19 December 2022
EMADOC-1700519818-946771
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion for the iBox Scoring System as a secondary efficacy endpoint in clinical trials investigating novel immunosuppressive medicines in kidney transplant patients

Draft agreed by Scientific Advice Working Party (SAWP)	1 September 2022
Adopted by CHMP for release for consultation	15 September 2022[1]
Start of public consultation	6 October 2022[2]
End of consultation (deadlines for comments)	17 November 2022[3]
Adopted by CHMP	15 December 2022

Keywords	Qualification of Novel Methodology, iBox, composite biomarker panel, eGFR, proteinuria, renal allograft biopsy, DSA, time-post-transplant, secondary efficacy endpoint, kidney transplant clinical trials, immunosuppressive medicines, allograft failure

[1] Last day of relevant Committee meeting.
[2] Date of publication on the EMA public website
[3] Last day of the month concerned

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Address for visits and deliveries Refer to www.ema.europa.eu/how-to-find-us
Send us a question Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000 An agency of the European Union

Table of contents

Annexes to this Qualification Opinion published as separate documents as provided by the applicant:

- Validated Briefing Document providing background information
- Appendix to the Briefing Document (BD)
- Written Answers to List of Issues No. 1
- Written Answers to List of Issues No. 2

1 CHMP qualification Opinion statement

CHMP qualifies the iBox Scoring System (Composite Biomarker Panel) as a secondary endpoint prognostic for death-censored allograft loss (allograft failure) in kidney transplant recipients to be used in clinical trials to support the evaluation of novel immunosuppressive therapy applications.

This opinion applies to both the abbreviated and the full iBox Scoring System. Considering the minimal difference in the performance of these two scores and the requirement for an invasive procedure for the full iBox Scoring system, the abbreviated iBox Scoring System may be the preferred one. The scoring systems predict death censored allograft failure at 5 years. This is not the preferred primary clinical endpoint as the preferred primary estimand includes death as an observed event. This should be taken into consideration for the development of a surrogate endpoint and further work on an all-cause endpoint score should be pursued. It is acknowledged that prediction of all-cause death events may be challenging at an early time point post transplantation. Finally, in order to increase the number of trials fulfilling the criteria for validation studies, the Applicant should consider an outcome reflecting the assessment of efficacy of chronic kidney disease, i.e. relative reduction in eGFR (30 to 57%) in addition to graft failure and death (EMA CKD guideline).

The focus of the analysis presented is to support use of the iBox score at 1 year post transplantation to assess 5-year risk of kidney graft failure. Nevertheless, the dataset supports a more flexible COU with the iBox score measured between 6- and 24-months post-kidney transplantation in pivotal or exploratory drug therapeutic studies for regulatory purposes. Additional material is provided to support this conclusion (Appendix to Briefing Document). The CHMP encourages the use of the iBox scoring system as a secondary endpoint in future trials of kidney transplantation and further development of the scoring system targeting a potential future qualification as a surrogate endpoint. Sponsors may consider using the iBox Scoring System as a secondary endpoint with Type 1 error control included in a procedure to address multiplicity in pivotal trials.

For a more detailed discussion of the CHMP assessment, please see '3. Questions posed by the applicant and CHMP answers'.

2 Executive summary as submitted by the applicant

2.1 The objective(s) of the 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

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).

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.

2. Mayo Clinic in Rochester, Minnesota, USA, North America.
3. Helsinki University Hospital in Helsinki, Finland, Europe.
4. 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.
5. 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. 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).

Table 1. 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 1. 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. $eGFR_{MDRD}$ 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. $eGFR_{MDRD}$ 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	Mayo Clinic Rochester[I]; Helsinki University Hospital; BENEFIT RCT;

	School of Medicine, Richmond, VA	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

[ł] 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).
- 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.
- 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).

- 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.
- 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.
- 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.

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 Questions from the Applicant and CHMP answers

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 as 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.

CHMP answer

It is agreed that there is a need to develop a reliable surrogate measure that performs across a heterogenous population and allow showing efficacy across therapies with multiple mechanisms of action (MoA).

