Clinical usefulness of gene-expression profile to rule out acute rejection after heart transplantation: CARGO II

Maria G. Crespo-Leiro1†, Jörg Stypmann2†, Uwe Schulz3, Andreas Zuckermann4, Paul Mohacsi5, Christoph Bara6, Heather Ross7, Jayan Parameshwar8, Michal Zakliczyński9, Roberto Fiocchi10, Daniel Hoefer11, Monica Colvin12, Mario C. Deng13, Pascal Leprince14, Barbara Elashoff15, James P. Yee15, and Johan Vanhaecke16*

1Instituto de Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitário de A Coruña (CHUAC), SERGAS, Universidade da Coruña (UDC), A Coruña, Spain; 2Department of Cardiovascular Medicine, University Hospital Muenster, Muenster, Germany; 3Ruhr University of Bochum, Bad Oeynhausen, Germany; 4Medical University of Vienna, Vienna, Austria; 5University Hospital Bern, Bern, Switzerland; 6Hannover Medical School, Hannover, Germany; 7Toronto General Hospital, Toronto, Canada; 8Papworth Hospital, Papworth Everard, Cambridge, UK; 9Silesian Center for Heart Disease, Zabrze, Poland; 10Ospedali Riuniti di Bergamo, Bergamo, Italy; 11Innsbruck Medical University, Innsbruck, Austria; 12University of Minnesota, Minneapolis, MN, USA; 13David Geffen School of Medicine, University of California, Los Angeles, CA, USA; 14Groupe Hospitalier Pitie-Salpêtrière, Paris, France; 15CareDx, Brisbane, CA, USA; and 16Department of Cardiology, University Hospital Gasthuisberg, Leuven, Belgium

Received 4 May 2015; revised 25 September 2015; accepted 25 November 2015

Aims
A non-invasive gene-expression profiling (GEP) test for rejection surveillance of heart transplant recipients originated in the USA. A European-based study, Cardiac Allograft Rejection Gene Expression Observational II Study (CARGO II), was conducted to further clinically validate the GEP test performance.

Methods and results
Blood samples for GEP testing (AlloMap®, CareDx, Brisbane, CA, USA) were collected during post-transplant surveillance. The reference standard for rejection status was based on histopathology grading of tissue from endomyocardial biopsy. The area under the receiver operating characteristic curve (AUC-ROC), negative (NPVs), and positive predictive values (PPVs) for the GEP scores (range 0–39) were computed. Considering the GEP score of 34 as a cut-off (6 months post-transplantation), 95.5% (381/399) of GEP tests were true negatives, 4.5% (18/399) were false negatives, 10.2% (6/59) were true positives, and 89.8% (53/59) were false positives. Based on 938 paired biopsies, the GEP test score AUC-ROC for distinguishing ≥3A rejection was 0.70 and 0.69 for ≥2–6 and >6 months post-transplantation, respectively. Depending on the chosen threshold score, the NPV and PPV range from 98.1 to 100% and 2.0 to 4.7%, respectively.

Conclusion
For ≥2–6 and >6 months post-transplantation, CARGO II GEP score performance (AUC-ROC = 0.70 and 0.69) is similar to the CARGO study results (AUC-ROC = 0.71 and 0.67). The low prevalence of ACR contributes to the high NPV and limited PPV of GEP testing. The choice of threshold score for practical use of GEP testing should consider overall clinical assessment of the patient’s baseline risk for rejection.

Keywords
Heart transplant • Organ rejection • Tests • Molecular diagnostics • AlloMap • Rejection surveillance

Introduction
A non-invasive gene-expression profiling (GEP) test (AlloMap®, CareDx, Brisbane, CA, USA) was first demonstrated to discriminate acute cellular rejection from quiescence in cardiac transplant recipients in a US-based clinical study, Cardiac Allograft Rejection Gene Expression Observational Study (CARGO). Estimates of the negative predictive value (NPV) associated with a GEP score

† These authors contributed equally to this work and shared co-first authorship.
*Corresponding author. Tel: +32 16 3 44251, Fax: +32 16 3 44240, Email: johan.vanhaecke@uzleuven.be
© The Author 2016. Published by Oxford University Press on behalf of the European Society of Cardiology
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
below a nominal threshold level (e.g. 34) were refined during re-
analyses of the CARGO dataset during the US Food and Drug Ad-
ministration (US FDA) regulatory clearance of AlloMap as a de novo
510 k in vitro diagnostic test (see Supplementary material online, Ap-
pendix S1). The randomized controlled Invasive Monitoring Attenu-
ation through Gene Expression (IMAGE) study showed the non-inferiority of clinical outcomes in patients managed with GEP
for rejection surveillance compared with patients managed with
conventional biopsy surveillance.\(^3\) The International Society of
Heart and Lung Transplantation (ISHLT) guidelines for care of heart
transplant recipients recommend that the GEP test can be used as a
non-invasive method for ruling out moderate to severe acute cellular
rejection in asymptomatic patients (i.e. without clinical suspicion
of allograft dysfunction).\(^4\) The CARGO II study was designed to
evaluate GEP in a different population from CARGO. There are
very few published clinical validation studies for commercially avail-
able diagnostic tests based on gene-expression profiling of periph-
eral blood, and AlloMap is the only product of this kind for use in
heart transplantation.\(^5\)

