Biomarker Qualification Letter of Intent Submission
Critical Path Institute
Transplant Therapeutics Consortium

Administrative Information

1. **Submission Title**: The Integrative Box (iBox) Scoring System as a reasonably likely surrogate endpoint for five-year risk of allograft loss in kidney transplant recipients for use in clinical trials evaluating the safety and efficacy of novel immunosuppressive drug therapies.

2. **Requesting Organization:**
Critical Path Institute
1730 E. River Rd.
Tucson, AZ 85718
520-547-3440
C-Path.org

Primary Contact: Inish O'Doherty
Executive Director, TTC
607-379-2336
iodoherty@c-path.org

Alternate Contact: Nicole Spear
Senior Project Manager, TTC
301-442-3406
nspear@c-path.org

3. **Submission Date:** December 13, 2019

   Resubmission Date: February 14, 2020

Drug Development Need Statement

In the United States, over 19,000 patients received a kidney transplant in 2017. With the general standard of care (SOC) immunosuppressive drug (ISD) therapy, according to the most recently available data from the Organ Procurement and Transplantation Network, one-year post-transplant allograft loss rates are remarkably low (2.5% and 7.8% for living and deceased donor transplants, respectively)("National Data - OPTN" n.d.). However, long-term allograft failure rates are unacceptably high, with 10-year all-cause allograft failure approaching 34% and 50% (Hart et al. 2019) for living and deceased donor transplants, respectively. Survival of the transplanted organ has been rated, by patients, as the most important outcome, including overall survival of the patient (Howell et al. 2012). There is a clinical need for novel individual ISDs or ISD regimens that will lead to improved long-term outcomes. However, the deterrents to ISD innovation are complex and multifactorial. Most notably is the lack of short-term predictors of long-term outcomes available for use in clinical trials. Such markers could serve as the basis for surrogate endpoints or reasonably likely surrogate endpoints that open FDA's Accelerated Approval Program, intended to incentivize and usher innovation.

The historically accepted clinical trial endpoint for novel ISDs in kidney transplantation is the equally weighted score of patient survival, graft-loss, biopsy-proven acute rejection (BPAR), and loss to follow up at one-year. This primary endpoint is a vestige of an era when graft loss and BPAR were significant issues in the first year following kidney transplantation. The short-term success of current ISD regimens for this outcome measure and the lack of markers capable of predicting long-term outcomes require clinical trials of novel agents to be lengthy, require large numbers of subjects, or be non-inferior in design to show superiority over the current standard. Large and lengthy trials are associated with prohibitively large costs while the non-inferiority determinations are associated with concerns of marketability for newly approved agents in the face of more affordable generically available SOC ISD regimens. In this Letter of Intent (LOI) we propose building on previous work in the field that has identified clinically relevant measures capable of predicting long-term kidney allograft failure. We aim to improve upon the limitations of the historically utilized clinical trial primary endpoint by developing a composite score capable of predicting long-term kidney transplant outcomes using measures available in the first year following transplantation.

While the underlying physiologic mechanisms leading to allograft loss are complex, recent studies have shown that certain key features present relatively early after transplantation (e.g., within the first year) can accurately predict which grafts are most likely to fail at later time points (e.g., at 5 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, and Meier-Kriesche 2003; Yilmaz et al. 2003; Lefaucheur et al. 2010). Recent efforts that have had access to large patient 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; Loupy et al. 2019). 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 loss, including kidney function, through estimated glomerular filtration rate (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 or absence of de novo (i.e., developed after the time of transplant) donor specific anti-human leukocyte antigen (HLA) antibodies. Other composite scores have incorporated pathophysiological measures along with 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 the Supporting Information section.

These risk prediction scores have focused on predicting long-term allograft survival at the individual patient level to inform clinical decision making. While progress has been made, none of these prognostic and predictive tools have been endorsed for use as a reasonably likely surrogate endpoint capable of supporting medical product registration studies or as surrogate or reasonably likely surrogate endpoints that can open FDA's Accelerated Approval Program.

The proposed composite marker in this submission is intended to be a reasonably likely surrogate endpoint for use in clinical trials evaluating the safety and efficacy of novel ISD therapies in kidney transplant patients as a marker for the probability of long-term allograft loss. This would significantly improve upon the current standard as it would allow drug sponsors the ability to design trials assessing the superiority, rather than non-inferiority, of a novel agent without the need for cumbersomely long, large, and prohibitively expensive clinical trials. As a reasonably likely surrogate for the long-term outcome of allograft survival, this composite score would allow drug sponsors to seek marketing approval of novel agents through FDA's Accelerated Approval Program, significantly improving the drug development landscape by encouraging drug sponsors to engage in this rare disease. Ultimately, patients will benefit from increased drug development activity by improving access to ISD therapies with better long-term outcomes.

This effort builds on previous work in the field that has identified clinically relevant measures capable of predicting long-term allograft failure by aggregating and standardizing multiple clinical trials, real-world clinical transplant center datasets, and long-term registry data. A list of prioritized and acquired datasets can be seen in Appendix 3. By leveraging a significant amount of patient level data from these sources, this effort intends to seek regulatory endorsement of a composite measure capable of predicting five-year risk of graft loss using data available within the first year after transplantation. Based on existing literature and the ongoing work of the Paris Transplant Group, the proposed components of this composite score will include eGFR calculated with SCr (referred to as 'eGFR'), measurement of protein excretion into the urine (referred to as 'proteinuria'), pathophysiological assessment of percutaneous kidney graft biopsy (referred to as 'biopsy histology'), and presence or absence of de novo anti-HLA DSA (referred to as 'dnDSA'). A semi-parametric or parametric survival model will be used to develop the composite score to estimate the probability of long-term allograft survival.

To acquire the necessary patient level data to develop a novel reasonably likely surrogate endpoint, the Critical Path Institute's Transplant Therapeutics Consortium (TTC) has led a large data collaboration effort across the field of kidney transplantation. Datasets from relevant clinical trials of ISDs and real-

world data from clinical transplant centers have been prioritized based on the presence of the variables of interest and of long-term outcomes. When possible, datasets that lack long-term follow up will be integrated, through established processes, with the long-term kidney transplant outcome registry managed by the Scientific Registry of Transplant Recipients.

Importantly, as the data collaboration effort across the transplant community is ongoing, the components of the final composite score will be dependent on the datasets ultimately available for inclusion and analysis. Based on preliminary analysis of the currently available datasets and existing literature, it is expected that the four components listed above (i.e., eGFR measured with serum creatinine, measures of protein excretion into the urine, pathological assessment of biopsy histology, and presence or absence of dnDSA) will be included in the final model. Thus, this LOI will include discussion of these four components. As some details regarding these variables or others may not be ascertained until the datasets that will underpin the development and assessment of the composite score have been fully analyzed, it will be noted in this submission where specific details will be provided in future submissions.

Biomarker Information and Interpretation

1. **Biomarker name:**
   The Integrative Box (iBox) Scoring System includes the following component biomarkers, taken together in the first year after transplantation:

   Estimated Glomerular Filtration Rate (‘eGFR’) [Serum Creatinine]: calculated with serum creatinine (a molecular biomarker) and certain patient characteristics using established equations (most commonly the Chronic Kidney Disease Epidemiology Collaboration, CKD-Epi equation);

   Measurement of protein excretion into the urine (‘proteinuria’): Molecular biomarker

   Pathophysiological assessment of percutaneous renal allograft biopsy, based on Banff scoring criteria (‘biopsy histology’): Histologic biomarker

   Presence or absence of de novo anti-human leukocyte antigen donor-specific antibodies. Additionally, the presence of the dnDSA will be refined into categories based on MFI values. The specific categorical cut-points in the MFI scale will be determined in future submissions once the appropriate patient-level data has been curated. It is envisaged these categorical cut-points for subjects with dnDSA will be based on those described in Loupy et al 2019 (Loupy et al. 2019) (Appendix 2): Molecular biomarker

2. **Analytical methods:**
   As stated above, this effort is the result of a significant ongoing data collaboration across multiple stakeholders in the field. Therefore, the specific analytical methods used in the raw measurement of the components of the composite score will be dependent on the sources of the data included in the final analysis and will be fully examined in future qualification plan submissions. The most common analytical method for each component is described here.

eGFR:

Calculated through established equations (most commonly the CDK-Epi equation ("Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C. - PubMed - NCBI" n.d.)) using measurements of serum creatinine, and patient characteristics. Creatinine is measured through Jaffe reaction assays or through enzymatic assays, both colorimetric assays.

Proteinuria:

Proteinuria may be measured in several quantitative or qualitative means. Proteinuria consists of excessive protein in the urine that can be categorized based on type of proteins, quantity of those proteins, and their concentration compared to creatinine. Total urinary protein over 24 hours, total urinary protein to creatinine ratio, and urine albumin to creatinine ratio are the various measures to estimate proteinuria. Assays measuring protein in the urine use turbidimetric, immuno-turbidimetric, or colorimetric. Turbidimetric assays include sulfosalicylic acid (SSA), sulfosalicylic acid with sodium sulphase (SSSS), trichloroacetic acid (TCA), and benzethonium chloride. Colorimetric assays involve the addition of a reagent, such as pyrogallol red molybdate or Coomassie brilliant blue, and spectrophotometric analysis. The ultimate determination of which measure of urinary protein and subsequent assay considerations is data dependent and will be discussed in future submission documents once patient level data have been assessed further (Yalamati, Karra, and Bhongir 2016).

Biopsy Histology:

Percutaneous renal biopsy (PRB) is the current SOC and is often guided by ultrasound and performed under local anesthesia. Automatic disposable spring-loaded devices use needles of various bore sizes to collect the renal sample. After the kidney sample is retrieved, it is manually sectioned, fixed, stained, and analyzed by a pathologist. The Banff Classification ("A 2018 Reference Guide to the Banff Classification of Renal... : Transplantation" n.d.) was developed to provide a schema to analyze signs of acute and chronic rejection through the scoring of kidney lesions. In total, 15 Banff Lesions Scores are defined, which document histopathological changes in the different compartments of the kidney.

dnDSA:

Measurement of dnDSA uses the Luminex Bead-based Multiplex Assay, currently available through two vendors. In this assay, insoluble dye-impregnated beads, which present a predefined HLA class I (HLA-A, HLA-B, HLA-Cw) or class II (HLA-DR, HLA-DQ, HLA-DP) molecule, are incubated with the serum of the allograft recipient. If DSAs are present, they will bind to the cognate antigen(s) on the bead. The bead-DSA complex(es) is then exposed to a fluorescent phycoerythrin-coupled-IgG which binds the complex and is used to measure the presence of DSAs. A negative control serum is used to establish the background value for each bead in a test batch. The measure of this assay is expressed as the mean fluorescent intensity (MFI) of the phycoerythrin-coupled-IgG that is bound to the DSA and is normalized using appropriate controls. MFI values are produced for each DSA type measured ("Detection of HLA Antibodies in Organ Transplant Recipients – Triumphs and Challenges of the Solid Phase Bead Assay" n.d.). Presence of dnDSA will be refined into categories based on MFI values. The specific categorical cut-points in the MFI scale will be determined in future submissions, once the appropriate patient-level data has been curated.

The iBox Scoring System will assess the presence or absence of any dnDSA, and does envisage

using MFI values in a semi-quantitative manner. It is anticipated these categorical cut-points for subjects with dnDSA will be based on those described in Loupy et al 2019 (Loupy et al. 2019)(Appendix 2). The appropriateness of these cut-points will be discussed in future qualification submissions.

3. Measurement units and limit(s) of detection:

Units and limits of detection will be dependent on the datasets included in the final analysis and will be fully described in future qualification submission documents.

4. Biomarker interpretation and utility:

The Integrative Box (iBox) Scoring System

The iBox Scoring System has been developed by estimating individual weights for each of the proposed components (i.e., eGFR, proteinuria, kidney biopsy histology, and the presence or absence of dnDSA). The component measures will be assessed at 12 months post transplantation in a clinical trial and entered into the iBox Scoring System. The determined weighting for each component will be a coefficient in a survival model. The composite score will be the linear combination of weights together with the individual patient features, or

$$\text{Composite Score} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N$$

where β_i is the estimated weight of the patient feature x_i, e.g. eGRF and $i = 1, \ldots, N$ where N is the total number of patient features used to compute the score. The output of the iBox Scoring System is a five-year graft loss risk prediction capable of indicating the long-term performance of therapeutic ISD interventions in a clinical trial. The score will allow for comparative efficacy assessments between study control and intervention study arms, guide regulatory decision making regarding the long-term efficacy of new therapeutic interventions, and open FDA's Accelerated Approval Program. A brief discussion of the interpretation of each component of the iBox follows, and more information can be found in the supporting information section of this submission.

eGFR:

Due to the impact of muscle mass on circulating creatinine levels, estimating equations have been developed to account for differences in muscle mass and translate creatinine concentration to estimated GFR. These equations account for population average differences by age, race, and sex, but cannot account for individual differences. The most widely used equation to estimate GFR is the CKD-Epi equation. The CDK-epi equation includes serum creatinine, gender (Male/female), Age, and race (black/non-black) as follows:

(Levey et al. 2009) **$CKD\text{-}EPI_{creatinine} = A \times (SCr/B)^{c} \times 0.993^{age} \times (1.159 \text{ if black})$**, where A, B, and C are the following:

Female		**Male**	
SCr ≤ 0.7	A = 144	**SCr ≤ 0.9**	A = 141
	B = 0.7		B = 0.9
	C = -0.329		C = -0.411
	A = 144		A = 141
	B = 0.7		B = 0.9

SCr > 0.7	C = -1.209	**SCr > 0.9**	C = -1.209

Current guidelines recommend reporting of creatinine results by clinical laboratories in terms of eGFR

Proteinuria:
UPCR is a common and clinically accepted alternative to 24-hour urine collection, which is otherwise required to assess 24-hour protein excretion rates (Akbari et al. 2014). Urine protein concentration is calculated by measuring the urinary protein and urinary creatinine in a spot first- or second-morning urine sample and dividing the urinary protein measure by the urinary creatinine measure. Units of urinary protein and urinary creatinine are both measured in mg/dL, yielding a unitless ratio. In Loupy et al 2019 (Appendix 2), the albumin to creatinine ratio was used to assess proteinuria. Yet to be published work by this group has demonstrated the performance of the iBox Scoring System with alternate measures of proteinuria, including dipstick, 24-hour urine collection, and urinary protein to creatinine ratio. The specific measure to be included in the iBox Scoring System will ultimately depend on the data available and will be discussed in detail in future submissions.

Biopsy Histology:
After retrieved kidney biopsy samples are appropriately sectioned, fixed, and stained, they are analyzed by a clinical pathologist. While kidney transplant biopsies are subject to issues of sample and inter- and intra-rater variability, the Banff allograft pathology criteria allows for a core series of semi-quantitated features that encompass specific histopathologic entities seen in kidney allografts. Fifteen specific lesions are scored, on a 0-3 scale. In clinical practice, lesion scores are then used in the diagnosis of antibody mediated rejection (AMR), suspicious or borderline acute T-Cell mediated rejection (TCMR), interstitial fibrosis and tubular atrophy (IFTA), or other tissue damages. This effort will consider including all or a pre-specified subset of biopsy lesion scores, depending on analysis of the final datasets available.

dnDSA:
Current assays are available through two vendors, which have somewhat different panels and a somewhat different range of responses. There are three types of kits to detect DSAs, which allow for assessment at varying levels of specificity (e.g., Class I HLA v. HLA-A, v. HLA-DQA). Assays can detect class I (HLA-A, HLA-B, HLA-Cw) and class II (HLA-DR, HLA-DQ, HLA-DP) antibodies using the following types of screening beads; 1) mixed antigen screen beads, wherein a single bead carries a mixture of purified class I and class II molecules from three or more donors; 2) phenotypic beads, that carry multiple HLA class I or class II phenotype antigens purified from a single donor; and 3) single antigen beads, where each bead carries a single recombinant HLA class I or class II antigen/allele.

