y. Vanrenterghem, L. Rostaing, B. Bresnahan, P. Darji, P. Massari, et al. 2010. "A Phase III Study of Belatacept-Based Immunosuppression Regimens versus Cyclosporine in Renal Transplant Recipients (BENEFIT Study)." *American Journal of Transplantation* 10 (3): 535–46. https://doi.org/10.1111/j.1600-6143.2009.03005.x.

Weaver, Robert G., Matthew T. James, Pietro Ravani, Colin G.W. Weaver, Edmund J. Lamb, Marcello Tonelli, Braden J. Manns, Robert R. Quinn, Min Jun, and Brenda R. Hemmelgarn. 2020. "Estimating Urine Albumin-to-Creatinine Ratio from Protein-to-Creatinine Ratio: Development of Equations Using Same-Day Measurements." *Journal of the American Society of Nephrology* 31 (3): 591–601. https://doi.org/10.1681/ASN.2019060605.

Webster, Angela C., Vincent W. S. Lee, Jeremy R. Chapman, and Jonathan C. Craig. 2006. "Target of Rapamycin Inhibitors (Sirolimus and Everolimus) for Primary Immunosuppression of Kidney Transplant Recipients: A Systematic Review and Meta-Analysis of Randomized Trials." *Transplantation* 81 (9): 1234–48. https://doi.org/10.1097/01.tp.0000219703.39149.85.

Weir, Matthew R., Fritz Diekmann, Stuart M. Flechner, Yvon Lebranchu, Didier A. Mandelbrot, Rainer Oberbauer, and Barry D. Kahan. 2010. "MTOR Inhibition: The Learning Curve in Kidney Transplantation." *Transplant International: Official Journal of the European Society for Organ Transplantation* 23 (5): 447–60. https://doi.org/10.1111/j.1432-2277.2010.01051.x.

Yilmaz, Serdar, Steven Tomlanovich, Timothy Mathew, Eero Taskinen, Timo Paavonen, Merci Navarro, Eleanor Ramos, Leon Hooftman, and Pekka Häyry. 2003. "Protocol Core Needle Biopsy and Histologic Chronic Allograft Damage Index (CADI) as Surrogate End Point for Long-Term Graft Survival in Multicenter Studies." *Journal of the American Society of Nephrology* 14 (3): 773–79. https://doi.org/10.1097/01.ASN.0000054496.68498.13.