The initially proposed Context of Use (COU) for the two composite biomarker panels was use as a surrogate endpoint to predict 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 target population are adult de novo kidney only transplant recipients from a living or deceased donor. Development of two scores (one with and one without histology) seems reasonable and could ease recruitment and maintenance of patients in future studies; biopsy may be associated with bleeding, renal fistulas and haematuria. The transplant recipient may refuse biopsy for study purposes only.

After discussion of two lists of issues provided by SAWP, the COU was modified and refined with a final proposal of the statement reading 'The iBox Scoring System (Composite Biomarker Panel) is a co-primary or secondary endpoint prognostic for death-censored allograft loss (allograft failure) in kidney transplant recipients to be used in clinical trials to support the evaluation of novel immunosuppressive therapy applications.' Additional information was provided by the Applicant that supports a more flexible COU with the iBox measured between 6- and 24-months post-kidney transplantation in pivotal or exploratory drug therapeutic studies for regulatory purposes. While the focus would likely be long-term prediction of death-censored graft loss, also shorter periods for prediction would be feasible with less events expected in a shorter time frame. From regulatory point of view the preferred primary clinical endpoint is to include death as an observed event. This should be taken into consideration for future development of the iBox Scoring System.

The more flexible COU would allow using the iBox scoring system in proof of concept or dose finding phase 2 studies and phase 3 studies. It is possible that iBox could add supportive evidence for CMA, provided requirements for CMA are fulfilled. These are outlined in EMA guideline (EMA/CHMP/509951/2006, Rev.1). For the iBox to support CMA, it will have to be able to support a positive benefit-risk balance of the medicine under investigation and it will have to be ensured that it is likely that comprehensive data post-authorisation will be generated. The timeframe to provide data post-authorization should not jeopardize the conduct of the study, e.g., in case of availability of a newly approved medical therapy.

The Applicant states (chapter 3.2) that when using the death-censored iBox score, it will always be necessary to determine if there is clinically meaningful decrease in transplant recipient survival with new therapy. This view is shared. Other post-authorisation requirements for CMA include the fulfilment of an unmet medical need and the benefit of the medicine's immediate availability to patients is greater than the risk inherent in the fact that additional data are still required.

There are several other regulatory approaches available to address safety, and/or efficacy, post approval. Such, post-authorisation measures (PAMs) may be aimed at collecting or providing data to enable the assessment of the safety or efficacy (see EMA website "Post-authorisation measures: questions and answers").

In conclusion, the initial COU proposed for iBox scoring system was a surrogate endpoint to support CMA. As explained, a surrogate endpoint is not a priori linked to a specific regulatory pathway within the EU. Please see the discussion regarding the assessment of iBox scoring system as a surrogate endpoint in the answer to Q3.

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 centres, 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.

CHMP answer

It is agreed that the clinical transplant population is heterogenous. This also poses a challenge to establishing surrogacy. The proposed target population is "Adult *de novo* kidney only transplant recipients from a living or deceased donor", i.e. the broad population of adult transplants. The efforts

of the TTC to acquire subject-level data for development of the proposed surrogate endpoint are acknowledged. Selecting studies (five out of 31) which contained all variables of interest is a reasonable approach. The variables with the composite panel are clinically relevant as they provide information on the health of the graft through measuring of renal function (proteinuria, eGFR), direct assessment of allograft health through histopathology, and the patient's immune response (DSA).

The five qualification datasets consist of one derivation dataset and four validation datasets; the latter comprised two prospective RCTs (the BENEFIT study and the BENEFIT EXT with a different target population). Subjects with grafts that never functioned (primary non-function) were excluded from the derivation data set. The broad range of patients and the variety data sources in the data set are acknowledged. The documentation of the laboratory assays used is adequate and supports reliability and adequacy of the analytical laboratory methods. There was reclassification applied to address the fact that different criteria for graft loss were used across the data sets. This led to a number of reclassifications and there was a considerable number in the BENEFIT and BENEFIT-EXT studies. Standardisation of criteria, using consensus criteria according to Levin et al. (Levin A et al., Kidney International 2020) was implemented during the validation procedure, but is in principle welcomed and obviously important. The ad hoc reclassification was discussed at the first discussion meeting (DM) and there was no impact on interpretation of calibration results.