These assays express units of measurement as MFI of each bead (generally 0 - >20,000) and are used semi-quantitatively in clinical practice. Based on the patient-level data that will be available to the consortium, the presence of the dnDSA will be refined into categories based on MFI values. The specific categorical cut-points will be determined in future submissions once the appropriate patient-level data has been curated. It is anticipated these categorical cut-points for subjects with dnDSA will be based on those described in Loupy et al 2019(Loupy et al. 2019) (Appendix 2). The appropriateness of these cut-points will be discussed in future

qualification submissions. The inter-operability of the assays available commercially from the two vendors is discussed in the analytical considerations section, and the analytical performance characteristics of the assay reagents will be evaluated in full detail in future submissions.

Future qualification submission documents will also assess which specific kits were used to generate dnDSA data in each data set supporting the qualification. These data will inform how dnDSA data will be incorporated into the final composite (i.e., any dnDSA versus a specific HLA Class versus specific HLA alleles).

Context of Use Statement

The Integrative Box (iBox) Scoring System of eGFR calculated with serum creatinine, proteinuria, kidney biopsy histology assessment using the Banff scoring criteria, and presence or absence of de novo anti-HLA donor specific antibodies, taken together in the first year after transplantation is a reasonably likely surrogate endpoint for the five-year risk of allograft loss in kidney transplant patients for use in clinical trial studies evaluating the safety and efficacy of novel immunosuppressive therapies for Accelerated Approval Program submissions.

The target population for this context-of-use will be dependent on patient populations of the datasets that underpin the iBox Scoring System. It is expected that the target population will be refined as data sets are made available to the consortium and analyzed. Analysis and discussion of the target population for the iBox Scoring System will thus be discussed in more detail in the Qualification Plan submission.

Analytical Considerations

As more data are acquired and aggregated, a deeper understanding of assay variation will be developed. Future qualification submission documents in support of the iBox Scoring System will contain full assessment of the analytical considerations for each assay based on the “Points of Consideration Document: Scientific and Regulatory Considerations for the Analytical Validation of the Assays used in the Qualification of Biomarkers in Biological Matrices”. The following represents a brief summary of measurement for each component of the composite marker. It is expected there will be some inter-site variation in the specific assay used for each component. Strategies to enable the integration of patient-level data from each dataset will be discussed in future qualification submissions.

eGFR

The most accepted and direct index of kidney function is the glomerular filtration rate (GFR), which provides a measure of the cumulative volume of plasma that is filtered by the glomeruli in a given time. Historically, the “gold standard” procedure for measuring GFR involved a continuous intravenous infusion of a marker (e.g., inulin, iothalamate), followed by analysis of repeated timed blood draws and urine collections to assess clearance of the marker. Directly measuring GFR with these procedures are burdensome for patients and staff and impractical in many clinical and research settings. As such, estimating GFR (eGFR) using endogenous markers is an attractive alternative and is most commonly used in current practice.

The most commonly used of these endogenous markers is serum creatinine. Measurement of serum creatinine is second only to glucose as the most common analyte in clinical chemistry.

Creatinine is a byproduct of muscle catabolism of creatine. As such, factors affecting muscle mass also impact circulating levels of creatinine (see below). Creatinine has several characteristics that make it a suitable marker of kidney function: it is freely filtered by the glomerulus at a constant rate, with little tubular reabsorption and minimal tubular secretion is most circumstances.

There are two types of assays commercially available and in widespread use for creatinine measurements: those based on the Jaffe reaction (alkaline picrate) and those based on enzymatic methods. In the former, creatinine forms a complex with picric acid in an alkaline solution. The concentration of the colored complex is proportional to the concentration of creatinine in plasma. There is some interference from other endogenous substances, so compensated assays have been developed which correct by a constant to improve accuracy. There are two types of enzymatic assays, with the most common converting creatinine to hydrogen peroxide, which reacts with a dye to generate a colored compound. The others use conversion of creatinine to ammonia.

Harmonization efforts by the College of American Pathologists, the National Kidney Disease Education Program, the European Federation of Chemistry and Laboratory Medicine, and others, and the creation of an international reference standard (SRM 967) by the National Institute of Standards and Technology have led to significant improvements in the accuracy of creatinine measurements. Nearly all assays are now traceable to isotope dilution-mass spectrometry reference standards. This standardization has substantially reduced the bias and minimized the influence of interfering substances in the assays.

Proteinuria

Urinary excretion of protein can be quantified by a 24-hour urine collection or by collection a spot urine sample and calculating UPCR (Yalamati, Karra, and Bhongir 2016). The 24-hour urine collection is the current gold standard method, but sample collection, storage, and transport are cumbersome for the patient and prone to errors in collections. As such, the use of spot urine collections is far more common in clinical practice. Excreted protein levels are indexed to urinary creatinine to correct for variations in urinary concentration due to varying levels of hydration. UPCR is highly correlated with total protein excretion from a 24-hour collection, with somewhat weaker correlation due to imprecision of all methods at very low levels. A first morning urine specimen most strongly correlates with 24-hour protein excretion.

Protein is detected using turbidimetric, immuno-turbidimetric, or colorimetric assays. Turbidimetry involves introduction of a protein precipitant and subsequent quantification of the denatured protein. Immuno-turbidimetry uses antibodies to isolate specific sizes (i.e. albumin) of proteins prior to turbidimetric analysis. The specific precipitant used, the specific proteins present, and the time from introduction of the precipitant to assessment may affect the results. Turbidimetric assays include SSA, SSSS, TCA, and benzethonium chloride. Colorimetric assay involves addition of a reagent, such as pyrogallol red molybdate or Coomassie brilliant blue, and spectrophotometric analysis. Turbidimetric methods of protein quantification are commonly used, but concerns of poor precision and sensitivity and variable response to different proteins in the urine have been reported. Colorimetric methods have better precision and sensitivity, but these assays are also subject to variation in dye-binding to various proteins (Yalamati, Karra, and Bhongir 2016).

Dipstick tests for urinary protein are also available and are inexpensive, easy to use, are highly specific, and provide rough estimates of severity of proteinuria. Dipstick tests, however, are insensitive and only provide semi-quantitative results.

A notable limitation of the precision of urinary protein assessment is the inherent variability in an individual over time. The within person standard deviation is approximately 40-50% of the mean value. Due to this variability, some clinical practice guidelines recommend repeat testing in a specified time period be used for diagnosis.

In Loupy et al 2019 (Appendix 2), the urine albumin to creatinine ratio was used to assess proteinuria. Yet to published work by this group has demonstrated the performance of the iBox Scoring System with alternate measures of proteinuria. Full assessment of the specific measures and assays used to assess urinary protein excretion in each dataset and the intra-operability of these assays will be provided in future qualification submissions.

Biopsy Histology

The PRB is the current SOC and is often guided by ultrasound and performed under local anesthesia. Automatic disposable spring-loaded devices use needles of various bore sizes to collect the renal sample. After the kidney sample is retrieved, it is manually sectioned, fixed, stained, and analyzed. The Banff Classification was developed to provide a schema to assess signs of acute and chronic rejection through the scoring of kidney lesions. In total, 15 Banff Lesions Scores are defined which document histopathological changes in the different compartments of the kidney.

An important consideration with kidney transplant biopsies is inter-observer variability. A recent study from the Mayo Clinic (Smith et al. 2019) involving scoring of biopsies by six different pathologists showed only fair to moderate agreement between any two pathologists (kappa values between 0.38 and 0.48) for Banff scoring of the key histologic lesions of Antibody-mediated rejection (ABMR), glomerulitis, peritubular capillaritis, and transplant glomerulopathy (TG), although kappa values for diagnosis of active and chronic active ABMR, based on pathologists' score of histologic lesions, were better (0.70 and 0.59, respectively). Having three pathologists grade each biopsy and using a "majority rules" approach resulted in improved kappa values for scoring the individual lesions (0.57 - 0.62) and diagnoses (0.82 and 0.70). Overall, however, the inter-observer variability for scoring of the individual Banff lesions is no worse than that reported for an experienced renal pathologist's scoring of individual histologic lesions comprising other histologic classifications, such as the International Society of Nephrology and the Renal Pathology Society classification of lupus nephritis and Oxford classification of IgA nephropathy (Furness and Taub 2006; Working Group of the International IgA Nephropathy Network and the Renal Pathology Society et al. 2009). Furthermore, a major focus at the 2019 Banff conference, with input from both pathologists and transplant clinicians, was the clarifications of definitions for and reporting of individual Banff lesions, as well as of TCMR (borderline, acute, chronic active) and ABMR (active, chronic active), and these modifications will be included in the Banff 2019 meeting report.

Appropriate assessments of inter- and intra-observer variability will be performed on the final data package. Specific details of this assessment will follow in future qualification submissions.

dnDSA

The development of dnDSA occurs in response to HLA antigens on the transplanted organ and do not exist in the recipient prior to transplantation. The details of the currently available Luminex Bead-based Multiplex Assays are described above. In short, the assay uses color coded beads that contain antibodies for each HLA allele protein in each MHC molecule class. Kits that measure different levels of specificity are available allowing for measuring the broader allele or more specific antigen level. The bead-antibody complex captures any dnDSAs present in the transplant recipient's serum. After washing, a luminescent phycoerythrin conjugate is then added and binds to the bound analyte specific antibodies. Dual lasers are then used to detect the identify of any present donor specific antibodies and the relative fluorescent intensity. There are several reagent and serum specific considerations that must be made with these assays.

As previously discussed, current assays are available through two vendors which have somewhat different panels and a somewhat different range of responses. There are three types of kits to detect dnDSAs, which allow for assessment of varying levels of specificity (e.g., antigen v. allele) depending on the types of beads being used. Bead types include mixed antigen screen beads, wherein a single bead carries a mixture of purified class I and class II molecules from three or more donors, phenotypic beads, that carry multiple HLA class I or class II phenotype antigens purified from a single donor, and single antigen beads, where each bead carries a single recombinant HLA class I or class II antigen/allele.

These assays have been cleared through CDRH's 501(k) process as in vitro diagnostics with FDA. MFI values reported by these assays are routinely used in transplant clinical practice in a semi-quantitative manner to inform clinical decision making. The presence of dnDSA will be refined into categories based on MFI values, according to the patient level data available to the consortium. The specific categorical cut-points in the MFI scale will be determined in future submissions once the appropriate patient-level data has been curated. It is anticipated these categorical cut-points for subject with dnDSA will be based on those describes in Loupy et al 2019(ref) (Appendix 2) but that the appropriateness of these cut-points will be discussed in future submissions. As part of the Consortium's ongoing data aggregation effort, the specific assay type, sample handling protocols, and raw MFI measures are being requested from data contributors. These data will allow an appropriate set of criteria for the comparison of assays between datasets to be established. These criteria will be developed with input of FDA and defined in future submissions.

The Consortium is aware of four limitations of this assay and will require the full documentation from data contributors before final determination of how these limitations will affect the qualification effort. This will be discussed in detail in a future submission.

There have been significant efforts to minimize the inter-laboratory variability of anti-HLA DSA testing. Each year the American Society for Histocompatibility and Immunogenetics proficiency testing provides clinically based samples to assess a participating laboratory's ability to accurately perform their analyses ("Proficiency Testing Program - American Society for Histocompatibility and Immunogenetics" n.d.). The results from these yearly assessments are made publicly available. The core laboratories of the Clinical Trials in Organ

Transplantation (CTOT), a collaborative clinical research organization headquartered at the National Institute of Allergy and Infectious Diseases, has also sought to improve anti-HLA DSA assay performance across the CTOT laboratories (Reed et al. 2013a; 2013b). These efforts have demonstrated and published intra- and inter-laboratory variability. This study determined several factors that contribute to overall variability in the inter-laboratory assay results, including the individual center, which manufacturer or kit was utilized, and reagent lots.

Furthermore, saturation of the bead-bound antibody has been observed in when higher levels of DSA are present than the number of HLA antigen targets on the beads (Anat R. Tambur and Wiebe 2018). This presents a significant a challenge to the use of this assay in a semi-quantitative or qualitative manner. This limitation can be overcome by serum dilution to ensure bead saturation does not occur.

In addition, it has been proposed that inhibition of the reporter antibody can occur due to high levels of complement in the serum that binds free DSA in solution and results in a reduced MFI value (prozone effect) (A. R. Tambur et al. 2015). This limitation is frequently overcome with pre-treatment of the serum with EDTA, other reagents, or dilution or titration studies.

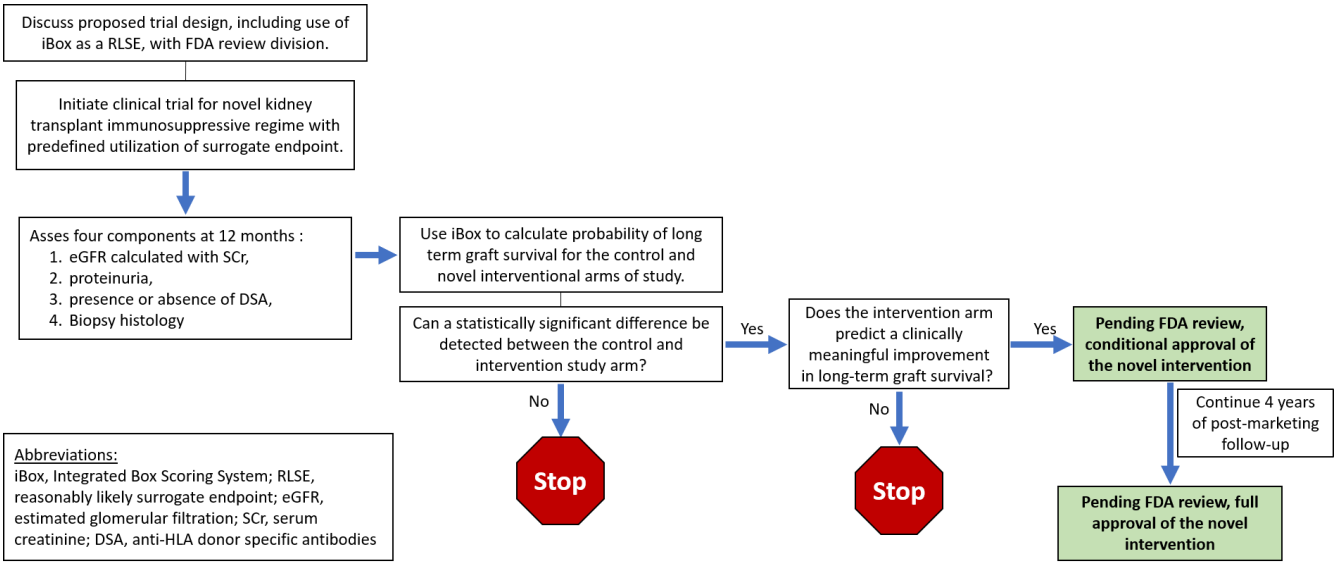
Finally, due to the nature of HLA antigen polymorphism, many HLA antigens share significant portions of their protein sequence (Anat R. Tambur et al. 2018). Thus, antibodies can recognize a target/epitope that is shared by several HLA antigens. As a result, a specific HLA antigen may bind to several beads. Since MFI is measured per single bead, the results may underestimate antibody strength. HLA recognition patterns are possible to detect to compensate for cross bead binding, but this limitation poses a significant challenge for interpreting the data from this assay in a semi-quantitative format.

Considerations will be given to each of these limitations in future submissions when the available patient-level data has been curated and the associated assay data has been gathered and documented.

Clinical Considerations

At the 12-month time point post-transplantation, the composite score will be calculated for each individual in both the study control and interventional arms. The mean composite scores between the two arms will then be calculated. The survival model will take as an input each mean composite score and will be used to predict the difference (if one exists) between the five-year failure rates between the two arms.

Figure 1: Decision tree for use of the Integrative Box (iBox) Scoring System as a reasonably likely surrogate endpoint



It is intended that the iBox Scoring System will be used as a reasonably likely surrogate endpoint in clinical trials evaluating the safety and efficacy of novel ISD therapies or regimens as part of FDA Accelerated Approval Program submissions. The patient population that will be included in the final context-of-use for this composite measure will ultimately be dependent on the data sets available to support the development of the tool. Future qualification submissions will include detailed analysis of the populations included in the data and therefore which populations will be included in the final context-of-use.