Therefore, the objective of the current report was to validate the
clinical performance of the gene-expression profiling technology in
an entirely independent large study population. The hypothesis of
this study is that the GEP test can discriminate acute cellular rejec-
tion from non-rejection status, with high negative predictive values
and overall accuracy similar to that previously demonstrated in
CARGO. Antibody-mediated rejection cases were not included in
these computations. The GEP test performance was evaluated in
two post-transplantation time windows: ≥2–6 months and >6
months, because prior evidence had indicated that the risk of rejec-
tion and performance of the GEP test may both be sensitive to time
post-transplantation.\(^1,6,7\)

Methods

CARGO II study design

The CARGO II was a prospective, observational, multi-centre study
with blood samples and associated clinical data collected from follow-up
visits of cardiac transplantation recipients from May 2005 through Feb-
ruary 2009 (ClinicalTrials.gov Identifier: NCT00761787). The 17 partici-
pating centres (13 Western European and 4 North American) obtained
Institutional Review Board approvals and signed written informed con-
ets from participating patients. The study was conducted in accord-
ance with the Good Clinical Practices and Declaration of Helsinki.

Venous blood was collected at clinic visits during which heart trans-
plant recipients underwent long-term ongoing rejection surveillance,
with or without endomyocardial biopsy (EMB). The blood samples
were collected before EMB. In general, patients were asked not to drink
or eat anything within 8 h before the procedure. An outline of the clin-
ic information collected at each visit, including patient inclusion and ex-
clusion criteria, and sample collection procedures, is provided in
Supplementary material online, Appendix S2.

Reference standard for rejection status

The reference standard for rejection status was the histology-based re-
jection grade (using the ISHLT 1990 grades: 0, 1A, 1B, 2, 3A, 3B, or 4) for
heart tissue obtained from an EMB performed on the same day as the
collection of blood for the GEP test.\(^6\) In 2005, the grading system was
revised as follows: Grade 0R-no rejection (no change from the 1990
grading system), Grade 1R-mild rejection (1990 Grades 1A, 1B, and 2), Grade 2R-moderate rejection (1990 Grade 3A), and Grade
3R-severe rejection (1990 Grades 3B and 4).\(^9\) In routine clinical practice,
patient care decisions are most strongly influenced by findings of rejec-
tion grades ≥3A (2R), which usually lead to a treatment intervention.
The histopathology grade is rendered by the local pathologist, who
has viewed the histology slides with light microscopy. In the CARGO
II study, a subset (see below) of all locally graded biopsy slides was
also sent for review by three other pathologists. The intention of this
independent assignment of rejection grades by a central panel of pathol-
ogists was to enhance the accuracy of the reference standard for rejec-
tion that was used to judge the performance of the GEP test.\(^10,11\)

Because of the high proportion of local EMB scores with Grade 0, and
the limited volume capacity for independent scoring by the reviewing
central pathologists, only one local rejection Grade 0 sample from
each subject was sent for central pathology reading; if a subject had
more than one sample with a local biopsy grade of 0, then one of
the samples was chosen at random. At most, only one Grade 1A sample
from each subject was sent for central pathology reading; if a subject
had more than one sample with a local biopsy grade of 1A, then one
of the samples was chosen at random. All samples that had local biopsy
grades of 1B, 2, and 3A or higher were sent for the independent central
pathology grading. For this report, each rejection case was qualified for
use as a reference standard of rejection if a moderate or more severe
grade of rejection, [i.e. ISHLT 1990 grade of ≥3A (2R)] was reported by
the local pathologist and this grade ≥3A (2R) was independently con-
firmed by at least one of three central reviewing pathologists. All
other centrally reviewed biopsy slides were classified as quiescent (i.e.
Grade 0 assigned by the local pathologist and at least 2 of 3 central
pathologists) or other non-rejection cases (i.e. did not meet either
the criteria for classification as quiescent or rejection). All samples
that had central pathology results and GEP score results were included in
the analyses.

Methodology of gene-expression profiling

For the GEP test, a whole blood sample was collected into a BD Vacu-
tainer\(^TM\) Cell Preparation tube (Becton Dickinson, NJ, USA) with
sodium citrate anticoagulant (CPT tube). Peripheral blood mononuclear
cells were isolated, lysed by RLT buffer (Qiagen) and the released ribo-
nucleic acid (RNA) was stabilized and frozen. Following shipment to the
CareDx reference laboratory in Brisbane, CA, USA, RNA was further
purified from the lysate; finally, complementary deoxyribo nucleic acid
(cDNA) was generated for use in the GEP test. The GEP test was con-
ducted in the CareDx Clinical Laboratory Improvement Amendments
(CLIA)-certified laboratory. The test uses real-time polymerase chain
reaction (RT–PCR) to quantitate expression levels of a preselected
panel of 20 genes: 11 genes informative about allograft rejection and 9
genes for normalization and quality control, used in the calculation of
a GEP logarithmic score that ranges from 0 to 39, Supplementary mate-
rial online, Appendix S3.\(^1\)

Sample selection

The blood samples included in this study conformed to the following in-
clusion criteria: obtained at least 55 days post-transplantation; > 30 days
after transfusion of blood products; > 21 days after administration of
≥ 20 mg/day of prednisone; and > 60 days after treating a prior rejection.