The studies included in the qualification exercise represent subjects with varying underlying diagnoses, receiving living related as well as extended donor kidneys, receiving various induction therapies and either CNI or CNI free therapy. As such, the data sources included are generally acceptable. However, the size of the database of the external validation studies is too small to determine consistency of the data across subpopulations. Also, most of these subsets are limited to single treatment centres. A limitation of the data sources is the small number of patients included in therapeutic intervention trials that are important for assessing the change in treatment effects in the proposed surrogate and the clinical endpoint at 5 years. Outcome events derived from randomised controlled trials are too sparse to be fully informative for the surrogacy at trial level of the iBox biomarker panel. To illustrate, in the largest trial there were 416 subjects with full iBox data at one year in the BENEFIT RCT, of whom 12 graft losses at 5 years were recorded.

The low number of endpoint events in the available trials with an intervention limit establishing a correlation of treatment induced modification of the surrogate to treatment induced modification of the endpoint at 5 years. Such a relation is considered key for establishing full surrogacy of a biomarker-based endpoint. The correlation coefficients indicating the relation between treatment effect on iBox score and treatment effect on 5-year allograft survival were positive but low (0.0307-0.3054). Splitting the data into pseudo-trials per region as performed by the Applicant (p. 121 BD) was helpful in allowing further assessment of the correlation at (pseudo-)trial level but does expectedly not contribute much to improve precision of estimates for elucidating trial level surrogacy. Trial level surrogacy is assessed in the answer to Question 3. The ongoing efforts of the TTC to explore if additional RCTs exist that that may support the trial-level surrogacy (TLS) are acknowledged. The notion that there are insufficient completed RCTs in existence globally to execute a reasonable TLS analysis is noted.

During the DM several approaches were discussed which would potentially increase the number of trials fulfilling the criteria for validation of the iBox. These include using clinical trials that do not collect histology results to at least validate the abbreviated iBox, using outcome data at 3 years following transplantation, and redefining the outcome data to include relative reduction in eGFR. However, as per the Applicant, none of these measures were found to improve the number of trials available for validation of the iBox.

Taken together, the whole exercise would benefit from access to more data. Extensive global effort to collect clinical trials and real-world data on the side of the Applicant is understood and appreciated.

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 centres (one in Europe and one in the USA) and two Phase III multicentre, 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 centres 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.

CHMP answer

<u>Overall validation approach</u>

CHMP acknowledges the strengths of the current model development and validation approach and also the extensive and valuable initial work of the group led by Loupy (Loupy A et al, BMJ 2019). The initial prospective approach by Loupy et al. for derivation data collection led to a prediction model has good predictive performance for clinical endpoint events based on a number of variables included in a biomarker panel proposed as iBox. The model was internally and externally validated. Based on feedback from CHMP in a scientific advice on a proposal to use iBox as surrogate endpoint in a clinical phase 3 trial (EMA/CHMP/SAWP/650635/2019), the Applicant refined the approach and performed additional analysis. Inclusion of a new independent set of validation data for the refined development of the iBox score by TTC is welcomed by CHMP.

The differences to the initial approach by Loupy et al. are comprehensively explained in the BD (p. 21). These include a different approach to handling donor specific antibodies (DSA) and pertain to the fixed 1-year time point proposed by the Applicant for the COU, which was addressed by imputing data for patients who die or lose graft during the first year. Imputation or spot proteinuria to reflect UPCR

was performed for three of the four validation studies (BENEFIT, BENFIT-EXT and population from Helsinki University Hospital) and discussed below.

Two iBox models are proposed and this is in principle supported to allow flexibility in application in trial settings. The abbreviated iBox without biopsy information is supported by only a minimally larger number of subjects in the derivation data set (n=4000 vs. n=3941) and was retained after dropping the four kidney allograft biopsy histopathology variables in Table 38 of the BD. Backward elimination was not repeated after dropping the biopsy variables; the main goal with the abbreviated iBox was showing that dropping biopsy variables had minimal impact on model performance in the external datasets. In the external validation dataset, more data without biopsy information are available. The development approach of the abbreviated model was discussed at the first meeting, e.g., if an abbreviated iBox could be re-derived with omitting biopsy related information. The Applicant explained that the 31 candidate variables explored in the derivation of the iBox Scoring System are not consistently present in the qualification validation datasets. It can also be concluded that restricting the analysis to an abbreviated iBox Scoring System will not increase the available data for analyses.