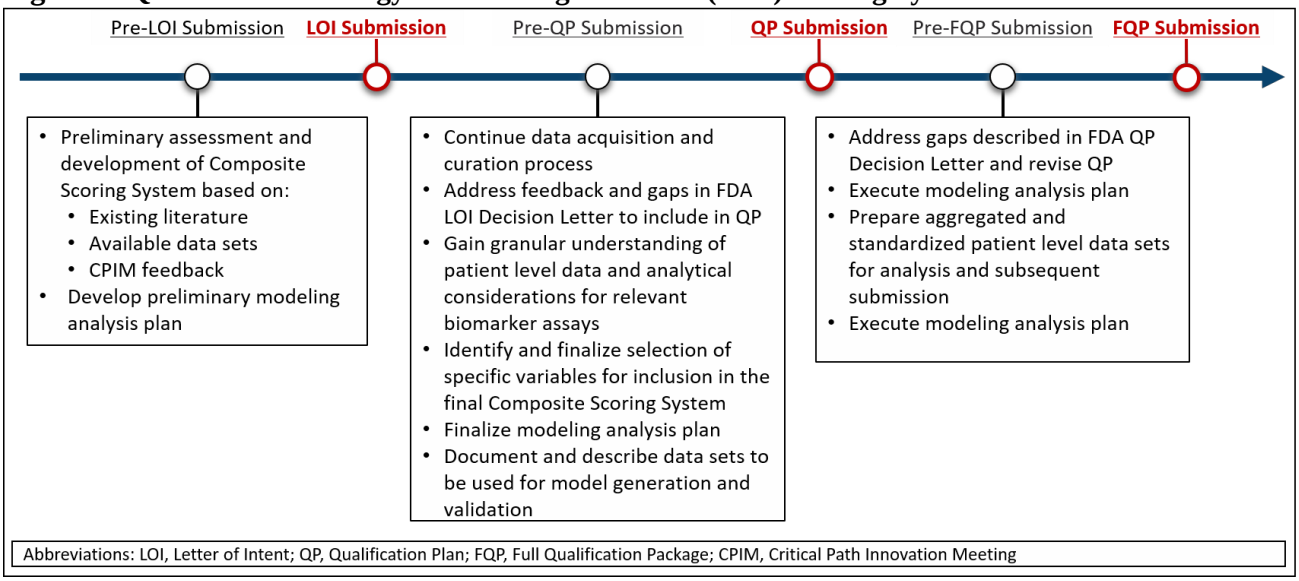
There are inherent risks to drug developers and to patients associated with the use of any surrogate or reasonably likely surrogate endpoint in clinical trials. If a surrogate does not actually reflect true long-term treatment effects, new therapeutic agents may be approved for use despite no evidence of long-term efficacy. This risk is mitigated through post-marketing confirmatory studies required by FDA's Accelerated Approval Program and by analysis of short-term outcomes at the time of long-term risk assessment. The proposed scoring system will improve upon the current standing by allowing for short-term and long-term assessments of drug efficacy.

The potential benefits of the iBox Scoring System are numerous. With ten-year graft survival rates as low as 50% in some kidney transplant populations, there is significant need for better ISDs with improved long-term outcomes. However, industry has be reticent to undergo the long-term clinical trials required to assess long-term efficacy. Thus, qualification of a reasonably likely surrogate endpoint would offer drug sponsors a mechanism to access FDA's Accelerated Approval Program, thus reinvigorating and incentivizing drug development in this therapeutic area. This directly translates to improved patient care by providing a mechanism for novel ISDs with improved outcomes to reach patients faster.

The current gaps of this effort are largely centered around the ongoing data sharing and collaboration work, discussed in more detail below. Specifically, information that will be required for future

qualification submission documents will ultimately depend on data sets available to be included in the final analysis. Further, while previous work in this field has identified measures important for predicting long-term outcomes, this effort must re-analyze the available data to confirm the predictive and prognostic capabilities of the individual components, as well as other predictors that should be included in the final composite score. Figure 2 provides a brief overview of this Consortium's plan to address these gaps in relation to future qualification submissions. Following submission of this LOI, the future Qualification Plan submission will include detailed analyses of the datasets that underpin the iBox Scoring System, including appropriate target population, predictive variables to be included in the Composite Scoring System, biomarker assay considerations of the identified variables, and aggregated data sets to be used for model generation and validation.

Figure 2. Qualification Strategy for the Integrative Box (iBox) Scoring System



Supporting Information

Summary of previous composite measures to predict kidney transplant outcomes

It is well established in the literature that individual markers of kidney transplant health are insufficient to predict long-term outcomes with acceptable accuracy. Thus, significant prior efforts have attempted to develop Composite Scoring Systems better able to predict the long-term health of the graft. A 2017 meta-analysis reviewed risk prediction models for graft failure in kidney transplantation (Kaboré et al. 2017). It is important to note that all of these risk prediction models have been geared towards clinical decision making, and none has been evaluated or validated for use in ISD development.

A recent meta-analysis identified 39 risk prediction models published in the scientific literature from 2005-2015. The majority of these studies aimed to predict graft failure (generally defined as dialysis, re-transplantation, or death with functioning graft) or death censored graft failure (defined as dialysis/re-transplantation).

Of these risk prediction models, 14 studies included predictors measured during the post-transplant period, with or without pre-transplant risk factors. Post-transplant predictors included in these studies

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 and carotid-femoral pulse wave velocity. Finally, previous risk prediction modeling efforts have attempted to predict short term (1-4 years) and long-term (≤ 5 years) outcomes.

Although important and significant work has been done to predict individual patient outcomes, as captured by Kabore et al, the validation and overall assessment of these efforts have been limited, and these efforts have not been validated for use in drug development.

More recently developed composite scores that have sought to predict long-term graft loss in individual patients have incorporated elements of the patient demographics and pathophysiological measures. Several recent composite scores are described below.

In 2014, Shabir et al (Shabir et al. 2014,) developed a new prediction instrument to predict five-year risk of kidney transplant failure using data available at one-year post transplant. This effort utilized clinical data from 651 patients from Birmingham, United Kingdom to develop a model capable of predicting death-censored and overall transplant failure at five-years post transplantation. This model, called the Birmingham Risk Score, incorporates recipient sex, age, and ethnicity, history of acute rejection, and one-year post transplant measurements of eGFR, albumin, and urine albumin to creatinine ratio. The model was then validated in 3 international cohorts, including 787 patients from Leeds, United Kingdom, 736 patients from Tours, France, and 475 patients 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.

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 histology and/or antibody evaluations. In this work, 1465 adults from the Mayo Clinic in Rochester, MN 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 donor-specific antibodies into the original model failed to improve predictability of the model, presence of glomerulitis or chronic interstitial fibrosis on a one-year surveillance biopsy improved the model predictability (C-statistic = 0.90), calibration, and resulted in the reclassification of the graft failure risk in 29% of patients. 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 patients with the presence of DSA. A recent effort has further validated the Birmingham-Mayo model in an additional cohort of patients with the presence of DSA (Bentall et al. 2019).

Building on these previous efforts, Loupy et al (Loupy et al. 2019) leveraged the nationalized health care system in France to prospectively follow long-term outcomes of kidney transplant patients in to develop a new risk prediction model capable of predicting risk of graft loss at 3, 5, and 7-years post-transplant. The model was subsequently validated in 2 international cohorts, three Phase II and III clinical trials, and in numerous clinical scenarios. The derivation cohort included 4000 consecutive patients over 18 years of age from four centers across France with a median follow up time of 7.65 years. Quantitative analyses

were performed to identify predictors of long-term outcomes. A scoring system, termed the iBox, was then developed using the identified predictors, which include eGFR, proteinuria, kidney biopsy histology, and presence or absence of donor specific antibodies. The performance of the iBox scoring system was then evaluated in two validation cohorts (n = 3557) from the United States and Europe. Overall, model performance showed good calibration and discrimination (C index 0.81, 95% confidence interval 0.79 to 0.83). The model was also validated against three phase II or III clinical trials, with C-indices determined to be 0.87, 0.82, and 0.92 in each of the three studies. Further, the risk score was shown to accurately predict the actual observations of graft loss in these studies. The model was then assessed in multiple clinical scenarios and in different subpopulations with acceptable performance characteristics in each scenario and population, with C-statistics that ranged between 0.78 and 0.84, depending on the scenario.

The iBox Scoring System represents the most advanced prediction system for determining long-term risk of graft-loss following a kidney transplantation. This work has recently been published and is available in Appendix 2. However, the significant and ongoing external validation efforts have focused on the use of the iBox Scoring System in clinical practice. To date, the iBox Scoring System has not yet been validated for use in drug development. The current effort seeks to build on the existing clinical validation of the iBox scoring system in order to expand its use into the drug development process.

Each component of the iBox Scoring System is individually biologically linked to key aspects of kidney health and kidney allograft function, as described below. However, taken together the composite gives broader biological insight into the current health of the kidney and the pathologies that lead to allograft loss than the individual components in isolation. The individual components provide distinct information on the health status of the graft through measurements of allograft function (eGFR and proteinuria), direct assessment of allograft health (biopsy histology), and the recipient's immune response to the transplanted organ (presence or absence of dnDSA). Combining these individual measures into one composite score allows for improved and more robust predictions of long-term graft survival than is possible with any individual component. A discussion of each individual component can be found below.

eGFR

The most accepted and direct index of kidney function is the glomerular filtration rate (GFR), which provides a measure of the cumulative volume of plasma that is filtered by the glomeruli in a given time. Historically, the "gold standard" procedure for measuring GFR involved a continuous intravenous infusion of a marker (e.g., inulin, iothalamate), followed by analysis of repeated timed blood draws and urine collections to assess clearance of the marker. Directly measuring GFR with these procedures are burdensome for patients and staff and impractical in many clinical and research settings. As such, estimating GFR (eGFR) using endogenous markers is an attractive alternative and this is most commonly used in current practice.
eGFR is a widely used marker of kidney function in clinical practice and clinical trials for most or all kidney related disorders, including kidney transplantation.

Proteinuria

Under normal conditions, waste products are filtered through the glomeruli, while larger proteins are selectively conserved. Smaller proteins that may be filtered through the glomeruli are

reabsorbed in the proximal tubule. As a result of these processes, excretion of protein in the urine is minimal in healthy kidneys. Excretion of more than small amounts of protein into the urine indicates excessive passage through the glomerulus, thought to be primarily due to endothelial cell injury, and/or decreased reabsorption by epithelial cells in the proximal tubule. Assessments of protein excretion (including urinary albumin) is recognized as an important identifier or chronic kidney disease. In kidney transplantation, even mild proteinuria has been demonstrated to be predictive of decreased long-term graft function (Hohage et al. 1997) and has been associated with overall patient mortality (Roodnat et al. 2001).

Biopsy Histology

Despite the inter- and intra-observer variability, several studies have shown that assessment of the kidney transplant biopsy histology predicts outcomes, and when included in composite models, has been shown to be predictive in multivariate analyses. In spite of its limitations, the grading of biopsies has provided a common ground in diagnosis and further provides opportunities to identify relationships between specific features and clinical outcomes.

Using the Banff criteria definitions, TCMR on one-year surveillance biopsy has been shown to be an independent risk factor for graft loss (Randhawa 2015). The persistence of TCMR has been associated with a substantial risk of graft failure (HR 4.88) in a Mayo Clinic Study that included nearly 800 kidney transplants, and several other studies (Gago et al. 2012; El-Zoghby et al. 2009; M. Naesens et al. 2013). When co-occurring with antibody mediated rejection (ABMR), the presence of TCMR is an independent risk factor for allograft failure (Matignon et al. 2012). Even when TCMR is characterized predominantly by vasculitis in the absence of extensive inflammation (i) or tubulitis (t), graft loss is significantly more frequent (Sis et al. 2015; Wu et al. 2014). The incidence of concomitant findings of TCMR and ABMR have been variable, but their combined presence is a poor prognostic feature, even when the level of tubulitis is minimal (Matignon et al. 2012; Rodrigues et al. 2014). This is further discussed below.

Another phenomenon now appreciated with poor prognosis is the presence of inflammation or tubulitis in areas of IFTA. Indeed, this finding led to the adoption of a total inflammation score (Mengel et al. 2009). Long-term risk outcome has been strongly associated with allograft biopsy histology and specifically the presence of IFTA, microvascular injury and tubulitis+inflammation in non-scarred areas (Loupy et al. 2019). Multiple studies have found the presence of inflammation in scarred areas (i-IFTA) of biopsies performed for allograft dysfunction or proteinuria (Mannon et al. 2010; Matas et al. 2019) or on surveillance biopsies (i.e., those performed per protocol, independent of transplant functional status) to be an independent predictor of allograft failure (Lefaucheur et al. 2018; Nankivell et al. 2018).

A number of investigators have identified specific histopathologic features of ABMR that have a negative prognosis for graft outcome. While the presence of C4d immunostaining in ABMR is associated with higher rates of allograft loss compared to C4d negative ABMR (Orandi et al. 2016; Gaston et al. 2010; Sis et al. 2009; Willicombe et al. 2011), the latter is also associated with shortened allograft survival. The presence of IFTA also portends a negative outcome for the graft. IFTA was an independent predictor of allograft loss (HR, 2.93; 95% CI, 1.62 to 5.29;

P<0.001) in a cohort of 278 kidney transplant recipients with active ABMR, as was the presence of TG (HR, 2.25;95% CI, 1.29 to 3.92; P=0.004), features often not responsive to treatment (Viglietti et al. 2018). Similarly, Moktefi (Moktefi et al. 2017) and Haas (Haas et al. 2017) identified IFTA at the time of ABMR biopsy to be a strong independent risk factor for death censored allograft failure.

The presence of coincident TCMR with ABMR also has a negative prognosis. This may be in part due to the frequency of both ABMR and cell mediated changes in kidney transplant recipients with non-adherence (Wiebe et al. 2012). Cell mediated rejection was identified as an independent risk factor for graft loss in C4d positive ABMR (Matignon et al. 2012), and with a trend to statistical significance in multivariable analysis in a similar sized ABMR cohort that included C4d negative biopsies (Haas et al. 2017). Similarly, the presence of vasculitis ("v") or interstitial inflammation ("i") plus tubulitis ("t") scores > 3 in ABMR biopsies were important independent risk factors for graft failure in a cohort of kidney transplant recipients with biopsy proven clinical rejection (Lefaucheur et al. 2013), highlighting the potential contribution of cell-mediated injury in the outcome of ABMR.

dnDSA

The development of dnDSA post transplantation is considered an important marker to predict the likelihood of allograft loss (Everly et al. 2013; Anat R. Tambur and Wiebe 2018). Physiologically, HLA antigens on the donor allograft represent the major target for the recipient's immune system. Recipient T-cell recognition of the allograft stimulates a cascade of immunologic events that ultimately results in a humoral immune response and the production of antibodies that specifically target HLA molecules on the transplanted organ, i.e., anti-HLA DSA. Thus, unlike the previously discussed features of the composite biomarker, measurements of the presence or absence of DSA are not necessarily indicators of current health or function of the allograft, but a marker of the recipient's immune response to the transplanted organ. The currently available assays allow for semi-quantitative analysis of the magnitude of immune response; however, the literature disagrees as to the overall predictability of the presence or absence of DSA compared to the magnitude of immune response as determined by MFI (Lee et al. 2009; Anat R. Tambur and Wiebe 2018).

Integrative Box (iBox) Scoring System

To build a robust data package capable of supporting the utility of the proposed surrogate marker, the Transplant Therapeutics Consortium has identified a diverse set of clinical trial data and real-world data from clinical transplant centers that capture a wide range of variables, including eGFR, proteinuria, biopsy histology, and dnDSA, which, when taken together within the first year post-transplantation are potentially predictive of long-term graft loss. The variety in datasets and scope of variability included will be fundamental to developing sufficient evidence to support the biological plausibility, causality, universality, proportionality, and specificity of the marker. When necessary, to demonstrate the linkage between the composite surrogate and the long-term outcome of allograft survival, shorter-term datasets (e.g., clinical trials of 24- month duration) will be linked to data from the long-term kidney transplant registry managed by the Scientific Registry of Transplant Recipients.