Statistical analysis

To calculate the GEP score sensitivity (ability to detect acute cellular re-
jection) and specificity (ability to detect the absence of rejection) and
generate an associated area under the receiver operating characteristic
(AUC-ROC) curve, positive predictive value (PPV), and NPV, each GEP sample was correlated to a paired, histology-based rejection grade (using the ISHLT 1990 grades: 0, 1A, 1B, 2, 3A, 3B, or 4 or equivalent ISHLT 2005 grades) for heart tissue obtained from an EMB performed on the same day.9

The main metric for validating the GEP test performance was the AUC-ROC, which is a widely used metric for new screening test discrimination.12,13 In early 2008, during the regulatory clearance of AlloMap as a FDA 510 k in vitro diagnostic test, the GEP test performance characteristics from the original CARGO study were reanalysed in cooperation with reviewing statisticians from the US FDA. The main change in the FDA-guided analyses compared with the methods used in the Deng 2006 publication was that all mild (i.e. < moderate [grade 3A (2R)]) rejection cases were included in the FDA-guided analyses which thereby include the full intended use population. Incorporating these mild rejection cases is called an intent-to-diagnosis (ITD) analysis (the analog of intent-to-treat or ITT analysis for randomized clinical trials of therapeutics). All centrally graded samples including all non-rejection samples (i.e. both quiescent and other samples that did not qualify to be classified as rejection) were included in the analysis; Supplementary material online, Appendix S1. The current CARGO II analyses use the same FDA-recommended methodology that was used in the analyses of the CARGO data: AUC-ROC of GEP score was estimated using a non-parametric method proposed by Obuchowski,14 altered slightly to account for the multiple records per subject.

The AUC-ROC and 95% confidence intervals were calculated using a bootstrap method for the data overall and also by time period: ≥ 2–6 months (≥ 55–182 days) post-transplantation and >6 months (>182 days) post-transplantation.

Clinical performance metrics were further characterized by computing the NPV and PPV for acute cellular rejection for each GEP ordinal score from 12 to 39. These characteristics were computed for two time periods post-transplantation: ≥ 2–6 months and >6 months.

The calculation of NPV and PPV involved an adjustment for prevalence of rejection, since not all local Grades 0 and 1A biopsies were sent for central pathology, as described. The adjustment was applied separately to the different time periods post-transplantation. The adjusted prevalence calculation method is shown in the statistical appendix, Supplementary material online, Appendix S4.

Logistic mixed-effects models were used to predict rejection status from GEP score, both in the absence and in the presence of other clinical variables (age, gender, primary indication for transplant, cytomegalovirus status, and pre-transplant use of a left ventricular assist device).

Minimization of bias

The study used a prospective design in which the blood samples and associated heart biopsy slides were selected from patients enrolled in the CARGO II Study. Sample labels were encrypted and the lab technicians were blinded to the patient data associated with each sample, minimizing any operator bias in determination of the GEP test scores.

Biopsy slides were de-identified prior to centralized pathology review and grading. Since patients may have more than one sample, the method of Emir et al.15 was used to analyse the data. This method assigns a weight to each sample based on the number of samples from the same patient within a category (rejection or non-rejection).

Results

From the total CARGO II enrollment, 499 patients had 1579 visits with paired EMB histopathology rejection grades and GEP scores that met inclusion criteria for the study analyses (Figure 1). Patients were predominantly Caucasian (89%), male (79%), and 50.8 years old on average, at the time of transplantation (Table 1). The first study visit occurred prior to 6 months post-transplantation in 88% of the patients. Moderate to severe acute cellular rejection was reported by the local pathologists in 53/1645 (3.2%) of biopsies at ≥ 2–6 months post-transplantation [representing 47/468 (10.0%) of patients] and in 53/1679 (3.2%) of biopsies at >6 months post-transplantation [representing 41/518 (7.9%) of patients].

The prevalence of moderate to severe grade rejection was 106/3325 (3.2%) in the population as a whole. When the rejection was also required to be confirmed by at least one central pathologist ([i.e. rejection was defined as 1 local grade ≥ 3A (2R) plus ≥ 1 central grade ≥ 3A (2R)], the prevalence was reduced to 2.0%.

The patient population with histopathology slides selected for independent central panel pathology rejection grading included 328 patients with 480 GEP test scores at ≥ 2–6 months post-transplantation, while 322 patients provided 458 GEP test scores at >6 months post-transplantation. Of a total of 938 centrally graded biopsies, 46 (4.9%) were deemed rejections, whereas 246 (26.2%) were classified as quiescent. The remaining 646 (68.9%) biopsies were intermediate (i.e. did not meet the criteria to qualify as either a rejection or a quiescent reference standard criteria). The implications of this subset with discordant grades from the pathologists are addressed in the discussion section of this paper. Of the 938 biopsies sent for central review, 919 were surveillance biopsies, 9 were for suspected rejection, and 10 were performed for other reasons. Out of the 46 centrally confirmed rejections, 43 were from surveillance biopsies, and 3 were biopsies performed because of suspected rejection. The three clinically suspected rejections that were confirmed by biopsy had GEP scores of 29, 34, and 37. The average GEP score of samples taken at the time of a biopsy-proven ≥ 3A (2R) rejection was significantly higher (28.3, SD 5.3) than the average GEP score of all samples taken at times when the biopsy revealed a non-rejection status (25.6, SD 5.9), P = 0.001. Of the 938 centrally graded biopsies, the distribution of local grades was as follows (n, %): Grade 0 (384, 40.5%); Grade 1A (245, 26.1%); Grade 1B (68, 7.2%); Grade 2 (131, 14.0%); Grade 3A (70, 7.5%); Grade 3B (5, 1%); Grade 4 (1, < 1%); local grades were not available for 34 (3.6%) biopsies. Of the 109 biopsies that were graded ≥ 3A (2R) by at least one central reviewer, 63 (58%) received a local grade of 2 or less (Table 2).

