# STATISTICAL ANALYSIS PLAN

August 20, 2017

A MULTICENTER STUDY TO EVALUATE PERFORMANCE OF AN AUTOMATED DEVICE FOR THE DETECTION OF DIABETIC RETINOPATHY

Version 2

[redacted]

For

IDx LLC
485 Highway 1 W
Iowa City IA 52246

## CONFIDENTIALITY STATEMENT

**Signature Page**

A MULTICENTER STUDY TO EVALUATE PERFORMANCE OF AN AUTOMATED DEVICE FOR THE DETECTION OF DIABETIC RETINOPATHY

Study: **A MULTICENTER STUDY TO EVALUATE PERFORMANCE OF AN AUTOMATED DEVICE FOR THE DETECTION OF DIABETIC RETINOPATHY**

**Written by:**

_____ _____

[redacted]

Date

**Approved by:**

_____ _____

[redacted]

Vice President of Quality & Legal Affairs

Date

_____ _____

[redacted]

Executive Vice President & Chief Operating Officer

Date

_____ _____

[redacted]

President

Date

## 1. BACKGROUND

The IDx-DR device is intended to be used by health care providers to automatically screen for more than mild diabetic retinopathy (mtmDR). The IDx-DR outputs correlate with levels of disease identified in the America Academy of Ophthalmology's Preferred Practice Pattern (AAO PPP), which are associated with well-characterized management recommendations [1]. According to the AAO PPP, patients presenting with mtmDR have a level of disease associated with management recommendations outside of the normal 12-month screening period for patients with diabetes, and the physician overseeing the exam is advised to refer them to an eye care provider.

The IDx-DR Device, will produce a report for the physician overseeing the exam. IDx-DR will produce **one** of three outputs:

1. The IDx-DR result MORE THAN MILD DIABETIC RETINOPATHY DETECTED: REFER TO AN EYE CARE PROFESSIONAL indicates the presence of more than mild diabetic retinopathy, which includes moderate non-proliferative diabetic retinopathy, severe non-proliferative retinopathy (NPDR), proliferative retinopathy (PDR), and/or macular edema. Per the AAO PPP, patients with mtmDR have a level of disease associated with management recommendations outside of the normal 12-month screening period for patients with diabetes, and the physician overseeing the exam is advised to *refer* them to an eye care provider.

2. The IDx-DR result of NEGATIVE FOR MORE THAN MILD DIABETIC RETINOPATHY: RESCREEN IN 12 MONTHS will indicate that IDx-DR did not detect macular edema, moderate non-proliferative retinopathy, severe non-proliferative retinopathy, or proliferative retinopathy. Per the AAO PPP, it is recommended that patients be *rescreened annually*.

3. The IDx-DR report will indicate EXAM QUALITY INSUFFICIENT when the IDx-DR Exam Quality Index determines that the result should not be provided to the user due to insufficient fundus image quality submitted for analysis. The Exam Quality Index threshold was designed to determine when disease cannot be effectively ruled out from a set of exam images. The device will give the user guidance on what images caused the problem so the user can take action to efficiently re-image the patient either non-mydriatically or in conjunction with the use of pharmacologic dilation. Guidance to the user will be expressed as "Insufficient Quality" (an indication of dark or obscured images) or "Protocol Failure" (an indication that the submitted images do not adhere to the protocol of one fovea centered and one macula centered image per eye). Patients who are unable to receive a result other than "EXAM QUALITY INSUFFICIENT" after reimaging will ultimately be recommended for *referral* to an eye care professional in product labeling.

In rare cases, the IDx-DR Device may also produce a result of "Technical Failure – Exam not Suitable for Analysis." This occurs when the exam submitted is of such poor quality that IDx-DR is unable to recognized submitted images as fundus photographs and therefore cannot analyze the image due to insufficient quality. Note that this result is different from "Exam Quality Insufficient," which is a result where the IDx-DR exam quality index is designed to override a diagnostic result that would otherwise be produced.

The IDx goal is to deliver a level of performance that aligns with the FDA's target that the IDx-DR Device rule out lower bounds of at least 75% sensitivity and at least 77.5% specificity for the mtmDR configuration, with study success defined as achieving at least 80% sensitivity and at least 80% specificity.

All diagnoses will be made according to the Fundus Photograph Reading Center (FPRC).