Key variables for analysis include, but are not limited to: core variables of interest (creatinine, eGFR, proteinuria, DSA, biopsy histology via Banff component scores), baseline parameters (cold ischemia time, delayed allograft function), outcome measures (cause of allograft loss, allograft loss date, death cause, death date, acute rejection status, loss to follow up, adverse event and safety data), recipient parameters (age at transplant, gender, ethnicity, end stage renal disease cause and classification, blood type, comorbid conditions, concomitant medications, dates of dialysis, time since dialysis), and donor parameters (age, gender, ethnicity, donor type: living or deceased), expanded criteria donor status, blood type, kidney donor profile index, kidney donor risk index, baseline creatinine, hypertension status, diabetes status, cause of death, hepatitis C status) and other supporting information (clinical trial ID numbers, publications).

Currently, 17 clinical trial datasets from eight pharmaceutical organizations, and datasets from five clinical transplant centers have been identified and prioritized by the TTC to support the regulatory development of the proposed surrogate marker. While the majority of these datasets are one to three years in duration, they will be linked through established processes to the long- term registry database in order to capture long term outcomes.

Data received from clinical transplant centers will contain an inherent heterogeneity reflective of the diverse kidney transplant recipient population, and thus represent a rich source of long-term data. It is expected that some data received will not be included in the final modeling analysis due to normal and expected issues with standardization and aggregation across datasets. To date, the TTC has acquired five clinical trial data sets and three from clinical transplant centers representing data from over 10,000 patients. A full list of datasets already acquired by the TTC can be found in Appendix 2.

Acquired individual datasets will be curated and integrated into one aggregated database. A subset of variables for analysis will be selected from the aggregated database and used to develop the composite risk scoring system. A semi-parametric or parametric survival modeling approach will be used to develop the composite score system, and the final model will be validated by an independent dataset or through cross-validation. The assessment of which datasets will be used for derivation and validation will be described in the Qualification Plan submission.

Previous Qualification Interactions and Other Approvals

The TTC, as a public-private partnership with FDA, has previously requested and been assigned an FDA-liaison. TTC's current FDA liaison has participated in consortium meetings, including those discussing the proposed biomarker and its use in drug development.

On March 26, 2019, the TTC held a Critical Path Innovation Meeting with FDA to discuss the potential utility of the composite score as part of a clinical trial simulation tool. In this meeting, there was broad agreement of the unmet clinical needs and of the drug development needs that act as barriers to the development of novel ISDs for use following kidney transplantation. Previous efforts at developing clinically focused risk prediction scores were discussed, and the Agency encouraged the TTC to pursue the formal qualification of a composite marker as a surrogate or reasonably likely surrogate for use in

clinical trials.

The TTC intends to engage European regulators at the European Medicines Agency (EMA) once sufficient data are acquired to support a Briefing Package submission towards the Qualification of Novel Methodologies in Drug Development program at EMA. As of the time of this submission, no interactions with EMA have been held.

List of Appendices:

- Appendix 1. References
- Appendix 2. BMJ Publication (Loupy et al. 2019)
- Appendix 3. TTC Data Summary Table
- Appendix 4. FDA-TTC CPIM Summary

Appendix 1. References

"A 2018 Reference Guide to the Banff Classification of Renal... : Transplantation." n.d. Accessed December 12, 2019. https://journals.lww.com/transplantjournal/Fulltext/2018/11000/A_2018_Reference_Guide_to_the_Banff_Classification.14.aspx.

Akbari, Ayub, Dean Fergusson, Madzouka B. Kokolo, Tim Ramsay, Andrew Beck, Robin Ducharme, Marcel Ruzicka, Amanda Grant-Orser, Christine A. White, and Greg A. Knoll. 2014. "Spot Urine Protein Measurements in Kidney Transplantation: A Systematic Review of Diagnostic Accuracy." *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 29 (4): 919–26. https://doi.org/10.1093/ndt/gft520.

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.

"Detection of HLA Antibodies in Organ Transplant Recipients – Triumphs and Challenges of the Solid Phase Bead Assay." n.d. Accessed December 12, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5146910/.

El-Zoghby, Z. M., M. D. Stegall, D. J. Lager, W. K. Kremers, H. Amer, J. M. Gloor, and F. G. Cosio. 2009. "Identifying Specific Causes of Kidney Allograft Loss." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 9 (3): 527–35. https://doi.org/10.1111/j.1600-6143.2008.02519.x.

"Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin C. - PubMed - NCBI." n.d. Accessed December 12, 2019. https://www.ncbi.nlm.nih.gov/pubmed/22762315.

Everly, Matthew J., Lorita M. Rebellato, Carl E. Haisch, Miyuki Ozawa, Karen Parker, Kimberly P. Briley, Paul G. Catrou, et al. 2013. "Incidence and Impact of de Novo Donor-Specific Alloantibody in Primary Renal Allografts." *Transplantation* 95 (3): 410–17. https://doi.org/10.1097/TP.0b013e31827d62e3.

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.

Furness, Peter N., and Nick Taub. 2006. "Interobserver Reproducibility and Application of the ISN/RPS Classification of Lupus Nephritis-a UK-Wide Study." *The American Journal of Surgical Pathology* 30 (8): 1030–35. https://doi.org/10.1097/00000478-200608000-00015.

Gago, M., L. D. Cornell, W. K. Kremers, M. D. Stegall, and F. G. Cosio. 2012. "Kidney Allograft Inflammation and Fibrosis, Causes and Consequences." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 12 (5): 1199–1207. https://doi.org/10.1111/j.1600-6143.2011.03911.x.

Gaston, Robert S., J. Michael Cecka, Bert L. Kasiske, Ann M. Fieberg, Robert Leduc, Fernando C. Cosio, Sita Gourishankar, et al. 2010. "Evidence for Antibody-Mediated Injury as a Major Determinant of Late Kidney Allograft Failure." *Transplantation* 90 (1): 68–74. https://doi.org/10.1097/TP.0b013e3181e065de.

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.

Haas, Mark, James Mirocha, Nancy L. Reinsmoen, Ashley A. Vo, Jua Choi, Joseph M. Kahwaji, Alice Peng, Rafael Villicana, and Stanley C. Jordan. 2017. "Differences in Pathologic Features and Graft Outcomes in Antibody-Mediated Rejection of Renal Allografts Due to Persistent/Recurrent

versus de Novo Donor-Specific Antibodies." *Kidney International* 91 (3): 729–37. https://doi.org/10.1016/j.kint.2016.10.040.

Hart, A., J. M. Smith, M. A. Skeans, S. K. Gustafson, A. R. Wilk, S. Castro, A. Robinson, et al. 2019. "OPTN/SRTR 2017 Annual Data Report: Kidney." *American Journal of Transplantation* 19 (S2): 19–123. https://doi.org/10.1111/ajt.15274.

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.

Hohage, H., U. Kleyer, D. Brückner, C. August, W. Zidek, and C. Spieker. 1997. "Influence of Proteinuria on Long-Term Transplant Survival in Kidney Transplant Recipients." *Nephron* 75 (2): 160–65. https://doi.org/10.1159/000189525.

Howell, Martin, Allison Tong, Germaine Wong, Jonathan C. Craig, and Kirsten Howard. 2012. "Important Outcomes for Kidney Transplant Recipients: A Nominal Group and Qualitative Study." *American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation* 60 (2): 186–96. https://doi.org/10.1053/j.ajkd.2012.02.339.

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: Official Publication of the European Dialysis and Transplant Association - European Renal Association* 32 (suppl_2): ii68–76. https://doi.org/10.1093/ndt/gfw405.

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

Lee, Po-Chang, Lan Zhu, Paul I. Terasaki, and Matthew J. Everly. 2009. "HLA-Specific Antibodies Developed in the First Year Posttransplant Are Predictive of Chronic Rejection and Renal Graft Loss." *Transplantation* 88 (4): 568–74. https://doi.org/10.1097/TP.0b013e3181b11b72.

Lefaucheur, Carmen, Clément Gosset, Marion Rabant, Denis Viglietti, Jérôme Verine, Olivier Aubert, Kevin Louis, et al. 2018. "T Cell-Mediated Rejection Is a Major Determinant of Inflammation in Scarred Areas in Kidney Allografts." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 18 (2): 377–90. https://doi.org/10.1111/ajt.14565.

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* 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., Lesley A. Stevens, Christopher H. Schmid, Yaping Lucy Zhang, Alejandro F. Castro, Harold I. Feldman, John W. Kusek, et al. 2009. "A New Equation to Estimate Glomerular Filtration Rate." *Annals of Internal Medicine* 150 (9): 604–12. https://doi.org/10.7326/0003-4819-150-9-200905050-00006.

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*, September, l4923. https://doi.org/10.1136/bmj.l4923.

Mannon, R. B., A. J. Matas, J. Grande, R. Leduc, J. Connett, B. Kasiske, J. M. Cecka, et al. 2010. "Inflammation in Areas of Tubular Atrophy in Kidney Allograft Biopsies: A Potent Predictor of Allograft Failure." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 10 (9): 2066–73. https://doi.org/10.1111/j.1600-6143.2010.03240.x.

Matas, Arthur J., Ann Fieberg, Roslyn B. Mannon, Robert Leduc, Joe Grande, Bertram L. Kasiske, Michael Cecka, et al. 2019. "Long-Term Follow-up of the DeKAF Cross-Sectional Cohort Study." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 19 (5): 1432–43. https://doi.org/10.1111/ajt.15204.

Matignon, Marie, Thangamani Muthukumar, Surya V. Seshan, Manikkam Suthanthiran, and Choli Hartono. 2012. "Concurrent Acute Cellular Rejection Is an Independent Risk Factor for Renal Allograft Failure in Patients with C4d-Positive Antibody-Mediated Rejection." *Transplantation* 94 (6): 603–11. https://doi.org/10.1097/TP.0b013e31825def05.

Mengel, M., J. Reeve, S. Bunnag, G. Einecke, G. S. Jhangri, B. Sis, K. Famulski, L. Guembes-Hidalgo, and P. F. Halloran. 2009. "Scoring Total Inflammation Is Superior to the Current Banff Inflammation Score in Predicting Outcome and the Degree of Molecular Disturbance in Renal Allografts." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 9 (8): 1859–67. https://doi.org/10.1111/j.1600-6143.2009.02727.x.

Moktefi, Anissa, Juliette Parisot, Dominique Desvaux, Florence Canoui-Poitrine, Isabelle Brocheriou, Julie Peltier, Vincent Audard, et al. 2017. "C1q Binding Is Not an Independent Risk Factor for Kidney Allograft Loss after an Acute Antibody-Mediated Rejection Episode: A Retrospective Cohort Study." *Transplant International: Official Journal of the European Society for Organ Transplantation* 30 (3): 277–87. https://doi.org/10.1111/tri.12905.

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, M., D. R. J. Kuypers, K. De Vusser, Y. Vanrenterghem, P. Evenepoel, K. Claes, B. Bammens, B. Meijers, and E. Lerut. 2013. "Chronic Histological Damage in Early Indication Biopsies Is an Independent Risk Factor for Late Renal Allograft Failure." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 13 (1): 86–99. https://doi.org/10.1111/j.1600-6143.2012.04304.x.

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.

Nankivell, Brian J., Meena Shingde, Karen L. Keung, Caroline L.-S. Fung, Richard J. Borrows, Philip J. O'Connell, and Jeremy R. Chapman. 2018. "The Causes, Significance and Consequences of Inflammatory Fibrosis in Kidney Transplantation: The Banff i-IFTA Lesion." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 18 (2): 364–76. https://doi.org/10.1111/ajt.14609.

"National Data - OPTN." n.d. Accessed December 12, 2019. https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/#.

Orandi, B. J., N. Alachkar, E. S. Kraus, F. Naqvi, B. E. Lonze, L. Lees, K. J. Van Arendonk, et al. 2016. "Presentation and Outcomes of C4d-Negative Antibody-Mediated Rejection After Kidney Transplantation." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 16 (1): 213–20.

https://doi.org/10.1111/ajt.13434.
"Proficiency Testing Program - American Society for Histocompatibility and Immunogenetics." n.d. Accessed November 22, 2019. https://www.ashi-hla.org/page/PT.
Randhawa, Parmjeet. 2015. "T-Cell-Mediated Rejection of the Kidney in the Era of Donor-Specific Antibodies: Diagnostic Challenges and Clinical Significance." *Current Opinion in Organ Transplantation* 20 (3): 325–32. https://doi.org/10.1097/MOT.0000000000000189.
Reed, E. F., P. Rao, Z. Zhang, H. Gebel, R. A. Bray, I. Guleria, J. Lunz, et al. 2013a. "Comprehensive Assessment and Standardization of Solid Phase Multiplex-Bead Arrays for the Detection of Antibodies to HLA." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 13 (7): 1859–70. https://doi.org/10.1111/ajt.12287.
———. 2013b. "Comprehensive Assessment and Standardization of Solid Phase Multiplex-Bead Arrays for the Detection of Antibodies to HLA-Drilling down on Key Sources of Variation." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 13 (11): 3050–51. https://doi.org/10.1111/ajt.12462.
Rodrigues, C. A., M. F. Franco, M. P. Cristelli, J. O. Pestana, and Jr H. Tedesco-Silva. 2014. "Clinicopathological Characteristics and Effect of Late Acute Rejection on Renal Transplant Outcomes." *Transplantation* 98 (8): 885–92. https://doi.org/10.1097/TP.0000000000000145.
Roodnat, J. I., P. G. Mulder, J. Rischen-Vos, I. C. van Riemsdijk, T. van Gelder, R. Zietse, J. N. IJzermans, and W. Weimar. 2001. "Proteinuria after Renal Transplantation Affects Not Only Graft Survival but Also Patient Survival." *Transplantation* 72 (3): 438–44. https://doi.org/10.1097/00007890-200108150-00014.
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, Banu, Serena M. Bagnasco, Lynn D. Cornell, Parmjeet Randhawa, Mark Haas, Belinda Lategan, Alex B. Magil, et al. 2015. "Isolated Endarteritis and Kidney Transplant Survival: A Multicenter Collaborative Study." *Journal of the American Society of Nephrology : JASN* 26 (5): 1216–27. https://doi.org/10.1681/ASN.2014020157.
Sis, Banu, Gian S. Jhangri, Sakarn Bunnag, Kara Allanach, Bruce Kaplan, and Philip F. Halloran. 2009. "Endothelial Gene Expression in Kidney Transplants with Alloantibody Indicates Antibody-Mediated Damage despite Lack of C4d Staining." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 9 (10): 2312–23. https://doi.org/10.1111/j.1600-6143.2009.02761.x.
Smith, Byron, Lynn D. Cornell, Maxwell Smith, Cherise Cortese, Xochiquetzal Geiger, Mariam P. Alexander, Margaret Ryan, et al. 2019. "A Method to Reduce Variability in Scoring Antibody-Mediated Rejection in Renal Allografts: Implications for Clinical Trials - a Retrospective Study." *Transplant International: Official Journal of the European Society for Organ Transplantation* 32 (2): 173–83. https://doi.org/10.1111/tri.13340.
Tambur, A. R., N. D. Herrera, K. M. K. Haarberg, M. F. Cusick, R. A. Gordon, J. R. Leventhal, J. J. Friedewald, and D. Glotz. 2015. "Assessing Antibody Strength: Comparison of MFI, C1q, and Titer Information." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 15 (9): 2421–30. https://doi.org/10.1111/ajt.13295.
Tambur, Anat R., Patricia Campbell, Frans H. Claas, Sandy Feng, Howard M. Gebel, Annette M. Jackson, Roslyn B. Mannon, et al. 2018. "Sensitization in Transplantation: Assessment of Risk (STAR) 2017 Working Group Meeting Report." *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 18 (7): 1604–14. https://doi.org/10.1111/ajt.14752.

Tambur, Anat R., and Chris Wiebe. 2018. “HLA Diagnostics: Evaluating DSA Strength by Titration.” *Transplantation* 102 (1S Suppl 1): S23–30. https://doi.org/10.1097/TP.0000000000001817.

Viglietti, Denis, Alexandre Loupy, Olivier Aubert, Oriol Bestard, Jean-Paul Duong Van Huyen, Jean-Luc Taupin, Denis Glotz, et al. 2018. “Dynamic Prognostic Score to Predict Kidney Allograft Survival in Patients with Antibody-Mediated Rejection.” *Journal of the American Society of Nephrology: JASN* 29 (2): 606–19. https://doi.org/10.1681/ASN.2017070749.