The mean AUC-ROC for the GEP test scores in the ≥ 2–6 months post-transplantation period was 0.70 with 95% confidence interval from 0.67 to 0.73. The mean AUC-ROC for the GEP test scores for the >6 months post-transplantation period was 0.69 with 95% confidence interval from 0.66 to 0.72. The GEP scores by time post-transplant are shown in Figure 2.

The CARGO II-based GEP NPV and PPV performance characteristics are shown in Tables 3 and 4. A GEP test score of ≥ 34 (often used in clinical practice as a threshold score for patients who are ≥ 6 months post-transplantation) corresponded to histology-based grade ≥ 3A (2R) rejection with a PPV of 4.0% at months ≥ 2–6 post-transplantation, and 4.3% at >6 months post-transplantation. The NPVs were 98.4% at months ≥ 2–6 post-transplantation and 98.3% at >6 months post-transplantation. In both time windows, the NPVs increased from 98.3 to >99.0% for decreasing threshold
values below 34. The corresponding PPVs decreased from 4.3 to 2.1. The number (and %) of patients with centrally adjudicated ≥3R (2R) rejection below or above GEP score of 34 is presented in Table 5. Using the score of 34 as a cut-off for patients who are 6 months post-transplantation, 95.5% (381/399) of GEP tests were true negatives, 4.5% (18/399) were false negatives, 10.2% (6/59) were true positives, and 89.8% (53/59) were false positives.

The probability of rejection using a threshold of ≥34 as a negative test moves from 3.2% pretest to 1.7% post-test, i.e. the NPV goes from 96.8% without the GEP test to 98.3–98.4% with a negative test, which occurred in 82.5–85% of the population. For ~15% of positive tests, the patients may be considered for a follow-up biopsy, 96% of which will be negative.

In the univariate model, GEP score and age were significantly associated with ≥3A (2R) acute cellular rejection (Table 6). In the multivariate model, after adjustment for clinical variables, a decreasing GEP score (P = 0.004) and age ≥54 years (P = 0.008) were associated with decreased probability of ≥3A (2R) rejection. Subgroup analyses of clinical variables for association of GEP score and ≥3A (2R) acute cellular rejection are shown in Table 7.

### Discussion

The performance characteristics of GEP scores to discriminate acute cellular rejection from non-rejection status of heart transplant recipients in CARGO II are very similar to the performance of GEP scores reported from CARGO, based on similar FDA-recommended analyses applied to these two independent data sets. This additional independent clinical validation is important, as such confirmatory studies are rare in the field of molecular diagnostic tests based on gene-expression profiling technologies; the hypothesis that the AlloMap GEP signature generated from assay of mRNA purified from circulating peripheral blood cells is informative of the immune status of the allograft is thus further validated by the results in this report.

The CARGO II study population is similar to most heart transplant study cohorts in demographic composition, but has a higher percentage of Caucasians than the US-based CARGO study. The 3.2% prevalence of moderate to severe rejection found in locally graded biopsies and estimated 13% incidence of rejection in patients is lower than the 66% rejection rate in patients reported during the...
The lower prevalence of acute cellular rejection in CARGO II results compared with estimates from experiences 20 years earlier may be attributable to more effective contemporary immunosuppression regimens. This prevalence of biopsy-based moderate to severe rejection grades in the current study was consistent with the more recent report of a similar population undergoing biopsies in the first year post-transplantation: biopsy-based acute rejections were reported to range from 1.9% (35 of 1875 biopsies) to 3.1% (58 of 1854 biopsies).

The current report finds that rates of rejection were similar in the early post-transplantation months (3.2% for months ≥ 2 to 6) compared with the later time post-transplantation (3.2% for Month 6). However, it is possible that the biopsies performed after Month 6 post-transplantation may have been performed more often due to clinical suspicion of rejection, whereas biopsies performed earlier...