## 2. PROTOCOL OVERVIEW

### 2.1. Study Design

#### 2.1.1. Prospective Cohort

This is a multi-center pivotal study for registration purposes. Study subjects (n=845 after exclusions) [redacted] will be recruited at primary care sites. Between 10 and 15 independent primary care sites will participate in the study.

#### 2.1.2. Enriched Cohort

At the start of the study all participants with diabetes who meet inclusion and exclusion criteria will be enrolled. To avoid excessive enrollment in any one stratum (no/mild, mDR, or vtDR), the totals will be monitored monthly [redacted].

The impact of enrichment [redacted] is not known. The goal will be to recruit enough subjects to satisfy each stratum. [redacted].

Monitoring of strata will be performed by the statistician while maintaining masking to IDx-DR results; [redacted]

#### 2.1.3. Precision Study

[redacted].

### 2.2. Independent Reading Center

The FPRC independent Reading Center will determine the reference standard for the severity of retinopathy and diabetic macular edema according to the Early Treatment of Diabetic Retinopathy Grading System (ETDRS) based on W-4 fundus photographs. Center involved macular edema will be assessed from OCT images according to the RC protocol, to be used for secondary endpoint analysis. Incidental findings and an indication of media opacity presence will also be documented.

### 2.3. Statistical Monitoring

The statistician will compare the performance of the IDx-DR device to FPRC readings in terms of sensitivity and specificity as well as Receiver Operator Characteristics for the device's mtmDR classifier algorithm. Results will also be monitored for enrichment outcomes.

## 3. STATISTICAL ASSUMPTIONS AND SAMPLE SIZE CALCULATIONS

### 3.1. [redacted]

[redacted]

Per the FPRC's Grading Charter, the following table summarizes the conversion of ETDRS scores to ground truth for the study:

| IDx-DR Output | | AAO PPP | ETDRS DR Grades | ETDRS Macular Edema Determination |
|---|---|---|---|---|
| No/Mild DR | | Normal or Minimal NPDR | 10, 12, 14A, 14B, 14C, 14Z, 15, 20 | |
| | | Mild NPDR | | |
| mtmDR | | Moderate NPDR | 35A, 35B, 35C, 35D, 35E, 35F, 43A, 43B, 47A, 47B, 47C, 47D | Definite Criteria 1: zone of retinal thickening > 1DA, part < 1 DD from center<br>Definite Criteria 2: retinal thickening or adjacent HE < 600µ from center |
| | vtDR | Severe NPDR | 53A, 53B, 53C, 53D, 53E, 60, 61A, 61B, 65A, 65B, 65C, 71A, 71B, 71C, 71D, 75, 81, 85A, 85B | |
| | | Non-high-risk PDR | | |
| | | High Risk PDR | | |
| | | Macular Edema | | |

Note that subjects with a grade of 90 will be excluded from primary analysis as being "ungradable" by FPRC for diabetic retinopathy severity level.

[redacted]

### 3.2. Co-primary Effectiveness Endpoints

Sensitivity and specificity will be the co-primary effectiveness endpoints to be evaluated against FPRC determination (truth) in the ITS (Intent to Screen) population for whom IDx-DR provides a diagnostic result (primary) against pre-established standards (null hypotheses) to be rejected if too low relative to these pre-defined thresholds.

The computations for the sensitivity and specificity measures are displayed in Table 4.7.1. An mtmDR subject with a positive algorithm counts towards the sensitivity numerator while a Mild + No DR subject with a negative algorithm will count towards the specificity numerator. Every case will be accounted for in the sensitivity and specificity calculations.

### 3.3. Closed Testing Strategy

The closed testing strategy requires statistical significance to be separately met for both sensitivity and specificity

### 3.4. [redacted]

The IDx-DR Device and its algorithm will be locked prior to study launch. A version of the tested software will be placed in third-party escrow during the study, and will be available for comparison to the algorithm used in the study if questions of algorithm alterations arise. [redacted]

### 3.5. Enrichment Strategy

Prospective enrichment will be used. [redacted]. Comparability will be assessed using two-sided 95% confidence intervals for the respective cohorts.

### 3.6. Key Assumptions

The following assumptions govern sample size:

- [redacted].
- The algorithms will be locked prior to subject enrollment.
- [redacted].
- The primary population for analysis will be patients from the Intention to Screen (ITS) population that have an IDx-DR result and a reference standard from the Reading Center.
- [redacted].
- To address incomplete data, prospective rules for case inclusion are summarized (see Section 4.7.7.2)

  Both sensitivity and both specificity must achieve statistical significance to support study success. This strategy preserves the overall Type I error.