Wiebe, C., I. W. Gibson, T. D. Blydt-Hansen, M. Karpinski, J. Ho, L. J. Storsley, A. Goldberg, P. E. Birk, D. N. Rush, and P. W. Nickerson. 2012. “Evolution and Clinical Pathologic Correlations of de Novo Donor-Specific HLA Antibody Post Kidney Transplant.” *American Journal of Transplantation: Official Journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 12 (5): 1157–67. https://doi.org/10.1111/j.1600-6143.2012.04013.x.

Willicombe, Michelle, Candice Roufosse, Paul Brookes, Jack W. Galliford, Adam G. McLean, Anthony Dorling, Anthony N. Warrens, Terry H. Cook, Tom D. Cairns, and David Taube. 2011. “Antibody-Mediated Rejection after Alemtuzumab Induction: Incidence, Risk Factors, and Predictors of Poor Outcome.” *Transplantation* 92 (2): 176–82. https://doi.org/10.1097/TP.0b013e318222c9c6.

Working Group of the International IgA Nephropathy Network and the Renal Pathology Society, Ian S. D. Roberts, H. Terence Cook, Stéphan Troyanov, Charles E. Alpers, Alessandro Amore, Jonathan Barratt, et al. 2009. “The Oxford Classification of IgA Nephropathy: Pathology Definitions, Correlations, and Reproducibility.” *Kidney International* 76 (5): 546–56. https://doi.org/10.1038/ki.2009.168.

Wu, K. Y., K. Budde, D. Schmidt, H. H. Neumayer, and B. Rudolph. 2014. “Acute Cellular Rejection with Isolated V-Lesions Is Not Associated with More Favorable Outcomes than Vascular Rejection with More Tubulointerstitial Inflammations.” *Clinical Transplantation* 28 (4): 410–18. https://doi.org/10.1111/ctr.12333.

Yalamati, Padma, Madhu Latha Karra, and Aparna V. Bhongir. 2016. “Comparison of Urinary Total Proteins by Four Different Methods.” *Indian Journal of Clinical Biochemistry* 31 (4): 463–67. https://doi.org/10.1007/s12291-016-0551-3.

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.

Appendix 2. BMJ Publication (Loupy et al. 2019), begins on following page:

Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study

Alexandre Loupy,[1,2] Olivier Aubert,[1,2] Babak J Orandi,[3] Maarten Naesens,[4] Yassine Bouatou,[1] Marc Raynaud,[1] Gillian Divard,[1] Annette M Jackson,[5] Denis Viglietti,[1,6] Magali Giral,[7] Nassim Kamar,[8] Olivier Thaunat,[9] Emmanuel Morelon,[9] Michel Delahousse,[10] Dirk Kuypers,[4] Alexandre Hertig,[11] Eric Rondeau,[11] Elodie Bailly,[11] Farsad Eskandary,[12] Georg Böhmig,[12] Gaurav Gupta,[13] Denis Glotz,[1,6] Christophe Legendre,[1,2] Robert A Montgomery,[14] Mark D Stegall,[15] Jean-Philippe Empana,[1,16] Xavier Jouven,[1] Dorry L Segev,[17] Carmen Lefaucheur[1,6]

For numbered affiliation see end of the article.

Correspondence to: A Loupy alexandre.loupy@inserm.fr (or @AlexandreLoupy on Twitter ORCID 0000-0003-3388-7747)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2019;366:l4923 http://dx.doi.org/10.1136/bmj.l4923

Accepted: 15 July 2019

ABSTRACT

OBJECTIVE

To develop and validate an integrative system to predict long term kidney allograft failure.

DESIGN

International cohort study.

SETTING

Three cohorts including kidney transplant recipients from 10 academic medical centres from Europe and the United States.

PARTICIPANTS

Derivation cohort: 4000 consecutive kidney recipients prospectively recruited in four French centres between 2005 and 2014. Validation cohorts: 2129 kidney recipients from three centres in Europe and 1428 from three centres in North America, recruited between 2002 and 2014. Additional validation in three randomised controlled trials (NCT01079143, EudraCT 2007-003213-13, and NCT01873157).

MAIN OUTCOME MEASURE

Allograft failure (return to dialysis or pre-emptive retransplantation). 32 candidate prognostic factors for kidney allograft survival were assessed.

RESULTS

Among the 7557 kidney transplant recipients included, 1067 (14.1%) allografts failed after a median post-transplant follow-up time of 7.12 (interquartile range 3.51-8.77) years. In the derivation cohort, eight functional, histological, and immunological prognostic factors were independently associated with allograft failure and were then combined into a risk prediction score (iBox). This score showed accurate calibration and discrimination (C index 0.81, 95% confidence interval 0.79 to 0.83). The performance of the iBox was also confirmed in the validation cohorts from Europe (C index 0.81, 0.78 to 0.84) and the US (0.80, 0.76 to 0.84). The iBox system showed accuracy when assessed at different times of evaluation post-transplant, was validated in different clinical scenarios including type of immunosuppressive regimen used and response to rejection therapy, and outperformed previous risk prediction scores as well as a risk score based solely on functional parameters including estimated glomerular filtration rate and proteinuria. Finally, the accuracy of the iBox risk score in predicting long term allograft loss was confirmed in the three randomised controlled trials.

CONCLUSION

An integrative, accurate, and readily implementable risk prediction score for kidney allograft failure has been developed, which shows generalisability across centres worldwide and common clinical scenarios. The iBox risk prediction score may help to guide monitoring of patients and further improve the design and development of a valid and early surrogate endpoint for clinical trials.

TRIAL REGISTRATION

Clinicaltrials.gov NCT03474003.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The transplant field lacks robust studies specifically designed for prediction of risk of long term allograft failure

Existing studies do not integrate a large spectrum of prognostic factors and validate scoring systems in multiple large cohorts worldwide with different transplant allocation systems

This represents a serious limitation for further improving patient care and drug development

WHAT THIS STUDY ADDS

This is the first international study of risk prediction in kidney transplant recipients, developed and validated across several large independent populations and in randomised controlled clinical trials

The iBox score represents a novel integration of demographic, functional, histological, and immunological factors that can be implemented in routine clinical practice

It has potential to upgrade the shared decision making process for transplant patients and represents a valid and early surrogate endpoint for clinical trials and drug development in transplantation

Introduction

End stage renal disease affects an estimated 7.4 million people worldwide.[1 2] According to data from the World Health Organization, more than 1 500 000 people live with transplanted kidneys, and 80 000 new kidneys are transplanted each year.[3] Despite the considerable advances in short term outcomes, kidney transplant recipients continue to experience late allograft failure, and little improvement has been made over the past 15 years.[4 5] Although the failure of a kidney allograft represents an important cause of end stage renal disease, robust and widely validated prognostication systems for the risk of allograft failure in individual patients are lacking.[6] Accurately predicting individual

patients' risk of allograft loss would help to stratify patients into clinically meaningful risk groups, which may help to guide monitoring of patients. Moreover, regulatory agencies and medical societies have highlighted the need for an early and robust surrogate endpoint in transplantation that adequately predicts long term allograft failure.[7] An enhanced ability to predict allograft outcomes would not only inform daily clinical care, counselling of patients, and therapeutic decisions but also facilitate the performance of clinical trials, which generally lack statistical power because of the low event rates during the first year after transplantation.[8]

Taken individually, parameters such as estimated glomerular filtration rate (eGFR),[9 10] proteinuria,[11] histology,[12] or human leukocyte antigen (HLA) antibody profiles,[13] fail to provide sufficient predictive accuracy. Previous efforts at developing prognostic systems in nephrology based on various combinations of parameters have been hampered by small sample sizes, the absence of proper validation, limited phenotypic details from registries, the absence of systematic immune response monitoring, and the failure to include key prognostic factors that affect allograft outcome (for example, donor derived factors, polyoma virus associated nephropathy, disease recurrence).[14-16] Finally, no scoring system has been evaluated in large cohorts from different countries with different transplant practices, allocation systems, and practice patterns, thereby limiting their exportability, which is an important consideration for health authorities to accept a scoring system as a surrogate endpoint.[17]

The objectives of this study (NCT03474003) were to develop a practical risk stratification score in a multicentre, prospective cohort of kidney transplant recipients that could be used to identify patients at high risk of future allograft loss; to validate the score on a large scale in geographically distinct independent cohorts with different allocation policies and types of transplant management; and to test the performance of the risk score for predicting graft failure in randomised controlled trials covering distinct clinical scenarios of transplant.

Methods

Study design and participants

Derivation cohort.

The derivation cohort consisted of 4000 consecutive patients over 18 years of age who were prospectively enrolled at the time of transplantation of a kidney from a living or deceased donor at Necker Hospital (n=1473), Saint-Louis Hospital (n=928), Foch Hospital (n=714), and Toulouse Hospital (n=885) in France between 1 January 2005 and 1 January 2014. We excluded patients with grafts that never functioned (primary non-functioning grafts; n=116). The clinical data were collected from each centre and entered into the Paris Transplant Group database (French data protection authority (CNIL) registration number: 363505). All data were anonymised and prospectively entered at the time of transplantation, at the time of post-transplant allograft biopsies, and at each transplant anniversary by using a standardised protocol to ensure harmonisation across study centres. We submitted data from the derivation cohort for an annual audit to ensure data quality (see the methods section and the study protocol in the supplementary material for detailed data collection procedures). We retrieved data from the database in March 2018. All patients provided written informed consent at the time of transplantation.

Validation cohorts.

The external validation cohorts comprised 3557 recipients of kidney transplants from a living or a deceased donor who were over 18 years of age and represented all patients eligible for post-transplant risk evaluation (that is, undergoing allograft biopsy as part of the standard of care of each centre with adequate biopsy according to the Banff criteria) from six centres: 2129 recipients recruited in Europe and 1428 recipients recruited in the US between 2002 and 2014. The European centres were Hôpital Hôtel Dieu, Nantes, France (n=632); Hospices Civils, Lyon, France (n=608); and the University Hospitals, Leuven, Belgium (n=889). The US centres were the Johns Hopkins Medical Institute, Baltimore, MD (n=580); the Mayo Clinic, Rochester, MN (n=556); and the Virginia Commonwealth University School of Medicine, Richmond, VA (n=292). Datasets from the validation centres were prospectively collected as part of routine clinical practice, entered in the centres' databases in compliance with local and national regulatory requirements, and sent anonymised to the Paris Transplant Group.

In France, the transplantation allocation system followed the rules of the French National Agency for Organ Procurement (Agence de la Biomédecine). The European centre outside France (Leuven) followed the rules of the Eurotransplant allocation system (https://www.eurotransplant.org), and the US centres (Johns Hopkins Hospital, Mayo Clinic, and Virginia) followed the rules of the US Organ Procurement and Transplantation System (https://unos.org/).

Additional external validation cohort.

Additional external validation was conducted in kidney transplant recipients previously recruited in three registered and published phase II and III clinical trials: a randomised, open label, multicentre trial that compared a cyclosporine based immunosuppressive regimen with an everolimus based regimen in kidney recipients (Certitem, NCT01079143); a randomised, multicentre, double blind, placebo controlled trial that investigated the efficacy of rituximab in kidney recipients with acute antibody mediated rejection (Rituxerah, EudraCT 2007-003213-13); and a randomised, double blind, placebo controlled, single centre trial that investigated the efficacy of bortezomib in kidney recipients with late antibody mediated rejection (Borteject, NCT01873157).[18-20] The details of the clinical trials including the population characteristics,

doi: 10.1136/bmj.l4923 | BMJ 2019;366:l4923 | the bmj

study design, inclusion criteria, and interventions are provided in supplementary table A.

Candidate predictors

Post-transplant risk evaluation times

Risk evaluation after transplantation was conducted at the time of allograft biopsy performed for clinical indication or as per protocol, which was performed after transplantation according to the centres' practices. In patients with multiple biopsies, risk evaluation used the date of the first biopsy. The distribution of post-transplant risk evaluation times is provided in supplementary figure A.

Risk evaluation after transplant comprised demographic characteristics (including recipients' comorbidities, age, sex, and transplant characteristics), biological parameters (including kidney allograft function, proteinuria, and circulating anti-HLA antibody specificities and concentrations), and allograft pathology data (including elementary lesion scores and diagnoses). All these factors are commonly and routinely collected in kidney transplant centres worldwide. See supplementary methods for the list of all prognostic determinants assessed from the derivation cohort.

Measurements performed at time of risk evaluation

Kidney allograft function was assessed by the glomerular filtration rate estimated by the Modification of Diet in Renal Disease Study equation (eGFR) and proteinuria level by using the protein/creatinine ratio in the derivation and validation cohorts. Circulating donor specific antibodies against HLA-A, HLA-B, HLA-Cw, HLA-DR, HLA-DQ, and HLA-DP were assessed using single antigen flow bead assays in the derivation cohort (see supplementary methods) and according to local centres' practice in the validation cohorts. Kidney allograft pathology data, including elementary lesion scores and diagnoses, were recorded according to the Banff classification in the derivation and validation cohorts (see supplementary methods). All the measurements (eGFR, proteinuria, histopathology, and circulating anti-HLA DSA) were performed on the day of risk evaluation.

Outcome

The outcome of interest was allograft loss defined as a patient's definitive return to dialysis or pre-emptive kidney retransplantation. This outcome was prospectively assessed in the derivation and validation cohorts at each transplant anniversary up to 31 March 2018.

Missing data

We excluded 59 (0.01%) patients in the derivation cohort from the final model owing to at least one data point being missing. We excluded 158 (7.4%) patients in the European validation cohort and 71 (5.0%) in the North American validation cohort from the final model owing to at least one data point being missing.

Statistical analysis

We followed the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement (supplementary methods) for reporting the development and validation of the multivariable prediction model.[21] We describe continuous variables by using means and standard deviations or medians and interquartile ranges. We compared means and proportions between groups by using Student's *t* test, analysis of variance (Mann-Whitney test for mean fluorescence intensity), or the χ^2 test (or Fisher's exact test if appropriate). We used the Kaplan-Meier method to estimate graft survival. The duration of follow-up was from the patient's risk evaluation (starting point) to the date of kidney allograft loss or the end of the follow-up (31 March 2018). For patients who died with a functioning allograft, allograft survival was censored at the time of death as a surviving or functional allograft.[22]

In the derivation cohort, we used univariable Cox regression analyses to assess the associations between allograft failure and clinical, histological, functional, and immunological factors measured at the patient's risk evaluation (see above). We used the log graphic method to test hazard proportional assumptions. The factors identified in these analyses were thereafter included in a final multivariable model.

We confirmed the internal validity of the final model by using a bootstrap procedure, which involved generating 1000 datasets derived from resampling the original dataset and permitting the calculation of optimism corrected performance estimates.[23] We tested the centre effect in stratified analyses. We investigated potential non-linear relations between continuous predictors and graft loss by using fractional polynomial methods (see supplementary methods).

We assessed the accuracy of the prediction model on the basis of its discrimination ability and calibration performance. We evaluated the discrimination ability (the ability to separate patients with different prognoses) of the final model by using Harrell's concordance index (C index) (see supplementary methods).[24] We assessed calibration (the ability to provide unbiased survival predictions in groups of similar patients) on the basis of a visual examination of the calibration plots by using the rms package in R. We used the SurvIDINRI package in R to calculate net reclassification improvement for censored survival data.[25 26] We then evaluated the external validity of the final model in the external validation cohorts, including discrimination tests and model calibration as mentioned above.

We calculated a risk prediction score (integrative box risk prediction score—iBox) for each patient according to the β regression coefficients estimated from the final multivariable Cox model. Allograft survival probabilities are given at three, five, and seven years after iBox risk evaluation. The seven year post-transplant iBox risk assessment was guided by the median follow-up after iBox risk assessment of 7.65 (interquartile range 5.39-8.21) years.

We used R version 3.2.1 foe all analyses and considered P values below 0.05 to be significant; all tests were two tailed. Details of the interpretation of important statistical concepts are given in the supplementary methods.

Patient and public involvement

The iBox initiative, including study design, study results, and potential for patient care, was presented and discussed among the two main French patients' associations, involving patients, nurses, and healthcare professionals.