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n = 499* rejections</th>
<th>n = 41 non-rejection</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years, mean (standard deviation)</td>
<td>50.8 (13.7)</td>
<td>46.1 (13.9)</td>
<td>51.2 (13.5)</td>
</tr>
<tr>
<td>Male sex</td>
<td>393 (79)</td>
<td>34 (83)</td>
<td>338 (78)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>443 (89)</td>
<td>35 (85)</td>
<td>386 (90)</td>
</tr>
<tr>
<td>African American</td>
<td>16 (3)</td>
<td>3 (7)</td>
<td>13 (3)</td>
</tr>
<tr>
<td>Asian</td>
<td>11 (2)</td>
<td>1 (2)</td>
<td>8 (2)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>9 (2)</td>
<td>1 (2)</td>
<td>8 (2)</td>
</tr>
<tr>
<td>Others</td>
<td>20 (4)</td>
<td>1 (2)</td>
<td>16 (4)</td>
</tr>
<tr>
<td>Non-Caucasian</td>
<td>56 (11)</td>
<td>6 (15)</td>
<td>45 (10)</td>
</tr>
<tr>
<td>Indication for cardiac transplantation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>185 (37)</td>
<td>16 (37)</td>
<td>160 (37)</td>
</tr>
<tr>
<td>Non-ischaemic cardiomyopathy</td>
<td>256 (51)</td>
<td>22 (54)</td>
<td>228 (53)</td>
</tr>
<tr>
<td>Other indications</td>
<td>38 (8)</td>
<td>3 (7)</td>
<td>43 (10)</td>
</tr>
<tr>
<td>Unknown indication</td>
<td>20 (4)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Interval between transplantation and study enrollment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolled prior to transplantation</td>
<td>70 (14)</td>
<td>6 (15)</td>
<td>62 (14)</td>
</tr>
<tr>
<td>&lt;6 months</td>
<td>367 (73)</td>
<td>29 (71)</td>
<td>323 (75)</td>
</tr>
<tr>
<td>6–12 months</td>
<td>24 (5)</td>
<td>2 (5)</td>
<td>15 (3)</td>
</tr>
<tr>
<td>13–36 months</td>
<td>20 (4)</td>
<td>2 (5)</td>
<td>18 (4)</td>
</tr>
<tr>
<td>37–60 months</td>
<td>2 (0)</td>
<td>0 (0)</td>
<td>2 (0)</td>
</tr>
<tr>
<td>&gt;60 months</td>
<td>16 (3)</td>
<td>2 (5)</td>
<td>11 (3)</td>
</tr>
<tr>
<td>Cytomegalovirus serology (IgG) status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donor and recipient positive</td>
<td>150 (30)</td>
<td>13 (32)</td>
<td>129 (30)</td>
</tr>
<tr>
<td>Donor and recipient negative</td>
<td>78 (16)</td>
<td>5 (12)</td>
<td>68 (16)</td>
</tr>
<tr>
<td>Donor positive and recipient negative</td>
<td>84 (17)</td>
<td>8 (20)</td>
<td>72 (17)</td>
</tr>
<tr>
<td>Donor negative and recipient positive</td>
<td>113 (23)</td>
<td>8 (20)</td>
<td>99 (23)</td>
</tr>
<tr>
<td>Unknown</td>
<td>74 (15)</td>
<td>7 (17)</td>
<td>63 (15)</td>
</tr>
<tr>
<td>Use of ventricular assist device before transplantation</td>
<td>118 (24)</td>
<td>6 (15)</td>
<td>110 (26)</td>
</tr>
<tr>
<td>Induction therapy (any)</td>
<td>384 (77)</td>
<td>34 (83)</td>
<td>295 (68)</td>
</tr>
<tr>
<td>Immunosuppressive therapy, at any time during study</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclosporin A</td>
<td>298 (60)</td>
<td>29 (71)</td>
<td>265 (59)</td>
</tr>
<tr>
<td>Tacrolimus</td>
<td>309 (62)</td>
<td>24 (59)</td>
<td>265 (61)</td>
</tr>
<tr>
<td>Mycophenolate mofetil or mycophenolic acid</td>
<td>421 (84)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prednisone</td>
<td>436 (87)</td>
<td>39 (95)</td>
<td>394 (91)</td>
</tr>
<tr>
<td>Left ventricular ejection fraction at first study visit, mean % (standard deviation)</td>
<td>64.1 (9.6)</td>
<td>63.1 (8.7)</td>
<td>64.7 (9.5)</td>
</tr>
</tbody>
</table>

Values are n (%) unless indicated otherwise.
IgG, immunoglobulin G.
*Twenty-seven patients did not have central pathology readings.
than Month 6 were probably more often conducted for surveillance of asymptomatic patients. This ascertainment bias is discussed further in the limitations section below.

The local biopsy-based rejection cases [grades $\geq 3A$ (2R)] that were confirmed by the independent panel of central pathologists in the current study served as the key reference standard for rejection that was used to judge the diagnostic performance of the GEP score. The reason these central pathologist confirmed rejections represent the most important factor in determining the clinical performance of the GEP test is that a biopsy finding of a rejection grade $\geq 3A$ (2R) generally calls for treatment of the patient, whereas for lower rejection grades or quiescence, there is the option to observe the patient’s interval clinical progress at a future visit. All 646 (68.9%) centrally graded biopsies that were by default assigned to an intermediate class (i.e. those biopsies that did not meet criteria to serve as rejection or quiescence reference standards) were included in the computation of the NPV and PPV; these cases were categorized as ‘non-rejection’. The relatively large number of biopsies assigned to this intermediate class reflects the high level of discordance between pathologists’ readings in this large proportion of biopsies with disparate reader scores of Grade 0, Grade 1A, Grade 1B, and Grade 2. Previous analysis of the CARGO

<table>
<thead>
<tr>
<th>Local grades</th>
<th>Central grades</th>
<th>$\geq 3A$ (2R) rejection</th>
<th>$&lt; 3A$ (2R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 3A$ (2R), n/n (%)</td>
<td>46/76 (60.5)</td>
<td>30/76 (39.5)</td>
<td></td>
</tr>
<tr>
<td>$&lt; 3A$ (2R), n/n (%)</td>
<td>63/862 (7.3)</td>
<td>799/862 (92.7)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2** Agreement between local and central pathology grades for rejection