Missing data is minimal in derivation data set, any covariate imputation approach (opposed to imputation of iBox for patients that do not reach the 1-year time point) has no considerable input. Model development and analysis for internal validation was mainly data driven and this is acceptable in the given setting with pre-planned external validation based on additional independent data sets. The step of establishing trial level surrogacy for full validation of iBox has limitations, mainly due to the available datasets with a low number of observed events (please see below).

Modelling and statistical methods

The modelling approach can be endorsed. As primary event of interest, graft loss was defined and death and loss to follow up were censored, assuming that these events are non-informative. As death as competing event could be informative, a competing risks analysis was performed. This is considered adequate. Subjects who died/withdrew/lost their graft before the first year after transplantation have missing iBox score values. These subjects were assigned imputed iBox score values. This is deemed a reasonable approach to avoid survivor bias. Incorporating scores for subjects who died for application of iBox with censoring for death using worst case scenario values for iBox at 1 year can be supported in principle (p. 56 BD).

A separate modelling approach using all-cause graft survival was also pursued to assess model performance when death is included in the model. The process of model derivation is appropriate. Univariate and multivariate analysis was used for variable selection from the 31 candidate variables. Backward elimination due to clinical considerations and rationale for categorical breakdown of variables in the univariate and multivariate models is comprehensively explained and can be supported. Overall, model assumption assessment for the Cox proportional hazard model and assessments of linearity of covariate using martingale residuals are endorsed. Testing the discriminatory properties for patients with and without graft loss (e.g., by ROC curve, p. 78 BD) is considered adequate. Using log transformed proteinuria values due to skewed data distribution appears adequate.

For performance assessment, Harrell's c-statistic was used (Harrell F, Stat Med 1996). This is an appropriate metric. Based on this measure, performance in patients without CNIs was assessed, as the training data used mainly subject treated with CNIs. Additionally, model performance with center as stratification factor was explored. Both steps are adequate and contribute to the validation. For an assessment of the predictive properties of the model with regard to accurately predicting the absolute risks, for calibration the number of predicted clinical endpoints were derived based on a Poisson model

and compared to the observed events (Crowson C et al., Stat Methods Med Res 2016). The method of assessment of calibration is supported.

Supplementary assessment to assess the proteinuria conversion, death as competing risk for graft loss in the death-censored model and trial level surrogacy was performed. All these analyses are appropriately conducted and comprehensively described. It should be noted that imputation of urine-dipstick reflects spot concentration of urine-albumin and may change, e.g., with increased fluid intake, which is not the case for 24-hour proteinuria or UPCR. It is understood that the extrapolation of spot urine albumin by dipstick was based on a German population with both UPCR and dip stick results. The approach was further discussed at the first discussion meeting, as fit of the data was not clearly demonstrated and the IQR (middle 50%) presented is very wide (figure 16). It can be concluded from the results and discussion that the imputation of urine-dipstick for albumin for the three validation cohorts does not adequately reflect UPCR. However, given that the level of proteinuria in chronic transplant nephropathy is generally mild, it is not expected that the imputation has major impact on the overall performance of the iBox score. During the discussion meetings (DMs) with the Applicant it was also evident that the qualification and validation exercise were tested separately for two different eGFR equations with equal performance (MDRD and SCr based CKD-EPI).

Model validation

Model diagnostics and Schoenfeld residual analysis for influential/outlier observations are adequate and do not cause concerns. The final model retained 8 variables in the full iBox score.

Internal validation

Internal validation focused on the full iBox panel. The abbreviated iBox Scoring System was not internally validated (p. 99 of the BD). The c-statistics for the derivation dataset were 0.809 and 0.803 for the full and abbreviated iBox Scoring Systems, respectively (table 42, p. 100 BD). The c-statistics for the abbreviated iBox Scoring System showed that it is not significantly different than the c-statistics for the full iBox Scoring System. This is acknowledged and supports the use of both score variants.

Various scenarios and subpopulations were examined in the qualification dataset for their c-statistic using the iBox Scoring System (table 43, p. 102 BD). The full iBox Scoring System showed a good ability to discriminate the between higher and lower risk subjects for various important scenarios and subpopulations, with c-statistic values ranging from 0.76 to 0.87.