Results

Characteristics of derivation and validation cohorts

The derivation cohort (n=4000) and the two validation cohorts (n=3557) comprised a total of 7557 participants with 1067 (14.1%) allograft failures after a median post-transplant follow-up time of 7.12 (interquartile range 3.51-8.77) years. The characteristics of the derivation and validation cohorts (overall, European, and US validation cohorts), as well as the transplant procedures, policies and allocation systems, are detailed in table 1 and supplementary tables B-D. The distribution of the time of the post-transplant risk evaluation is provided in supplementary figure A. The median time from kidney transplantation to post-transplant risk evaluation was 0.98 (0.27-1.07) years in the derivation cohort and 0.99 (0.18-1.04) years in the validation cohort. The median follow-up after transplantation was 7.65 (5.39-8.21) years in the derivation cohort. The cumulative numbers of graft losses in the development cohort were 332 at three years, 449 at five years, and 549 at seven years.

Prediction of kidney allograft failure in derivation cohort

We first investigated the prognostic factors measured at the time of post-transplant risk evaluation that were associated with long term kidney allograft failure in a univariable analysis. These factors included recipient's demographics, characteristics of transplant, allograft functional parameters, immunological parameters, and allograft histopathology (table 2). In the multivariable analysis, the following independent predictors of long term allograft failure were identified: time of post-transplant risk evaluation (P=0.005); allograft functional parameters, including eGFR (P<0.001) and proteinuria (logarithmic transformation, P<0.001); allograft histological parameters, including interstitial fibrosis and tubular atrophy (P=0.031), microcirculation inflammation defined by glomerulitis and peritubular capillaritis (P=0.001), interstitial inflammation and tubulitis (P=0.014), and transplant glomerulopathy (P=0.004); and recipient's immunological profile as defined by the presence and concentration of the immunodominant circulating anti-HLA donor specific antibodies (P<0.001) (table 3). We used a Cox model stratified by centre to test the effect of centre. We obtained stratified estimates (with equal coefficients across centres but with a baseline hazard unique to each centre). We confirmed that the eight prognostic parameters identified in the primary analysis remained independently associated with allograft survival (supplementary table E).

We calculated the prognostic score, named iBox, for each patient according to the β regression coefficients estimated from the final multivariable Cox model. On the basis of this score, we built a ready to use online interface for the clinician to provide allograft survival estimates for individual patients (http://www.paristransplantgroup.org). We are also providing, in supplementary figure B, examples of clinical use of iBox risk prediction scoring in daily practice.

Prediction model performance in internal and external validation cohorts

We first internally validated the final multivariable model via a bootstrapping procedure with 1000 samples from the original dataset of the derivation cohort (supplementary methods). Using this approach, we confirmed the robustness of the final multivariable model: the internal validity of the final model using a bootstrap procedure, which involved generating 1000 datasets derived from resampling the original dataset, thus permitting the calculation of optimism corrected performance estimates. Models were fitted for each of the 1000 samples by using backwards elimination. The eight independent predictors identified in the final multivariable Cox model were replicated in more than 85% of the 1000 estimated models. We also confirmed the discrimination ability of the model at three, five, and seven years (C index 0.835 (95% confidence interval 0.813 to 0.856), 0.819 (0.799 to 0.839), and 0.808 (0.790 to 0.827), respectively) by internally validating it using bootstrap resampling with optimism corrected C index 0.831 (0.813 to 0.854), 0.816 (0.799 to 0.837), and 0.806 (0.790 to 0.827) at three, five, and seven years, respectively.

We then used several independent validation cohorts and confirmed the transportability of the iBox risk score in these geographically distinct cohorts. The cumulative number of allograft losses were 72 (3.4%), 155 (7.3%), and 206 (9.7%) in the European validation cohort and 73 (5.1%), 108 (7.6%), and 148 (10.4%) in the US validation cohort at three, five, and seven years after iBox risk evaluation.

Overall, we showed good discrimination performance in the external validation cohorts with a C statistic of 0.81 (95% bootstrap percentile confidence interval 0.78 to 0.84) in Europe and 0.80 (0.76 to 0.84) in the US. Visual inspection of the calibration plots showed good agreement between the iBox risk score predicted probabilities of allograft survival at three, five, and seven years after risk evaluation and actual kidney allograft survival (fig 1).

Effect of therapeutic interventions on iBox risk score

We applied the iBox risk score to patients with therapeutic interventions, including 844 kidney transplant recipients from the derivation cohort who

Table 1 | Patients' characteristics by cohort. Values are numbers (percentages) unless stated otherwise

	Derivation cohort (n=4000)	European validation cohort (n=2129)	US validation cohort (n=1428)	P value*
Recipient demographics				
Mean (SD) age, years	49.83 (13.7)	50.58 (13.66)	50.42 (14.17) (n=1420)	0.09
Male sex	2450 (61.3)	1333 (62.6)	830 (58.1)	0.02
Cause of end stage renal disease:				
Glomerulonephritis	1086 (27.2)	584 (27.4)	365 (25.6)	
Diabetes	438 (11.0)	316 (14.8)	271 (19.08)	<0.001
Vascular	296 (7.4)	139 (6.5)	249 (17.4)	
Other	2180 (54.5)	1090 (51.2)	543 (38.0)	
Transplant characteristics				
Mean (SD) donor age, years	51.68 (16.33)	48.24 (15.79) (n=2122)	41.01 (14.75) (n=1420)	<0.001
Male donor	2151 (53.8)	1225/2124 (57.7)	694/1420 (48.9)	<0.001
Donor with hypertension	1005/3903 (25.7)	450/1876 (24.0)	189/1287 (14.7)	<0.001
Donor with diabetes mellitus	231/3861 (6.0)	47/1713 (2.7)	47/1276 (3.7)	<0.001
Donor with serum creatinine >1.5 mg/dL	422/3962 (10.7)	193/1936 (10.0)	284/1075 (26.4)	<0.001
Donor type:				
Deceased donor	3327 (83.2)	1974 (92.7)	620 (43.4)	<0.001
Death from cerebrovascular disease	1864/3327 (56.0)	993/1974 (50.3)	194/618 (31.4)	<0.001
Expanded criteria donor	1409/3995 (35.3)	628/2010 (31.2)	72/1425 (5.1)	<0.001
Prior kidney transplant	605 (15.1)	322 (15.1)	235/1408 (16.7)	0.34
Mean (SD) cold ischaemia time, hours	16.20 (8.99) (n=3976)	15.50 (7.30) (n=2093)	9.51 (11.81) (n=1212)	<0.001
Delayed graft function†	1046/3897 (26.8)	476/2127 (22.40)	158/1424 (11.1)	<0.001
Mean (SD) No with HLA-A/B/DR mismatch	3.817 (1.36)	3.15 (1.39) (n=2083)	3.54 (1.79) (n=1427)	<0.001

HLA=human leucocyte antigen.
*Based on comparison of all cohorts.
†Defined as use of dialysis in first postoperative week.

received standard of care treatment for antibody mediated rejection, standard of care treatment for T cell mediated rejection, and calcineurin inhibitor weaning for calcineurin inhibitor toxicity with belatacept (characteristics, protocols, and treatment interventions detailed in supplementary table F). Overall, we found that the therapeutic interventions were associated with significant changes in the iBox risk scores (supplementary figure C). The iBox prediction capability after treatment was accurate in these three therapeutic scenarios (C index 0.81, 95% bootstrap percentile confidence interval 0.77 to 0.85). The calibration plot showed a good agreement between the iBox prediction model after therapeutic intervention and the actual observation of kidney allograft loss.

Performance of iBox risk prediction score in therapeutic randomised controlled clinical trials

We tested the performance of the iBox risk prediction score in three registered and published phase II and III clinical trials.[18-20] The details of the clinical trials including the population, intervention, clinical scenario, and follow-up times are presented in supplementary table A. We calculated the iBox risk prediction scores of all patients included in the trials and compared them with the actual allograft failures. The iBox risk prediction score applied in the three trials showed accurate discrimination overall (C index 0.87, 0.82 to 0.92). The calibration plot showed a good agreement between the risk prediction score based on predicted allograft loss and the actual observations of kidney allograft loss.

Sensitivity analyses

We did various sensitivity analyses to test the robustness and generalisability of the iBox risk score in different clinical scenarios and subpopulations.

iBox integrative risk prediction score using allograft monitoring (eGFR/proteinuria) parameters

We showed that the iBox risk score using the full model was superior in terms of prediction capability to a simplified iBox model including eGFR, proteinuria, and circulating anti-HLA DSA (C index 0.79, 0.77 to 0.81; P<0.001). This was further demonstrated by a continuous net reclassification improvement of 0.228 for the full iBox model compared with the simplified iBox model (95% confidence interval 0.174 to 0.290; P<0.001). To account for potentially different medico-economic contexts limiting the availability of allograft biopsies, we are providing a simplified iBox score based on functional-immunological parameters. The calibration plot showed a good agreement between allograft loss predicted by the simplified iBox model and the actual observations of kidney allograft loss.

Added value of iBox risk prediction score compared with previously reported risk scores

We did a systematic review (supplementary table G) and compared the iBox risk prediction score with previously published risk scores assessing long term allograft outcomes. This showed that the iBox prediction score outperformed other risk scores (supplementary table G).

Table 2 | Factors assessed at time of post-transplant risk evaluation associated with kidney allograft failure in derivation cohort: univariable analysis

	No of patients	No of events*	Hazard ratio (95% CI)	P value
Recipient characteristics				
Age (per 1 year increment)	4000	549	1.00 (1.00 to 1.01)	0.46
Sex:				
Female	1550	214	1	
Male	2450	335	1.00 (0.85 to 1.19)	0.97
Transplant characteristics				
Donor age (per 1 year increment)	4000	549	1.02 (1.01 to 1.02)	<0.001
Donor sex:				
Female	1849	254	1	
Male	2151	295	0.98 (0.83 to 1.16)	0.83
Donor type:				
Living	673	51	1	
Deceased	3327	498	2.06 (1.54 to 2.74)	<0.001
Donor after cardiac death†:				
No	3234	489	1	
Yes	93	9	1.51 (0.78 to 2.92)	0.22
Donor hypertension:				
No	2898	340	1	
Yes	1005	195	1.84 (1.54 to 2.20)	<0.001
Donor diabetes mellitus:				
No	3630	491	1	
Yes	231	31	1.392 (1.01 to 1.93)	0.05
Creatinine concentration:				
<1.5 mg/dL	3540	467	1	
≥1.5 mg/dL	422	75	1.43 (1.12 to 1.82)	0.004
Expanded criteria donor:				
No	2586	285	1	
Yes	1409	263	1.90 (1.60 to 2.24)	<0.001
Previous kidney transplant:				
No	3395	421	1	
Yes	605	128	1.86 (1.53 to 2.27)	<0.001
Cold ischaemia time:				
<12 hours	1120	106	1	
12-24 hours	2099	319	1.61 (1.30 to 2.01)	
≥24 hours	757	121	1.73 (1.33 to 2.25)	<0.001
Thymoglobulin induction immunosuppression:				
No	1643	109	1	
Yes	2104	316	1.25 (1.05 to 1.49)	0.012
No of HLA-A/B/DR mismatches	4000	549	1.03 (0.97 to 1.10)	0.29
Delayed graft function†:				
No	2851	362	1	
Yes	104	246	1.94 (1.63 to 2.30)	<0.001
Pre-existing anti-HLA donor-specific antibody:				
No	3278	425	1	
Yes	722	124	1.51 (1.23 to 1.84)	0.001
Time of risk evaluation				
Time from transplant to evaluation (per 1 year increment)	3996	549	1.26 (1.21 to 1.33)	<0.001
Functional parameters				
eGFR (mL/min/1.73 m^2)	4000	549	0.94 (0.94 to 0.95)	<0.001
Proteinuria at 1 year (log transformation)	4000	549	1.99 (1.86 to 2.13)	<0.001
Structural-histopathology parameters				
Interstitial fibrosis/tubular atrophy:				
0-1	3099	331	1	
2	555	116	2.15 (1.74 to 2.66)	
3	321	95	3.36 (2.67 to 4.22)	<0.001
Arteriosclerosis:				
0	1365	137	1	
≥1	2446	386	1.62 (1.33 to 1.97)	<0.001
Hyalinosis:				
0	1567	149	1	
≥1	2360	381	1.74 (1.44 to 2.10)	<0.001
Interstitial inflammation and tubulitis:				
0-2	3610	546	1	
≥3	390	93	1.97 (1.58 to 2.46)	<0.001
Transplant glomerulopathy:				
0	3702	449	1	
≥1	260	94	3.70 (2.96 to 4.62)	<0.001

Table 2 | Continued

	No of patients	No of events*	Hazard ratio (95% CI)	P value
Endarteritis:				
0	3794	506	1	
≥1	96	27	2.26 (1.54 to 3.33)	<0.001
C4d graft deposition:				
No	3452	416	1	
Yes	548	133	2.45 (2.01 to 2.98)	<0.001
Microcirculation inflammation (g+ptc):				
0-2	3616	261	1	
3-4	308	92	3.07 (2.45 to 3.85)	
5-6	76	35	4.99 (3.53 to 7.04)	<0.001
Polyomavirus associated nephropathy:				
No	3902	518	1	
Yes	97	31	2.82 (1.96 to 4.05)	<0.001
Nephropathy recurrence:				
No	3868	510	1	
Yes	130	38	2.55 (1.84 to 3.55)	<0.001
Antibody mediated rejection:				
No	3398	368	1	
Yes	600	181	3.36 (2.81 to 4.02)	<0.001
T cell mediated rejection:				
No	3812	503	1	
Yes	187	46	1.96 (1.45 to 2.66)	<0.001
Immunological parameters				
Anti-HLA donor specific antibody mean fluorescence intensity				
<500	3312	394	1	
≥500-3000	483	82	1.66 (1.31 to 2.11)	
≥3000-6000	82	24	3.11 (2.06 to 4.70)	
≥6000	123	49	4.56 (3.38 to 6.14)	<0.001

C4d=C4d stain; eGFR=estimated glomerular filtration rate; g=glomerulitis score; HLA=human leukocyte antigen; ptc=peritubular capillaratis score.
*Number of events at 7 years after iBox risk evaluation.
†Among deceased donors.