The statistical assumptions that were used to plan the study are as follows:

- Sample size calculations were performed at the subject-level.
- Use of conservative one-sided hypothesis testing with overall one-sided 2.5% Type I error [redacted].
- The null hypotheses will be tested separately for mtmDR sensitivity and mtmDR specificity.
- All combined mild DR + no DR subjects will be included in the mtmDR specificity calculation. All mtmDR [redacted] subjects will be included in the more than mild DR sensitivity calculation.
- Subjects that received a IDx-DR result of "Insufficient Exam quality" will be excluded from primary analysis. [redacted]

[redacted]. As with all subjects, these subjects will also have been imaged using a 4-W protocol, with the wide field photos used by the FPRC to determine the reference standard.

### 3.7. Hypothesis Tests

The null hypotheses to be tested against a higher alternative hypothesis for the ITS population follow:

- mtmDR sensitivity: ≤75% null hypothesis vs >75% alternative hypothesis ([redacted])
- mtmDR specificity: ≤77.5% null hypothesis vs >77.5% alternative hypothesis ([redacted])

[redacted]

## 4. STATISTICAL METHODS

This analysis plan describes the planned data analyses.

### 4.1. General Considerations

- This Statistical Analysis Plan (SAP) is based on the Clinical Trial Protocol for **A MULTICENTER STUDY TO EVALUATE PERFORMANCE OF AN AUTOMATED DEVICE FOR THE DETECTION OF DIABETIC RETINOPATHY AND DIABETIC MACULAR EDEMA** dated May 2017.
- All calculations will be performed using SAS statistical software, version 9.1 or later, and StatXact, version 10 or later.
- Final analysis will be conducted after the last subject completes evaluation by the RC.
- Analyses will be conducted at the subject-level (since both eyes are required to run the IDx algorithm).
- Study data will be used to simultaneously evaluate the mtmDR algorithm with study success defined to be rejecting both study hypothesis (see Section 4.7).
- All hypothesis tests will account for all subjects as previously stated (see Section 3.6).
- Testing will be performed for the ITS population for whom IDx-DR delivers a screening result and full ITS (secondary) to support labeling.
- To be included in ITS effectiveness analyses, subjects must be declared evaluable according to the FPRC and have a valid IDx-DR test.
- To be included in the primary effectiveness analyses, subjects must be declared evaluable according to the FPRC and have an IDx-DR screening test result.
- One-sided 2.5% Type 1 error was used to plan sample size.
- One-sided p=0.025 will be required to achieve statistical significance for effectiveness; no p-value adjustment will be made for testing effectiveness endpoints since both specificity and sensitivity criteria must be met for study success.
- There are 18 effectiveness endpoints of interest (see Section 4.7.4) with two primary and six secondary endpoints.
- The literature suggests that there will be five times as many no DR cases as mild DR cases.
- The effect of subject losses will be assessed using robustness analyses for the two primary effectiveness endpoints (specificity and sensitivity).
- Verification bias will be investigated in excluded subjects.
- All effectiveness analyses will be performed using the ITS population for who IDx-DR delivers a screening result. Additional analyses will support labeling claims in the absence of a diagnostic result due to low quality imaging in a PCP setting by previously untrained ophthalmic photographers.
- Enriched subjects will be analyzed in combination with other subjects as well as separately to assess their impact.

### 4.2. Study Objectives

The study objectives are to establish mtmDR specificity and the corresponding sensitivity.

### 4.3. Photo Acquisition

The study population is described at the subject level. All subjects will be evaluated by photography and these photos will be subsequently evaluated via IDx-DR so no formal safety events are expected. Subjects may be reimaged per the IDx-DR imaging procedure if the initial image quality is not sufficient and the subject is willing to be retested.

### 4.4. Analysis Populations

The following database components will be used to determine effectiveness analysis population inclusion:

- Subject eligibility
- IDx-DR results

- FPRC evaluation.

Each data source will be reviewed by the Contract Research Organization (CRO) for the algorithm processing and results reflecting input from appropriate medical, clinical, data management, and statistical personnel. Subject inclusion / exclusion status with respect to the study populations will be made without knowledge of IDx-DR outcome.