Three subsets showed significantly different c-statistic values from the c-statistic of 0.809 for the qualification derivation dataset (i.e., the 3,941 subjects for the full iBox Scoring System). This includes subjects transplanted with kidneys from elderly (c-statistic, 95% CI: 0.777, 0.746 to 0.808) and hypertensive donors (c-statistic, 95% CI: 0.771, 0.737 to 0.805). 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) in the mTORi subset of the derivation population shows also 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. Overall, c-statistics in the derivation subsets suggests that the full iBox Scoring System performs well in various clinically relevant scenarios and subpopulations.

External validation

External validation was performed using the four external qualification datasets: Mayo Clinic Rochester and Helsinki University Hospital observational transplant center data, and the 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 the c-statistic censored at five-years plus 28 days post-transplant. All c-statistic values in table 45 are 0.70 or greater for each qualification validation dataset. The Applicant pointed out some shift in c-statistics score for the full- and the abbreviated iBox scores between datasets. This is explained by two participants with high eGFR at one year who lost their grafts. Similar change in c-statistic score was noted between the full- and abbreviated iBox in the Mayo cohort due to graft losses in two individuals at low risk of graft loss. The calibration results show that overall the predicted number of events is reasonably well matching the number of observed events with some over- and underprediction when using single data sets, but with somewhat higher margins of error. This pertains to the full and abbreviated iBox score and also to subpopulations with treatment that is CNI based and without CNIs.

Overall, the data are considered encouraging. However, due to the limited number of data sets for validation and the limited number of graft loss events, the model assessment is subject to uncertainty and predicted event numbers show some variability. This also precluded assessment of the model in different subgroups, as was done for the derivation cohort.

Supplementary analysis for validation

Competing risks analysis

The sponsor 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. Second, a Fine-Gray sub distribution survival model was built those accounts for death and compared to the iBox Scoring System, which is a Cox survival model that does not account for death. The sponsor gave detailed explanations and references for the two applied methods which are agreed upon.

The result of the analysis is that censoring deaths has little to no impact on predictions of graft loss in the derivation dataset.

Trial level surrogacy

The focus of the TLS analysis was to: (1) estimate the treatment effect for each trial on full and abbreviated iBox Scoring System and graft loss, and (2) compute the correlation coefficient and/or the surrogate threshold effect (STE).

Due to limited availability of RCT, the two RCTs BENEFIT and BENEFIT-EXT were split into pseudo trials based on regions to support the TLS method. Splitting the data into pseudo-trials per region as performed by the Applicant (p. 121 BD) was helpful in allowing further assessment of the correlation at (pseudo-) trial level but does expectedly not contribute too much to improve precision of estimates for elucidating trial level surrogacy. Additionally, a subset of their derivation dataset was used, consisting of subjects who were on a CNI-free mTORi-based therapy, sirolimus or everolimus, versus CNI-based therapy at the time of transplant. To reduce potential confounding issues that can be present when examining non-RCT data propensity score techniques were used to reweight subjects in the derivation dataset. The addition of retrospective data from non-randomised comparisons in patients of the Loupy et al. cohort is only acceptable as supportive analysis. Inverse probability weighting based on propensity scores was used to allow comparisons. Results suggest that not all potential prognostic factors could be included, a stabilisation approach for the weights was necessary and it was not possible to generate bootstrap estimates for variance and correlation. While these

issues raise concerns on the addition of non-randomised data to the exercise, even when these issues were not present, conclusions from the TLS analysis would not change.

No precise estimation of the trial-level correlation coefficient could be derived from these data. 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. 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.

The low number of endpoint events in the available trials with an intervention limit establishing a correlation of treatment induced modification of the surrogate to treatment induced modification of the endpoint at 5 years. Such a relation is considered key for establishing full surrogacy of a biomarker-based endpoint.

The TLS correlation analysis of treatment effects shows the limitations. The attempt to establish a trial level coefficient using a hierarchical Bayesian bivariate model shows a wide credible interval for the trial level coefficient including zero and therefore indicates the limitations for the precision of the estimate.