Prediction model performance using histological diagnoses instead of Banff international classification histological lesion grading

When we included histological diagnoses in the multivariable model instead of histological lesions graded according to the international Banff classification, antibody mediated rejection (P<0.001), T cell mediated rejection (P=0.04), primary nephropathy recurrence (P=0.003), and BK virus nephropathy (P=0.05) showed significant and independent associations with allograft failure. In this model, the set of non-histological predictors of allograft failure identified in the primary analyses remained unchanged (hazard ratios are shown for each parameter in supplementary table H). The discrimination ability of the histological diagnosis based model showed a C index of 0.81 (0.79 to 0.83).

iBox performance when applied at time of clinically indicated biopsies versus protocol biopsies

We tested and confirmed the performance of the iBox risk prediction score when risk evaluation started at the time of clinically indicated allograft biopsies performed at any time after transplantation (n=1598; 40%), as well as at the time of one year protocol biopsies (n=2402; 60%) (table 4). Similarly, the iBox risk score showed accurate discrimination ability for long term allograft loss when risk evaluation started before one year post-transplant or after one year post-transplant (mean post-transplantation time of 0.89 (SD 0.23) years and 2.31 (1.66) years, respectively; table 4).

iBox risk score performance versus risk score based on parameters assessed at time of transplantation

When we tested the parameters assessed at time of transplantation (recipient's age, recipient's sex, donor's age, donor's sex, deceased donor, donor's cause of death, donor's diabetes, donor's hypertension, expanded criteria donor, previous kidney transplant, HLA mismatches, and anti-HLA donor specific antibody), none of them remained independently associated with allograft survival after adjustment for post-transplant parameters assessed at the time of iBox risk evaluation. Similarly, when we added day 0 parameters to the multivariable model including risk factors evaluated post-transplantation, we saw no improvement in its discrimination ability. Lastly, when we ran the Cox model with these parameters assessed at the time of transplantation, the C index was 0.62 (0.593 to 0.643).

iBox assessed in other clinical scenarios and subpopulations

Finally, we confirmed the performance of the iBox risk prediction score when applied in different subpopulations and clinical scenarios including living and deceased donors, according to recipient's ethnicity, in highly sensitised (high immunological risk) and non-highly sensitised (low immunological

Table 3 | Independent determinants of kidney allograft loss assessed at time of post-transplant risk evaluation in derivation cohort: multivariable analysis

Factor	No of patients	No of events*	Hazard ratio (95% CI)	P value
Time from transplant to evaluation (years)	3941	538	1.08 (1.02 to 1.14)	0.005
eGFR (mL/min/1.73 m^2)	3941	538	0.96 (0.95 to 0.96)	<0.001
Proteinuria (log)	3941	538	1.51 (1.40 to 1.63)	<0.001
Interstitial fibrosis/tubular atrophy:				
0/1	3074	330	1	
2	550	115	1.14 (0.92 to 1.42)	
3	317	93	1.39 (1.08 to 1.77)	0.03
Microcirculation inflammation (g+ptc):				
0-2	3568	414	1	
3-4	299	90	1.45 (1.12 to 1.88)	
5-6	74	34	1.83 (1.24 to 2.71)	0.001
Interstitial inflammation and tubulitis (i+t):				
0-2	3559	447	1	
≥3	382	91	1.34 (1.06 to 1.68)	0.01
Transplant glomerulopathy (cg)				
0	3684	445	1	
≥1	257	93	1.47 (1.13 to 1.90)	0.004
Anti-HLA donor specific antibody mean fluorescence intensity				
<500				
≥500-3000	477	80	1.25 (0.97 to 1.61)	
≥3000-6000	80	23	1.72 (1.13 to 2.66)	
≥6000	119	48	2.05 (1.47 to 2.86)	0.001

cg=transplant glomerulopathy score; eGFR=estimated glomerular filtration rate; g=glomerulitis score; HLA=human leukocyte antigen; i=interstitial inflammation score; ptc=peritubular capillaratis score; t=tubulitis score.
*Number of events at 7 years after iBox risk evaluation.

risk) recipients, and in patients receiving induction by anti-interleukin-2 receptor or anti-thymocyte globulin (table 4). When parameters assessed at the time of transplant (such as HLA mismatches), recipient blood pressure at the time of risk assessment (log scale), and calcineurin inhibitor through blood concentration at the time of risk assessment were forced in the risk prediction score, we saw no significant improvement in its prognostic performance (table 4).

Discussion

The iBox, a risk prediction score combining functional, histological, and immunological allograft parameters together with HLA antibody profiling, showed good performance in predicting the risk of long term kidney allograft failure. We confirmed the generalisability of the iBox risk prediction score by showing its external validity in six geographically distinct cohorts recruited in Europe and the US with distinct allocation systems, patients' characteristics, and management practices. The iBox risk prediction score also showed its accuracy when measured at different times after transplantation, which permits updating of the score on the basis of new events that patients might encounter in their long term course. We also showed that the iBox risk prediction score outperformed other available risk scores applied in kidney transplant patients. Lastly, we confirmed the predictive accuracy of the risk score in the data reported from three published randomised therapeutic trials covering different clinical scenarios encountered after transplantation, further enhancing its value as a potential surrogate endpoint in transplantation.[18-20]

Overall, the predictor variables used in the iBox risk prediction score are easily available after transplantation in most centres worldwide, making it feasible for implementation in routine clinical practice. The iBox risk prediction system assessed the risk at a given time point, but we have shown that it can be re-evaluated at different time points after transplantation, enabling clinicians to calculate a new risk that takes into account the updated values of eGFR, proteinuria, allograft scarring, allograft inflammation, damage, and presence and concentration of anti-HLA DSA. Therefore, we confirmed the iBox system's transportability for additional and updated evaluations in the patient's long term course. To account for different potential medico-economic contexts limiting the availability of allograft biopsies, we also provide an abbreviated iBox score based on clinical-functional-immunological parameters.

Comparison with other prognostic scores

Current prognostic scores implemented in clinical practice in transplant medicine mostly predict allograft survival at the time of transplantation; thus, their use is limited to allograft allocation because they do not inform post-transplant clinical decision making and monitoring of patients.[27] The few attempts to develop post-transplant prognostic scores have failed to provide useful tools for transplant clinicians. According to a systematic review without date restrictions for publications up to 28 September 2018, for allograft survival scoring systems among kidney transplant recipients (see supplementary table G), no study has developed and externally validated a post-transplant prognostic score usable at any time after transplantation that shows accuracy in clinical trials. The main limitations to achieving a robust and validated scoring system depend on multiple factors including the insufficient data quality of the previously

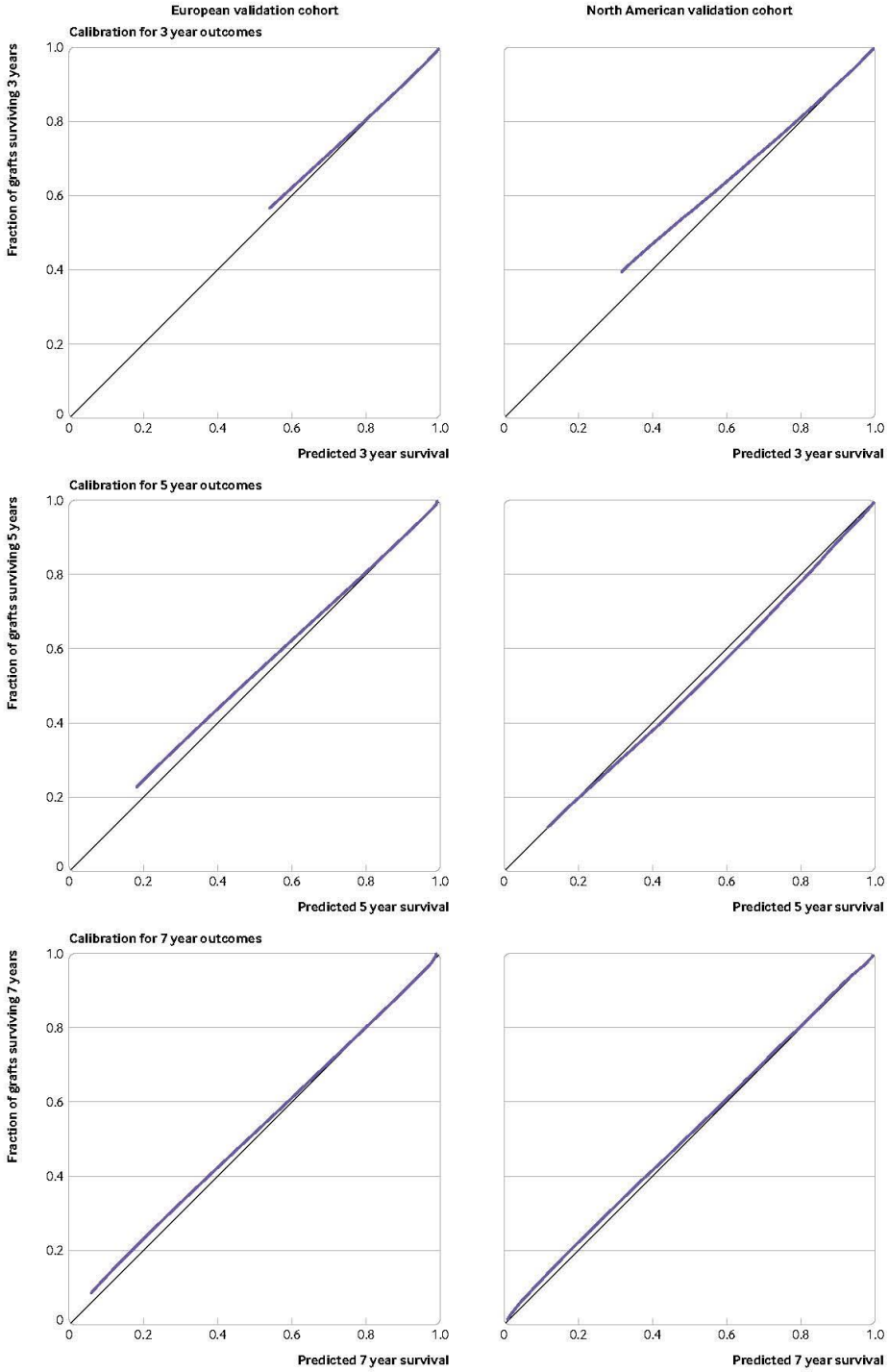


Fig 1 | Calibration plots at three, five, and seven years of iBox risk scores for validation cohorts: three year (A, B), five year (C, D), and seven year (E, F) predictions. Data are from European validation cohort (A, C, E) and US cohort (B, D, F). Vertical axis is observed proportion of grafts surviving at time of interest. Average predicted probability (predicted survival; x-axis) was plotted against Kaplan-Meier estimate (observed overall survival; y-axis). Black line represents perfectly calibrated model, and blue line represents optimism corrected iBox model

Table 4 | iBox risk prediction score performance when assessed in different clinical scenarios and subpopulations

Clinical scenarios and subpopulations	No of patients	No of events	Risk model performance: C statistic (95% bootstrap percentile CI)
Using functional and immunological parameters	3941	538	0.79 (0.77 to 0.81)
Using histological diagnoses* instead of Banff lesions grading	3997	548	0.81 (0.79 to 0.83)
In stable patients (protocol biopsy)	1160	85	0.81 (0.77 to 0.86)
In unstable patients (biopsy for cause)	2781	453	0.80 (0.78 to 0.82)
In first year after transplant	2300	291	0.78 (0.72 to 0.81)
After 1 year post-transplant	1641	247	0.84 (0.82 to 0.87)
In living donors	662	51	0.82 (0.75 to 0.88)
In deceased donors	3279	487	0.80 (0.78 to 0.82)
In highly sensitised recipients†	715	121	0.80 (0.76 to 0.84)
In non-highly sensitised recipients	3226	417	0.81 (0.79 to 0.83)
Adding transplant baseline characteristics‡	3735	573	0.81 (0.79 to 0.83)
In patients with anti-IL2 receptor induction	1621	206	0.79 (0.76 to 0.82)
In patients with anti-thymocyte globulin induction	2069	308	0.83 (0.80 to 0.85)
In African-American population§	371	62	0.80 (0.74 to 0.85)
In non-African-American population§	986	77	0.84 (0.80 to 0.89)
Adding recipient blood pressure profile post-transplant¶	3973	541	0.80 (0.78 to 0.82)
Adding CNI blood trough concentration at time of evaluation	3822	525	0.81 (0.78 to 0.83)

CNI=calcineurin inhibitor; IL=interleukin.
*Histological diagnoses defined by last update of Banff international classification: antibody mediated rejection, T cell mediated rejection, BK virus nephropathy, primary nephropathy recurrence.
†Highly sensitised patients defined by panel of reactive antibodies >90%.
‡Donor's age, donor's sex, donor's hypertension, donor's diabetes, recipient's age, recipient's sex, human leukocyte antigen (HLA) mismatches, retransplantation, and anti-HLA DSA at time of transplantation.
§Status was retrieved in US participating centres' databases (no ethnicity data allowed in French development cohort database according to the French law and regulation). African-Americans in US validation cohort represented 390 (27.3%) patients; Non-African-Americans in US validation cohort represented 1038 (72.7%) patients.
¶Blood pressure profile defined by systolic blood pressure measured at time of risk assessment on log scale.

studied cohorts and the fact that no registry or database system has been primarily designed to tackle the specific aspect of prognostication. An even more important aspect is external validation in different populations, which prompted us to conduct a large external validation in multiple centres worldwide. Despite some expected loss of discriminative performance, models are typically considered useful for clinical decision making when the C statistic is greater than 0.70 and strong when the C statistic exceeds 0.80, suggesting that the iBox risk prediction score could support decision making.[28] For prognostication systems in other fields such as oncology (for example, locally advanced pancreatic cancer and metastatic colonic cancer), the C index is typically closer to 0.60 or 0.70.[29] Taken together, these results confirm not only the robustness and validity of the iBox risk prediction score but also its generalisability to other transplant cohorts with different kidney allocation systems, donor and recipient profiles, and distinct patient management and healthcare environments.

Strengths of study

In this study, we have shown that the iBox risk prediction score outperformed the current gold standard (eGFR and proteinuria) for the monitoring of kidney recipients. In particular, compared with previous attempts at developing a prognostication system, we found that allograft histological lesions such as microcirculation inflammation, interstitial inflammation-tubulitis (reflecting active rejection process) and atrophy-fibrosis, and transplant glomerulopathy (reflecting chronic allograft damage), in addition to measuring allograft functional parameters and recipient antibody profiles, improved the overall discrimination capacity of the model and that a multidimensional risk prediction score performs better than its individual components. This risk prediction score reflects the main patterns of allograft deterioration leading to failure, represented by alloimmune processes and allograft scarring.[30] Two other prognostic scores have attempted to combine several transplant diagnostic dimensions, including allograft function and pathology and alloantibodies; however, these scores were outperformed by the iBox risk prediction score.[16 31]

Importantly, our results and the parameters included in the final model reinforce the potential of the iBox to be implemented into contemporary clinical practice by using automated approaches within electronic medical record systems (an online electronic risk calculator is provided at http://www.paristransplantgroup.org, and examples are provided in supplementary figure B).

In addition, the combination of major drivers of allograft failure in the iBox risk prediction score allowed us to evaluate the early effect of clinical interventions on long term allograft outcomes. In this study, we tested and validated the iBox risk prediction score in the setting of therapeutic clinical trials covering different clinical scenarios and showed accurate performance overall. We found that the prediction of allograft failure assessed by the iBox score accurately fits with the actual graft failures observed in these trials at five years after risk evaluation. Importantly, the accuracy of the iBox risk prediction score was conserved regardless of the therapeutic intervention and popula-

doi: 10.1136/bmj.l4923 | BMJ 2019;366:l4923 | the bmj

tion in those trials, with accurate performance in the Certitem (NCT01079143) calcineurin inhibitor minimisation trial and in rejection treatment trials (EudraCT 2007-003213-13; NCT01873157).[18-20] This finding reinforced the potential of the iBox risk prediction score for defining a valid surrogate endpoint. In our study, a well validated, strong, and robust association existed between the surrogate endpoint and the true endpoint, and this association was consistent across different treatment settings. Finally, because the criteria for defining a surrogate endpoint also include the capacity of a surrogate to be modified by therapeutics, we tested the iBox across three prototypic therapeutic interventions and showed that the iBox score was significantly modified by these therapeutic interventions and showed good performance in this setting as well. Thus, the iBox risk prediction score fulfils all the Prentice criteria for a satisfactory surrogate endpoint.[17 32]

As a development perspective, implementation of patient reported experience data would probably be very relevant in future, so that quality of life predictions can complement those on graft survival, around indicators such as the experience of treatments, the relationship with the transplant doctor, adherence to the therapeutic strategy, engagement, participation in decisions, fatigue, anxiety, depression, and so on. This would imply that other sources of data can be mobilised, from collections made from the patients themselves.

Limitations of study

Regarding the limitations of this study, we acknowledge that statistical significance as a criterion to select variables may not be ideal as it may exclude confounding factors. However, the multiple external validations performed consistently confirm the robustness of our final model. Emerging predictors post-transplant might be also missing in our model. Despite the already high performance achieved by the iBox risk prediction score, future studies should evaluate the added value of new non-invasive biomarkers or genetic factors in addition to those currently reported regarding discriminative capability, generalizability, and overcoming the need for an invasive procedure (kidney allograft biopsy). Although intragraft gene measurements may improve diagnostic accuracy in T cell mediated rejection and antibody mediated rejection, their additive value for allograft survival compared with classical prognostic factors has not yet been demonstrated in large unselected populations.

Another limitation is that information on the adherence to drug treatment of individual patients was lacking in our dataset. Although non-adherence is inherently difficult to capture, especially at a population level,[30] the iBox score, because its mechanistically informed design could likely capture the consequences of non-adherence (development of de novo donor specific anti-HLA antibodies, allograft injury, scarring, inflammation, and diminished glomerular filtration rate).

Although the iBox risk prediction score was primarily generated using a large, prospective, unselected cohort, a prospective validation of the iBox in daily clinical practice remains desirable. Finally, despite the validation of the iBox risk prediction score in an interventional setting, future trials are needed to determine whether a strategy based on a systematic risk evaluation compared with an empirical approach might improve clinical management.