### 4.4.1. Intent-to-screen (ITS) Population

The ITS population is described at the subject level and represents all eligible subjects with a Reading Center reference standard result. Study eligibility will be ascertained to exclude ineligible subjects from the ITS population. The primary study analyses will include the ITS population for whom IDx-DR delivers a screening output (i.e., a result of mtmDR or no/mild DR). Additional ITS population analyses will be confirmatory relative to these primary analyses for assessment of effectiveness, with reference standard established by the FPRC, which will determine the DR category each subject belongs to: [redacted].

Rules regarding ITS exclusion decisions will be prospectively established prior to database lock and reviewed by individuals without knowledge of the IDx-DR outcome by subject.

## 4.5. Subject Disposition

All enrolled subjects will be included in summaries of disposition and evaluation for analysis sets. Accounting will be provided for the following parameters:

1. Number of subjects by study site including those that receive a result of insufficient exam quality.
2. Numbers of subjects according to eligibility status overall and by study site.
3. Number of subjects able to be evaluated for diagnostic category (including enriched subjects).
4. Number of subjects with useable photos for the IDx-DR evaluation.
5. Construction of the ITS populations by subject from exclusions.
6. Study completion status and reasons for discontinuation by subject.

Subject exclusions from analysis sets and associated reasons will be tabulated.

The protocol deviations leading to ITS exclusions will be listed.

## 4.6. Demographic and Other Baseline Characteristics

Sex, age, race, and ethnicity will be presented using descriptive statistics by sites and by category for the overall ITS population and for the respective cohorts. Baseline characteristics will be compared first within the prospective enrichment cohort; [redacted] and the comparability of baseline characteristics corresponding to any change will be presented for ethnicity and race, cataract, incidental findings, and mtmDR characteristics. Cohort comparability will be assessed across cohorts using two-sided 95% confidence intervals for the paired differences between cohorts; a logistic regression model will be used for binary baseline measures as the dependent outcome, while an ANOVA model will be used for continuous measures as the dependent outcome.

## 4.7. Effectiveness Analyses

### 4.7.1. Statistical Terminology

The measures of screening accuracy are defined as follows:

| | | Reference Standard Panel | |
|---|---|---|---|
| | Outcome | Positive | Negative |
| Test | Positive | TP | FP |
| | Negative | FN | TN |

- Sensitivity = TP/(TP+FN)
- Specificity = TN/(TN+FP)
- Positive predictive value (PPV) = TP/(TP+FP)
- Negative predictive value (NPV) = TN/(FN+TN)
- Positive likelihood ratio (PLR) = [TP/(TP+FN)]/[FP/(FP+TN)]
- Negative likelihood ratio (NLR) = [FN/(TP+FN)]/[TN/(FP+TN)]
- Odds ratio (OR) = (TP/FP)/(FN/TN)

### 4.7.2. Effectiveness Analysis Strategy

A frequentist analysis approach will be used. One-sided 97.5% confidence intervals will be computed. P-values will be computed.

Analysis will be run in SAS in a GCP controlled environment.

### 4.7.3. Final Analysis

The generalized hypotheses of interest are

$$H_0: p \leq p_0 \text{ vs. } H_A : p > p_0$$

where $p$ is the sensitivity or specificity of the proposed diagnostic test [redacted] One-sided testing will be performed for both sensitivity and specificity; one-sided 2.5% Type I error will be used which results in a one-sided 97.5% rejection rule per hypothesis.

### 4.7.4. Primary Effectiveness Endpoints

The primary effectiveness endpoints for this study are the sensitivity and specificity. One-sided 97.5% confidence intervals for the lower bounds will be calculated for these binary measures. Descriptive statistics will be presented for the percent positive separately for each of mtmDR and no/mild DR. Results will be displayed per cohort and combined. All subjects from both cohorts will be placed into and analyzed according to the respective [redacted] classifications established according to the FPRC grading charter.

In calculating the mtmDR sensitivity and specificity, the estimation analyses will be carried out on the combined cohorts. [redacted] Results will be reported as posterior means, medians and with corresponding one-sided 97.5% confidence intervals (CI).

Analyses will be run for the ITS population with and without imputation for "Insufficient Exam Quality" cases. The ITS population analysis will be primary based on subjects with an IDx-DR screening result,

with secondary analyses preforming different imputations for those patients with a result of "Insufficient Exam Quality".