**Figure 2** Gene-expression profiling score by the time post-transplant. The number of gene-expression profiling tests between Month 12 and Month 13 post-transplant ranged from 24 (Month 11) to 128 (Month 3). The number of gene-expression profiling tests between Month 14 and Month 24 post-transplant ranged from 2 (Month 24) to 11 (Month 15).
II data set demonstrated that both within the panel agreement and between the panel and local pathologists agreement decreased with increasing category of severity of rejection grade (from \(\approx 65\%\) to \(\approx 28\%\)). In that study, an EMB graded \(\geq 3A\) (2R) by one pathologist was unlikely to be equally graded by a given second pathologist. The observed low agreement between pathologists of 28% for grade \(\geq 3A\) (2R) EMB readings indicates the uncertainty of biopsy-based diagnosis of rejection, albeit this method remains the ‘gold’ standard for assessing the rejection status of a transplanted heart. In the current study, in \(\approx 40\%\) of cases (Table 2), a central pathologist ‘downgraded’ the local pathologists’ grades of \(\geq 3A\) (2R) to a grade \(< 3A\) (2R). The tendency for local pathologists to assign grades of \(\geq 3A\) (2R) more often than central pathologists may be attributable to the local pathologist being influenced by clinical information that is not available to the central reviewers. The discordance of grade assignment by local and central pathologists is also attributable to the diagnosis depending on subjective pattern recognition as well as potential tissue sampling errors and/or the non-specificity of the histology findings.

A major reason the ISHLT 2005 rejection grading re-classification combined the several 1990 subcategories of intermediate rejection grades is that there is uncertainty of the clinical relevance and prognostic implications of these intermediate grades. Most often, a biopsy finding of an intermediate grade will not lead to treatment or other interventional actions: follow-up depends on the clinical context. The GEP test is not designed to correlate with or classify intermediate grades of rejection. However, a sub-analysis of the original CARGO data demonstrated that GEP scores discriminate moderate to severe rejection from mild rejection (Grades 1A and 2). A subgroup of mild rejection cases, defined as Grade 1B according to the 1990 grading system, share a molecular signature more consistent with moderate to severe rejection. The clinical relevance of these intermediate grades of rejection is not part of the current CARGO II study and remains to be defined.

Because the AUC-ROC is easy to understand and it does not rely on specific thresholds, it is commonly used to measure the accuracy of a diagnostic test. Values for the AUC-ROC range from 0.5 (uninformative) to 1.0 (perfect discrimination). An example of AUC-ROC in the cardiovascular area is the Framingham Risk Score, which is based on traditional cardiovascular risk factors. This score yields an AUC-ROC of 0.68, based on age and gender, and goes up...
to 0.75 when other risk factors are added. The CARGO II GEP AUC-ROC of 0.70 (95% confidence interval 0.67 to 0.73) and 0.69 (95% confidence interval 0.66 to 0.72) for patient samples from the ≥2–6 months and ≥6 month time windows, respectively, were comparable to the CARGO results (see Supplementary material online, Appendix S1). The limited sensitivity and PPV of the GEP score for a histology diagnosis of acute cellular rejection are similar to the sensitivity and PPV reported between expert pathologists. Observed concordance between local and collaborating study pathologists for EMB scoring was 28% for grade ≥3A (2R) in the CARGO II study and 40% for grade ≥3A (2R) using 1990 ISHLT grading in the CARGO Study. This circumstance of a low concordance of rejection grades between expert pathologists has also been reported with interpretation of biopsies from transplanted kidneys. Kidney transplant researchers who are seeking to develop molecular signatures that are correlated to histology reference standards of rejection have observed a problem that applies also to heart transplant: the ‘reference standard-related bias’. The new test, whether based on a molecular classifier or the opinion of a second pathologist, will not agree perfectly with the existing reference standard (the opinion of a different pathologist), when there is a high degree of subjectivity involved in the assignment of the reference standard. ‘Noise’ and the associated range of uncertainty within heart (or other transplanted organ such as kidney) tissue

<table>
<thead>
<tr>
<th>Table 4</th>
<th>CARGO II GEP test performance results in the &gt;6 months post-transplantation period</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEP score</td>
<td>Scores below threshold n (%)</td>
</tr>
<tr>
<td>12</td>
<td>6 (1.3)</td>
</tr>
<tr>
<td>13</td>
<td>9 (2.0)</td>
</tr>
<tr>
<td>14</td>
<td>14 (3.1)</td>
</tr>
<tr>
<td>15</td>
<td>19 (4.1)</td>
</tr>
<tr>
<td>16</td>
<td>27 (5.9)</td>
</tr>
<tr>
<td>17</td>
<td>32 (7.0)</td>
</tr>
<tr>
<td>18</td>
<td>43 (9.4)</td>
</tr>
<tr>
<td>19</td>
<td>55 (12.0)</td>
</tr>
<tr>
<td>20</td>
<td>62 (13.5)</td>
</tr>
<tr>
<td>21</td>
<td>79 (17.2)</td>
</tr>
<tr>
<td>22</td>
<td>95 (20.7)</td>
</tr>
<tr>
<td>23</td>
<td>108 (23.6)</td>
</tr>
<tr>
<td>24</td>
<td>130 (28.4)</td>
</tr>
<tr>
<td>25</td>
<td>146 (31.9)</td>
</tr>
<tr>
<td>26</td>
<td>165 (36.0)</td>
</tr>
<tr>
<td>27</td>
<td>198 (43.2)</td>
</tr>
<tr>
<td>28</td>
<td>220 (48.0)</td>
</tr>
<tr>
<td>29</td>
<td>251 (54.8)</td>
</tr>
<tr>
<td>30</td>
<td>294 (64.2)</td>
</tr>
<tr>
<td>31</td>
<td>331 (72.3)</td>
</tr>
<tr>
<td>32</td>
<td>357 (77.9)</td>
</tr>
<tr>
<td>33</td>
<td>378 (82.5)</td>
</tr>
<tr>
<td>34</td>
<td>399 (87.1)</td>
</tr>
<tr>
<td>35</td>
<td>418 (91.3)</td>
</tr>
<tr>
<td>36</td>
<td>439 (95.9)</td>
</tr>
<tr>
<td>37</td>
<td>450 (98.3)</td>
</tr>
<tr>
<td>38</td>
<td>454 (99.1)</td>
</tr>
<tr>
<td>39</td>
<td>455 (99.3)</td>
</tr>
</tbody>
</table>