Conclusions

We have developed and validated a risk prediction score that accurately predicts allograft failure after kidney transplantation. We have shown its generalisability and transportability across centres in Europe and the US and its performance in therapeutic clinical trials. The risk prediction score provides an accurate but simple strategy that can be easily implemented to stratify patients into clinically meaningful risk groups and that can be time updated after transplant, which may help to guide monitoring of patients in everyday practice and upgrade the shared decision making process. Lastly, as the risk score fulfils the Prentice criteria, it may represent a valid surrogate endpoint that could open avenues for improving the design of clinical trials and development of drugs in transplantation.

AUTHOR AFFILIATIONS

[1]Université de Paris, INSERM, Paris Translational Research Centre for Organ Transplantation, Paris, France
[2]Kidney Transplant Department, Necker Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France
[3]Department of Surgery, University of California San Francisco School of Medicine, San Francisco, CA, USA
[4]Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium
[5]Department of Surgery, Duke University School of Medicine, Durham, NC, USA
[6]Kidney Transplant Department, Saint-Louis Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France
[7]Department of Nephrology, Centre Hospitalier Universitaire de Nantes, Nantes, France
[8]Université Paul Sabatier, INSERM, Department of Nephrology and Organ Transplantation, CHU Rangueil & Purpan, Toulouse, France
[9]Department of Transplantation, Nephrology and Clinical Immunology, Hospices Civils de Lyon, France
[10]Department of Transplantation, Nephrology and Clinical Immunology, Foch Hospital, Suresnes, France
[11]Kidney transplant department, Tenon Hospital, Assistance Publique - Hôpitaux de Paris, Paris, France
[12]Division of Nephrology and Dialysis, Department of Medicine III, General Hospital Vienna, Vienna, Austria
[13]Division of Nephrology, Department of Internal Medicine, Virginia Commonwealth University School of Medicine, Richmond, VA, USA
[14]New York University Langone Transplant Institute, New York, NY, USA
[15]William J. von Liebig Centre for Transplantation and Clinical Regeneration, Mayo Clinic, Rochester, MN, USA
[16]Cardiology and Heart Transplant department, Pompidou hospital, Assistance Publique - Hôpitaux de Paris, Paris, France
[17]Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA

We thank the participants who made this study possible, the advocacy groups that participated in evaluating the iBox, and the two patients' associations Renaloo and France REIN which gave feedback on the applications of the risk score system and its translation to patient care. We plan to share the results to the wider community and our participants.

Contributors: AL and OA designed the study, did the data analysis, and wrote the manuscript. BO, MN, AMJ, DV, CLegendre, MG, NK, OT, EM, MD, DK, AH, ER, EB, FE, GB, GD, MR, GG, DG, JPE, XJ, RAM, MDS, DLS, and CLefaucheur contributed to data acquisition and interpretation. AL, OA, BO, MN, YB, DV, DLS, DG, CLegendre, JPE, and XJ interpreted the data. BO, MN, and YB participated in data interpretation and critically reviewed the manuscript. All authors revised the manuscript for important intellectual content. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. AL, OA, BO, and MN contributed equally as first authors. XJ, DLS, and CLefaucheur contributed equally as last authors. AL is the guarantor.

Funding: INSERM–Action thématique incitative sur programme Avenir (ATIP-Avenir) provided financial support; OA received a grant from the Fondation Bettencourt Schueller; MN received grants from the Research Foundation, Flanders (FWO; IWT.150199), the Flanders Innovation and Entrepreneurship of the Flemish government (IWT.130758), and the Clinical Research Foundation of the University Hospitals Leuven.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: support as detailed above for the submitted work; AL holds shares in Cibiltech, a company that develops software; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Each patient from the Paris Transplant Group cohort provided written informed consent to be included in the Paris Transplant Group database. This database has been approved by the National French Commission for Bioinformatics, Data, and Patient Liberty: CNIL registration number: 363505, validated 3 April 1996. The institutional review boards of the Paris Transplant Group participating centres approved the study.

Data sharing: Technical appendix is available from the corresponding author at alexandreloupy@gmail.com.

Transparency: The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Mills KT, Xu Y, Zhang W, et al. A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010. *Kidney Int* 2015;88:950-7. doi:10.1038/ki.2015.230
2 Hill NR, Fatoba ST, Oke JL, et al. Global Prevalence of Chronic Kidney Disease - A Systematic Review and Meta-Analysis. *PLoS One* 2016;11:e0158765. doi:10.1371/journal.pone.0158765
3 Evans RW, Manninen DL, Garrison LPJr, et al. The quality of life of patients with end-stage renal disease. *N Engl J Med* 1985;312:553-9. doi:10.1056/NEJM198502283120905
4 Meier-Kriesche HU, Schold JD, Srinivas TR, Kaplan B. Lack of improvement in renal allograft survival despite a marked decrease in acute rejection rates over the most recent era. *Am J Transplant* 2004;4:378-83. doi:10.1111/j.1600-6143.2004.00332.x
5 Coemans M, Süsal C, Döhler B, et al. Analyses of the short- and long-term graft survival after kidney transplantation in Europe between 1986 and 2015. *Kidney Int* 2018;94:964-73. doi:10.1016/j.kint.2018.05.018
6 Perl J. Kidney transplant failure: failing kidneys, failing care?*Clin J Am Soc Nephrol* 2014;9:1153-5. doi:10.2215/CJN.04670514
7 Stegall MD, Morris RE, Alloway RR, Mannon RB. Developing New Immunosuppression for the Next Generation of Transplant Recipients: The Path Forward. *Am J Transplant* 2016;16:1094-101. doi:10.1111/ajt.13582
8 Vincenti F, Rostaing L, Grinyo J, et al. Belatacept and Long-Term Outcomes in Kidney Transplantation. *N Engl J Med* 2016;374:333-43. doi:10.1056/NEJMoa1506027
9 Kaplan B, Schold J, Meier-Kriesche HU. Poor predictive value of serum creatinine for renal allograft loss. *Am J Transplant* 2003;3:1560-5. doi:10.1046/j.1600-6135.2003.00275.x
10 He X, Moore J, Shabir S, et al. Comparison of the predictive performance of eGFR formulae for mortality and graft failure in renal transplant recipients. *Transplantation* 2009;87:384-92. doi:10.1097/TP.0b013e3181900 4a1
11 Naesens M, Lerut E, Emonds MP, et al. Proteinuria as a Noninvasive Marker for Renal Allograft Histology and Failure: An Observational Cohort Study. *J Am Soc Nephrol* 2016;27:281-92. doi:10.1681/ASN.2015010062
12 Yilmaz S, Tomlanovich S, Mathew T, et al. Protocol core needle biopsy and histologic Chronic Allograft Damage Index (CADI) as surrogate end point for long-term graft survival in multicenter studies. *J Am Soc Nephrol* 2003;14:773-9. doi:10.1097/01.ASN.0000054496.68498.13
13 Lefaucheur C, Loupy A, Hill GS, et al. Preexisting donor-specific HLA antibodies predict outcome in kidney transplantation. *J Am Soc Nephrol* 2010;21:1398-406. doi:10.1681/ASN.2009101065
14 Moore J, He X, Shabir S, et al. Development and evaluation of a composite risk score to predict kidney transplant failure. *Am J Kidney Dis* 2011;57:744-51. doi:10.1053/j.ajkd.2010.12.017
15 Shabir S, Halimi JM, Cherukuri A, et al. Predicting 5-year risk of kidney transplant failure: a prediction instrument using data available at 1 year posttransplantation. *Am J Kidney Dis* 2014;63:643-51. doi:10.1053/j.ajkd.2013.10.059
16 Gonzales MM, Bentall A, Kremers WK, Stegall MD, Borrows R. Predicting Individual Renal Allograft Outcomes Using Risk Models with 1-Year Surveillance Biopsy and Alloantibody Data. *J Am Soc Nephrol* 2016;27:3165-74. doi:10.1681/ASN.2015070811
17 Schold JD, Kaplan B. The elephant in the room: failings of current clinical endpoints in kidney transplantation. *Am J Transplant* 2010;10:1163-6. doi:10.1111/j.1600-6143.2010.03104.x
18 Eskandary F, Regele H, Baumann L, et al. A Randomized Trial of Bortezomib in Late Antibody-Mediated Kidney Transplant Rejection. *J Am Soc Nephrol* 2018;29:591-605. doi:10.1681/ASN.2017070818
19 Sautenet B, Blancho G, Büchler M, 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* 2016;100:391-9. doi:10.1097/TP.0000000000000958
20 Rostaing L, Hertig A, Albano L, et al, CERTITEM Study Group. Fibrosis progression according to epithelial-mesenchymal transition profile: a randomized trial of everolimus versus CsA. *Am J Transplant* 2015;15:1303-12. doi:10.1111/ajt.13132
21 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55-63. doi:10.7326/M14-0697
22 Lamb KE, Lodhi S, Meier-Kriesche HU. Long-term renal allograft survival in the United States: a critical reappraisal. *Am J Transplant* 2011;11:450-62. doi:10.1111/j.1600-6143.2010.03283.x
23 Efron B. Bootstrap Methods: Another Look at the Jackknife. *Ann Stat* 1979;7:1-26. doi:10.1214/aos/1176344552
24 Harrell FEJr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-87. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
25 Pencina MJ, D'Agostino RBSr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;30:11-21. doi:10.1002/sim.4085
26 Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat Med* 2013;32:2430-42. doi:10.1002/sim.5647
27 Rao PS, Schaubel DE, Guidinger MK, et al. A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index. *Transplantation* 2009;88:231-6. doi:10.1097/TP.0b013e3181ac620b
28 Hosmer DWLS. *Applied Logistic Regression*. 2nd ed. John Wiley & Sons, 2000. doi:10.1002/0471722146
29 Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Intern Med* 2015;175:1389-98. doi:10.1001/jamainternmed.2015.2829
30 Sellarés J, de Freitas DG, Mengel M, et al. Understanding the causes of kidney transplant failure: the dominant role of antibody-mediated rejection and nonadherence. *Am J Transplant* 2012;12:388-99. doi:10.1111/j.1600-6143.2011.03840.x
31 Prémaud A, Filloux M, Gatault P, et al. An adjustable predictive score of graft survival in kidney transplant patients and the levels of risk linked to de novo donor-specific anti-HLA antibodies. *PLoS One* 2017;12:e0180236. doi:10.1371/journal.pone.0180236
32 Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431-40. d doi:oi:10.1002/sim.4780080407

Supplementary materials

Appendix 3. TTC Data Summary Table

Datasets In-House: Number of Subjects		
Study/Dataset Name	**Sponsor**	**Subject Numbers**
Mayo Arizona	--	3,160
Mayo Rochester	--	1,640
TRANSFORM	Novartis	2,226
Elevate	Novartis	992
US-92	Novartis	613
Houston Methodist	--	1,676
Total		**10,307**
6-Month Data Outlook: Number of Subjects		
Study/Dataset Name	**Sponsor**	**Subject Number**
BENEFIT, BENEFIT-EXT	BMS	1,333
LCPTacro3001 and 3002	Veloxis	863
Northwestern data	--	TBD
CTOT08	--	372
Paris Transplant Centers	--	~4,800
The 1010, TRIMS	Sanofi Genzyme	~428
Helsinki Finland Center Study	--	~600-700
Total		**~8,500**
Total Data Expected		**~18,807**

Appendix 4. FDA-TTC CPIM Summary, begins on following page:

Memorandum

Date: 7/25/2019

Subject: Critical Path Innovation Meeting: Transplant Therapeutics Consortium

Date of meeting: 3/26/2019

Requestor: Critical Path Institute

Note: Discussions at Critical Path Innovation Meetings are informal. All opinions, recommendations, and proposals are unofficial and nonbinding on FDA and all other participants.

FDA Representatives

Center for Drug Evaluation and Research

Office of the Center Director
• Professional Affairs and Stakeholder Engagement

Office of Translational Sciences (OTS)
• OTS, Office of Pharmacology
• OTS, Office of Biostatistics

Office of New Drugs (OND)
• OND, Office of Antimicrobial Products (OAP), Division of Transplant and Ophthalmology Products
• OND, OAP, Division of Anti-Infective Products
• OND, Office of Drug Evaluation (ONDE) I, Division of Cardiovascular and Renal Products
• ODEII, Division of Pulmonary, Allergy, and Rheumatology Products
• OND, ONDE III

REQUESTER: Critical Path Institute, Transplant Therapeutic Consortium (TTC)

Inish O'Doherty – Executive Director, TTC
Stephen Karpen – Scientific Director, TTC
Nicole Spear – Project Manager, TTC
Klaus Romero – Director of Clinical Pharmacology and Quantitative Medicine, C-Path
Jagdeep Podichetty – Scientific Director, Quantitative Medicine, C-Path
Ken Newell, American Society of Transplantation and TTC Workgroup Co-chair
Ulf Meier-Kriesche, Veloxis and TTC Workgroup Co-chair
Alex Loupy, Paris Transplant Group and TTC member

1. BACKGROUND

Approximately 19,000 patients received kidney transplants in the United States each year. Nearly all patients who receive kidney transplants are prescribed immunosuppressive drugs (ISD) therapy; however, a significant number of patients experience allograft loss by 5 and 10 years after transplantation. There is a need for development of ISDs with better long-term outcomes. The Critical Path Institute's Transplant Therapeutic Consortium (TTC) requested the CPIM to discuss current needs to accelerate drug development for kidney transplant patients and to obtain FDA's feedback on the development of a clinical trial simulation tool to support the evaluation of immunosuppressive drugs in clinical trials.

2. DISCUSSION

Representatives of TTC discussed the challenges associated with the development of novel immunosuppressive drugs (ISD). It was noted that the standard of care, immunosuppressive therapy for transplant patients have not changed in twenty years and that first-year post-transplant success rates were high. Clinical trials for ISDs must aim to be non-inferior to the standard of care regimen or lengthy if they are to show superiority to the standard of care. There is a lack of understanding of the complex mechanisms that lead to graft loss and diversity of kidney transplant donors and recipients. In addition, there are no publicly available drug development tools to optimize clinical trial design for ISDs for better long-term graft survival.

Representatives of the TTC described the development of the clinical trial simulation (CTS) tool. The proposed context-of use for the CTS tool is to optimize phase II and III clinical trial design for evaluating therapeutic candidates for immunosuppression following kidney transplantation. The goal is to develop a mathematical representation of longitudinal changes derived from 1-year post-transplant characteristics. The Integrative Box (iBox) Scoring System presented as a component of the CTS tool is based on two sets of multivariate models. The first set of models describes the 1-year dynamics of disease progression measured by proteinuria, eGFR, and donor specific antibodies. These models are enhanced with the Banff lesion score to derive the total iBox score, an integrative scoring system that predicts kidney allograft loss. The iBox system is based on a large international study of kidney transplant recipients and takes into account important allograft loss risk factors including baseline donor and recipient characteristics, transplant characteristics, post-transplant injuries, treatment and anti-HLA donor specific antibody measurements. TTC representatives noted that the performance of the iBox score was validated in multiple cohorts and was able to successfully predict graft survival beyond 1-year post transplantation. As a result they concluded that the iBox score is correlated with treatment performance. The results from the first set of multivariate models will be used to develop a second parametric time to event (TTE) model that describes the time-varying probability of kidney graft failure up to five years.

Additional discussion focused on the potential utility of the proposed CTS tool in ISD clinical trials. There were questions as to whether the specific CTS tool would have utility given that longitudinal data for 1 year would need to be accrued prior to entry into a clinical trial. Instead, there seemed to be significant interest in iBox score as an endpoint to support accelerated approval or as a surrogate endpoint. FDA noted that there is a need to understand whether iBox score changes in response to intervention, and additional discussion would be needed to support the utility of the iBox as a surrogate endpoint. However, FDA expressed openness to discussing the use of iBox as a part of accelerated approval pathways in individual drug programs. FDA representatives were interested in seeing the current TTC manuscript under peer review, along with the peer reviewers' comments. TTC representatives discussed the consortium's future plans to compare the iBox score's ability to predict long-term outcomes in registry data.

3. NEXT STEPS

1. TTC will work internally to determine a proposed context of use for iBox and/or the CTS.
2. TTC will provide FDA with the iBox manuscript under review along with the peer reviewers' comments.