Validation of an All-cause Endpoint score

ACE score development

The primary event of interest in the ACE is all-cause allograft loss (including death). The abbreviated iBox composite score assessed at one year was used to assess all-cause graft survival. As can be expected, the model considerably underpredicts events. The model was therefore refined based on prior knowledge. Originally, known predictors for all-cause graft loss were delayed graft function (DGF) and rejection in the first year. These potential risk indicators were however not included in the model for predicting all-cause graft loss based on assessments at one-year post-transplant due to non-availability of rejection in the first year data in the PTG derivation dataset and due to "a non-substantial improvement" in risk prediction when DGF was included in the model compared to the use of the scoring system without DGF (= abbreviated iBox). The resulting model with eGFR, proteinuria and DSA was therefore taken forward for the ACE model. The considerations for model development are acknowledged.

With external validation datasets (p. 144 BD), C-statistics showed variable performance in moderate to good range of the ACE score on the discriminatory ability across the validation datasets (lowest C-statistic in Helsinki University Hospital 0.67 and Benefit-EXT RCT 0.67; C- statistic range from 0.67 to 0.78). When excluding high risk donors, the model performed only slightly better (improvement in Helsinki University Hospital data plus 0.02, from 0.67 to 0.69).

C-statistics in the qualification derivation dataset (Loupy et al. 2019) showed moderate performance of the ACE score, again with a better performance when excluding high risk donors (C-statistics 0.75 with and 0.77 without high risk donors). Consequently, the model was adapted to exclude high risk donors. However, results showed moderate improvement in performance (table 82, p. 147 BD). The distribution plot of ACE scores for the derivation dataset without high-risk donors is illustrative (figure 28, p. 147), showing separately the resulting counts for patients at 5 years discriminating patients alive with functional graft and deaths with functional graft and deaths with graft failure. This figure shows that the discriminatory ability for deaths with functional graft and deaths with graft failure of the ACE scoring system is modest.

The trial level surrogacy analysis (p. 149 BD) was repeated for the ACE. Treatment effect analyses were performed to investigate whether the treatment effect was significant on both the surrogate (ACE score) and the five-year all-cause graft survival. Two RCTs (Benefit-EXT RCT and BENEFIT RCT) and

the mTORi derivation subsets were used. Concordance (significant treatment effect on ACE score and significant treatment effect on five-year all-cause survival) was found in one dataset (BENEFIT RCT), but not in the two others, where a directional effect on survival was found, but without statistical significance (likewise shown in analyses with and without high-risk donors). Like the surrogacy analysis in the iBox score systems, these results may be due to lack of statistical power.

Albeit not all predictors for all-cause graft survival were included in the ACE score, identity between this score and the abbreviated iBox score enables comparison of results. The performance of the ACE is less good than the iBox and this may be expected since predicting death events with functional graft may be difficult based on information tailored to predict renal events. Considering the above and the observed results, from a performance and sensitivity perspective, the iBox score should be preferred over the ACE score.

Conclusions

The Applicant initially proposed the iBox scoring system with full and abbreviated score without biopsy information as surrogate endpoint assessed at 1 year for prediction of outcomes at 5 years specifically tailored to settings with a conditional marketing authorisation application.

Overall, an extensive validation exercise has been performed, comprising internal validation based on prospectively collected data, external validation including randomised clinical studies and a trial-level surrogacy analysis. Previous work by Loupy et al. and the work done by the Applicant are comprehensively documented. Results show that the proposed iBox score models are suitable for individual predictions of graft loss events with good performance based on c-statistics and with the ability to predict numbers of graft loss events with reasonable, but not small margins of error. However, trial level surrogacy could not be established due to limited data in terms of available studies and event numbers. This is acknowledged by CHMP and also by the Applicant. Therefore, the iBox scores can currently not be formally qualified as surrogate endpoint for use as a primary endpoint. However, the use of the iBox as a secondary endpoint could be encouraged in order to further stimulate robust assessment of the iBox score and efficiency of drug development for treatments to prevent kidney graft failure.

During the discussion meetings with the Applicant it was evident that further data are needed in order to validate the iBox scoring system as a surrogate endpoint. This is understood and supported.

4 Background as submitted by the applicant

Please refer to the validated Briefing Document (BD) and other submitted documentation published as separate documents for the evidence presented.