GEP test scores range from 0 to 39.

Table 5 True positive, false positive, true negative, and false negative rates based on GEP threshold score ≥34

<table>
<thead>
<tr>
<th>GEP score</th>
<th>≥2–6 months post-transplant</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥34, n/n (%)</td>
<td>5/55 (9.1) TP</td>
</tr>
<tr>
<td>&lt;34, n/n (%)</td>
<td>17/425 (4.0) FN</td>
</tr>
</tbody>
</table>

Example: 5/55, 5 = number of cases with GEP score ≥34 and ≥3A (2R) rejection; 55 = sum of ≥3A (2R) and <3A (2R) rejections with GEP score of ≥34.

GEP, gene-expression profiling; FN, false negative; TP, true positive.
histology is intrinsic, due to a variety of causes as summarized above. Transcending this limitation of histology, it may be useful to correlate the performance of GEP to alternative clinical endpoints such as organ function; this approach is outside the scope of the current paper, but the potential of such approaches has been reported.

Although CARGO II, like CARGO, was a landmark study in cumulative number of patients and surveillance samples/visits collected, there was a low cumulative incidence of observed rejections. The relatively small number of rejection episodes and the spread of these cases over a wide post-transplantation time range likely constrain the GEP test performance characteristics (i.e. failure of GEP test to achieve statistical significance in some of our subgroup analyses could be attributed to imbalance in the dataset and to the very low sample size for some covariates, Table 7). The GEP score was an equally significant predictor of acute cellular rejection (≥2R) in both univariate and multivariate models (Table 6). Nevertheless, the discovery of asymptomatic rejection may be enhanced if a GEP score exceeding threshold (i.e. 34) is used to lead to performance of a biopsy, since the GEP score PPV in this circumstance may be approximately two-fold higher than the prevalence of rejection in the surveillance population.

The AUC-ROC was analysed in two time windows, because the mean GEP score in a large, unselected, patient population tends to gradually rise from Month 2 to Month 6; this upward trend is less pronounced from Month 7 to 12; the mean GEP score reaches a

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Univariate and multivariate analyses for variables predictive of acute cellular rejection ≥3A (2R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics</td>
<td>Univariate Coefficient</td>
</tr>
<tr>
<td>Gene-expression profiling score (incremental unit increase by one)</td>
<td>0.08</td>
</tr>
<tr>
<td>Age, ≥54</td>
<td>−0.02</td>
</tr>
<tr>
<td>Gender, male</td>
<td>0.09</td>
</tr>
<tr>
<td>Primary indication for cardiac transplantation, ischaemic</td>
<td>−0.02</td>
</tr>
<tr>
<td>Cytomegalovirus serology status, donor or recipient positive</td>
<td>0.42</td>
</tr>
<tr>
<td>Left ventricular assist device use, yes</td>
<td>−0.59</td>
</tr>
</tbody>
</table>

*Positive coefficient indicates a probability of acute cellular rejection ≥3A (2R) increases with specified variable setting; negative coefficient indicates decreased risk of acute cellular rejection ≥3A (2R).

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Univariate analyses of clinical variables for significance of gene-expression profiling score in estimating risks of ≥3A (2R) acute cellular rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covariate</td>
<td>≥3A (2R) rejection (n)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>≥54 years</td>
<td>13</td>
</tr>
<tr>
<td>&lt;54 years</td>
<td>33</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>37</td>
</tr>
<tr>
<td>Female</td>
<td>9</td>
</tr>
<tr>
<td>Primary indication for cardiac transplantation</td>
<td></td>
</tr>
<tr>
<td>Ischaemic</td>
<td>17</td>
</tr>
<tr>
<td>Other</td>
<td>29</td>
</tr>
<tr>
<td>Cytomegalovirus serology status, donor or recipient positive</td>
<td></td>
</tr>
<tr>
<td>Donor or recipient positive</td>
<td>39</td>
</tr>
<tr>
<td>Donor or recipient negative or unknown</td>
<td>7</td>
</tr>
<tr>
<td>Left ventricular assist device use</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td>No</td>
<td>40</td>
</tr>
</tbody>
</table>