The following additional analyses will be performed separately for sensitivity and specificity per cohort and pooled with a single covariate included for the enriched cohort in the respective logistic regression models.

### 4.7.5. Secondary Effectiveness Endpoints

Additional, non-primary analyses will also be performed to assess the following overall and by cohort:

- The predictive value of a positive IDx test for mtmDR
- The predictive value of a negative IDx test for mtmDR:
- The positive and negative likelihood ratios for mtmDR
- The odds ratios for mtmDR:

[redacted] See below for endpoint calculations.

### 4.7.6. Effectiveness Analysis Methodology Summary

There will be two primary, six secondary, and 10 confirmatory effectiveness endpoints [redacted] The p-values will be computed for all subjects combined for the ITS population. [redacted]

| [redacted] | [redacted] |
|---|---|
| [redacted] | [redacted] |
| [redacted] | [redacted] |
| [redacted] | [redacted] |

[redacted]

### 4.7.7. Further Effectiveness Analyses

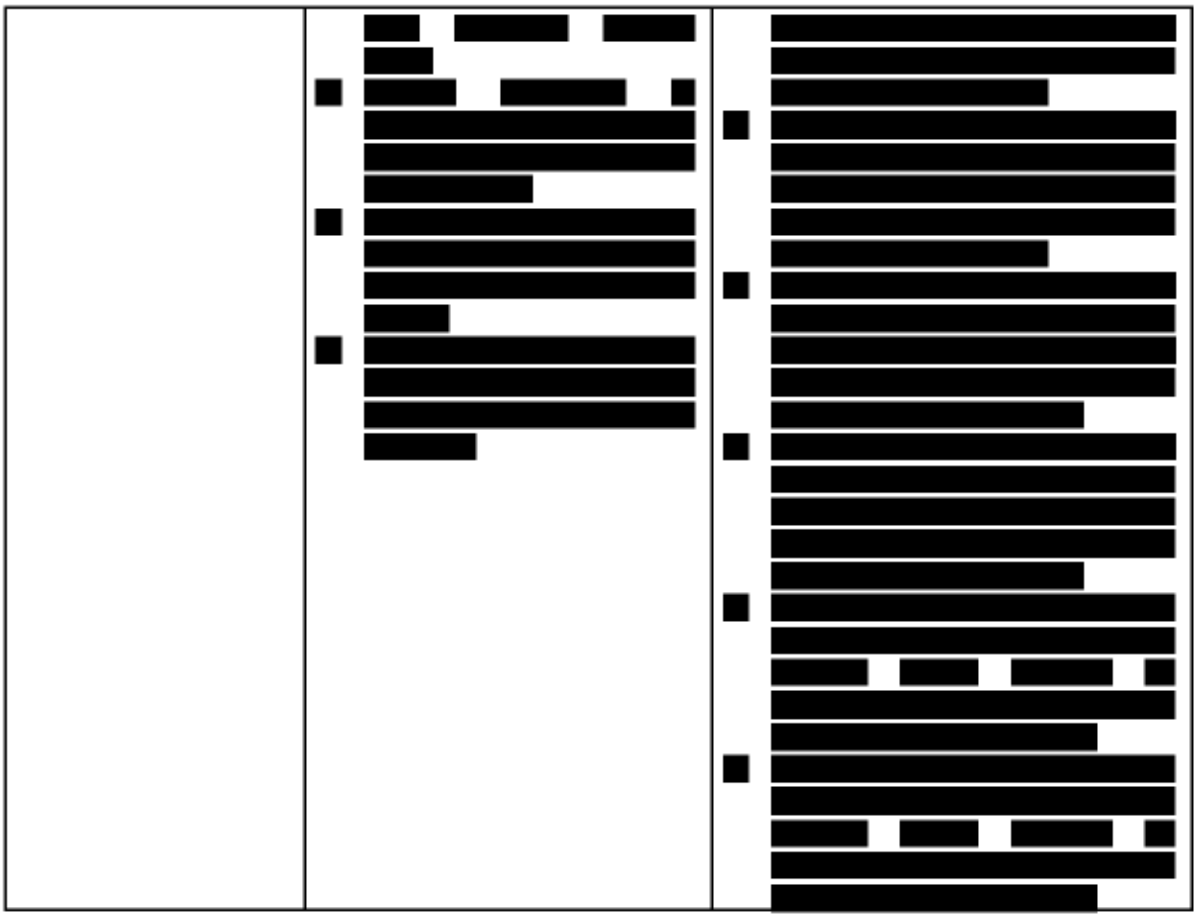
#### 4.7.7.1. Ineligibility Impact

There will be a detailed accounting of every subject to assess the impact of ITS exclusions. The nature of the loss being missing at random will be evaluated. Age, sex, race, and ethnicity will be assessed for any impact on exclusions due to ineligibility, as will an analysis based on incidental findings indicated by the Reading Center from either fundus photos or anterior segment images. These analyses focus on sensitivity and specificity.