Reported P-values represent significance of gene-expression profiling score in estimating risks of ≥3A (2R) acute cellular rejection within subsets of patients defined by each covariate.
plano and remains stable from Years 1 to 5 post-transplantation.7
Findings from this study confirmed the GEP score trajectory be-
tween Months 2 and 12 post-transplant. After 12 months, the CAR-
GO II mean GEP scores are much more variable and lower than
observed in the larger database. This discrepancy may be explained
by the small number of samples available from the current study
after 12 months (Figure 2) and patient selection differences between
the studies. The early time dependency of average GEP scores and
acquired practical experience has led many centres to use 34 as the
selected threshold to rule out rejection for patients who are
>6–12 months post-transplantation, and to consider a lower threshold
(e.g. 30) for patients who are 2–6 months post-transplantation.24
The GEP test is `intended to aid in the identification of heart trans-
plant recipients with stable graft function who have a low probability
of moderate/severe ACR at the time of testing in conjunction with
standard clinical assessment’ (FDA cleared intended use statement).
Specific threshold values are not designated in the approved label-
ing; rather, the clinician may choose how to select the test score
threshold and make an associated clinical interpretation based on
his/her experience with the test and the overall, comprehensive clin-
ical context of the patient under evaluation. Thus, the complete list-
ing of NPVs and PPVs for each GEP score from 0 to 39 is provided
for the clinician to refer to in considering how a specific test score
may aid in ‘ruling out’ ACR. The threshold of ≥34 was used in the
IMAGE outcomes trial: for patients at least 6 months post-
transplant, a score of ≥34 was assumed in IMAGE to be equivalent
to a biopsy diagnosis of no rejection. In addition, useful clinical information may also be obtained by considering findings from a series of
GEP scores in a patient, since each patient may serve as his/her own
control of a ‘normal’ GEP score: within patient variability of GEP
scores (after 1 year post-transplantation) may be less than that be-
tween patients.22
As previously characterized in the CARGO study,6 the current
CARGO II results confirmed that the high NPV associated with
GEP scores below a nominal threshold may be reasonably used in
lieu of surveillance biopsies to rule out rejection. The IMAGE3
and E-IMAGE24 studies have demonstrated that the number of sur-
veillance biopsies was safely reduced while the clinical outcomes of
patients managed with GEP surveillance were non-inferior to out-
comes in patients managed with the biopsy strategy.

The results of this study must be interpreted in the context of
some limitations. The study excluded GEP data and biopsies from
visits earlier than 55 days post-transplantation, so the analyses do
not inform us about GEP performance or rejection incidence in
the first 54 days post-transplantation. Also, because patients treated
for rejection within the prior 30 days were excluded, we may have
underestimated the incidence of rejection that might be found in the
entire population. Additionally, in this study, not all GEP tests had
paired biopsy results. Because of the very small number and propor-
tion of clinically suspected, biopsy confirmed, rejections of grade
≥3A (2R), rejections discovered during routine surveillance and re-
jections confirmed with biopsies performed for clinical suspicion
were pooled together in the analyses. These reference rejection
cases, a combination of clinically silent and clinically suspected rejec-
tion, may not have identical GEP signatures. Although the GEP test is
intended for use in a surveillance setting, the discovery of subclinical
rejection is dependent on the frequency of surveillance that varied
at the participating centres. Additionally, the histology grading is an
imperfect reference standard for diagnosis of rejection, although it
remains the clinical standard of care for diagnosis of rejection in
heart transplantation.

The central pathology feature of this study does not reflect prac-
tical clinical care: no clinician relies on obtaining additional histo-
pathology reports from ‘blinded’ independent pathologists to
decide on treatment of a patient. We utilized the central biopsy
reading in this study to enhance the accuracy of diagnostic classifi-
cation of the reference cases. Finally, the question remains whether
GEP scores can be potentially used to optimize a post-transplant immu-
nosuppression regimen. This study was not designed to address
that question. Future studies should be conducted to evaluate the
clinical utility of the GEP test to guide immunosuppression medication
intensity to improve long-term outcomes in transplant recipients.

Conclusion
The GEP test performance characteristics in this CARGO II study
are consistent with prior CARGO findings. The low prevalence of
ACR contributes to the high NPV and limited PPV of GEP testing.
Taken together with clinical trial evidence that equivalent clinical
outcomes may be attained in patients managed with GEP compared
with patients managed with biopsies for acute cellular rejection sur-
veillance, the CARGO II results further validate that GEP testing may
be used as part of a non-invasive surveillance strategy to rule out
acute cellular rejection in heart transplant recipients. The choice of
threshold score for practical use of GEP testing to rule out rejec-
tion should consider overall clinical assessment of the patient’s base-
line risk for rejection.

Supplementary material
Supplementary material is available at European Heart Journal online.

Authors’ contributions
B.E. performed statistical analysis. J.P.Y. handled funding and super-
M.C.D., P.L., and J.V. acquired the data. M.G.C.-L., J.S., A.Z., P.M., and
J.V. conceived and designed the research. J.V., M.G.C.-L., J.S., and
J.P.Y. drafted the manuscript. P.M., M.C., J.P.Y., and H.R. made critical
revision of the manuscript for key intellectual content.

Acknowledgements
David Hiller, PhD, provided statistical support and Helen Nicely,
PhD, provided editorial assistance.

Funding
This work was supported by CareDx Inc. (ClinicalTrials.gov Identifier:
NCT00761787). Funding to pay the Open Access publication charges
for this article was provided by CareDx, Inc.

Conflict of interest: M.C.D. serves on the advisory board of CareDx.
J.S., U.S., P.M., A.Z., and M.C.D. received study support from CareDx.
J.P.Y. is an employee owning equity in CareDx. B.E. has been a statistical
consultant to CareDx. No other potential conflict of interest relevant to this article was reported.

References