#### 4.7.7.2. Missing Data Handling

The primary analysis for ITS and the secondary analysis will be based on all available data without imputation for missing data. [redacted]

[redacted]



Further analyses will be performed to evaluate the impact of missing diagnostic data as well as missing IDx-DR outcomes by comparing demographic data such as age, sex, race, and ethnicity. Enrolled subjects will be evaluated for bias by examination of cross-tabulations of baseline demographics for IDx-DR test outcome (present vs. absent) and ITS population with and without imputation for results of "Insufficient Exam Quality" will be displayed to assess bias. In addition, the impact of population exclusions will be assessed for both sensitivity and specificity. The impact will be assessed overall and separate analyses per cohort.

#### 4.7.7.3. Study Completion

Subjects who complete their study-directed photos in primary care and who complete gold standard ophthalmic imaging will be considered complete. The date of study completion is the date of the completed 4-W imaging.

Subjects who do not undergo their IDx-DR evaluation in primary care evaluation, and/or do not present 4-W photos that can be evaluated by the Reading Center will be discontinued from the study. The reason for the discontinuation will be recorded. The study will end when all subjects have completed the study-directed evaluations or have discontinued including withdrawals.

#### 4.7.7.4. Verification Bias

Verification bias is a type of measurement bias in which the results of a diagnostic test affect whether the gold standard procedure is used to verify the test result. This type of bias is also known as "work-up bias" or "referral bias". In clinical practice, verification bias is more likely to occur when a preliminary

diagnostic test is negative. Here, the IDx test is not invasive nor does it carry a higher risk. Verification bias is minimized in this setting since all subjects will have a gold standard test. Thus neither sensitivity nor specificity estimates may be biased. Excluded subjects will be compared to included subjects as part of this bias assessment. [redacted].

#### 4.7.7.5. Spectrum Bias

Spectrum bias will be investigated for enriched vs. enrolled mtmDR as well as across study sites because each site may have a different case mix. This is a form of a sampling bias. Subgroup sensitivities and specificity will also be investigated. In addition, the positive and negative likelihood ratios will also be checked. [redacted]

[redacted]

Spectrum bias will be investigated for each cohort as well as across study sites because each cohort or site may have a different case mix. This is a form of a sampling bias. [redacted]

#### 4.7.7.6. Handling of Withdrawals

Subjects may be withdrawn from the study by any of the following mechanisms:

1. Voluntary withdrawal of consent to participate by the subject at any time during the study.
2. Determination by the clinical investigator that it is in the best interest of the subject.
3. Determination by the sponsor or CRO it is in the best interest of the subject.
4. Subject ineligibility.
5. Inability for the subject to come in for 4-W photos.

All data available from enrollment to subject withdrawal will be collected unless the subject withdraws consent from the study. Data will be collected for subjects who discontinue participation but who do not withdraw consent.

[redacted].

#### 4.7.7.7. Multiple Comparisons / Multiplicity

[redacted] P-values will be computed for the primary and secondary effectiveness endpoints. The other effectiveness endpoints will be displayed with distributions or one-sided 97.5% confidence intervals.

#### 4.7.7.8. Examination of Subgroups

[redacted]

#### 4.7.7.9. Performance Data Stratified by Media Opacities Grading

Results will include the grading of media opacities and performance data stratified by the media opacity grading provided by FPRC.

## 5. [redacted]

[redacted]

## 6. REFERENCES

1. SAS Version 9 or later, Cary NC.

2. StatXact Version 10 or later, Cambridge MA.

3. Abramoff, M. D., J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, M. Lamard, D. C. Moga, G. Quellec and M. Niemeijer (2013). "Automated analysis of retinal images for detection of referable diabetic retinopathy." JAMA Ophthalmol 131(3): 351-357.