University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Phu, J;Bui, BV;Kalloniatis, M;Khuu, SK

Title:
How many subjects are needed for a visual field normative database? A comparison of ground truth and bootstrapped statistics

Date:
2018-03-01

Citation:
Phu, J., Bui, B. V., Kalloniatis, M. & Khuu, S. K. (2018). How many subjects are needed for a visual field normative database? A comparison of ground truth and bootstrapped statistics. Translational Vision Science and Technology, 7 (2), https://doi.org/10.1167/tvst.7.2.1.

Persistent Link:
https://hdl.handle.net/11343/256167

License:
CC BY-NC-ND

tvst

https://doi.org/10.1167/tvst.7.2.1

Article

How Many Subjects are Needed for a Visual Field Normative Database? A Comparison of Ground Truth and Bootstrapped Statistics

Jack Phu[1,2], Bang V. Bui[3], Michael Kalloniatis[1,2], and Sieu K. Khuu[2]

[1] Centre for Eye Health, University of New South Wales, Kensington, NSW, Australia
[2] School of Optometry and Vision Science, University of New South Wales, Kensington, NSW, Australia
[3] Department of Optometry and Vision Science, University of Melbourne, Parkville, VIC, Australia

Correspondence: Sieu K. Khuu, School of Optometry and Vision Science, Rupert Myers Building North Wing, Barker St, University of New South Wales, Kensington 2052, NSW, Australia. e-mail: s.khuu@unsw.edu.au

Received: 19 November 2017
Accepted: 9 January 2018
Published: 1 March 2018

Keywords: perimetry; Humphrey Visual Field Analyzer; glaucoma; bootstrap; Gaussian

Citation: Phu J, Bui BV, Kalloniatis M, Khuu SK. How many subjects are needed for a visual field normative database? A comparison of ground truth and bootstrapped statistics. Trans Vis Sci Tech. 2018;7(2):1, https://doi.org/10.1167/tvst.7.2.1
Copyright 2018 The Authors

Purpose: The number of subjects needed to establish the normative limits for visual field (VF) testing is not known. Using bootstrap resampling, we determined whether the ground truth mean, distribution limits, and standard deviation (SD) could be approximated using different set size (x) levels, in order to provide guidance for the number of healthy subjects required to obtain robust VF normative data.

Methods: We analyzed the 500 Humphrey Field Analyzer (HFA) SITA-Standard results of 116 healthy subjects and 100 HFA full threshold results of 100 psychophysically experienced healthy subjects. These VFs were resampled (bootstrapped) to determine mean sensitivity, distribution limits (5th and 95th percentiles), and SD for different 'x' and numbers of resamples. We also used the VF results of 122 glaucoma patients to determine the performance of ground truth and bootstrapped results in identifying and quantifying VF defects.

Results: An x of 150 (for SITA-Standard) and 60 (for full threshold) produced bootstrapped descriptive statistics that were no longer different to the original distribution limits and SD. Removing outliers produced similar results. Differences between original and bootstrapped limits in detecting glaucomatous defects were minimized at $x = 250$.

Conclusions: Ground truth statistics of VF sensitivities could be approximated using set sizes that are significantly smaller than the original cohort. Outlier removal facilitates the use of Gaussian statistics and does not significantly affect the distribution limits.

Translational Relevance: We provide guidance for choosing the cohort size for different levels of error when performing normative comparisons with glaucoma patients.

Introduction

Statistical analysis of visual field (VF) sensitivity data reported by commercially available instruments (e.g., the Humphrey Field Analyzer, HFA) are used to determine whether a patient's sensitivity is statistically "normal," such as in the case of patients with glaucoma.[1,2] Normative data are typically generated by recruiting healthy subjects, and the distribution limits of the normative data are empirically determined using conventional statistics.[3–10] Thus, identification of VF defects is contingent upon the characteristics of the normative distribution, and may differ across instruments and various research- and clinically based populations.

A number of questions remain regarding the development of normative databases for VF testing. First, how many subjects are required to produce resultant descriptive statistics that are robust estimates of the total population from which the sample is drawn? Second, in order to mitigate the need for further data collection, how do the distribution limits change with an increase in the underlying normative

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

sample, and whether this can be altered using statistical resampling methods (e.g., bootstrapping)? Third, what is the contribution of natural statistical outliers to the normative distribution?

These questions all ultimately relate to the number of subjects needed to generate normative data that reflects the underlying population of interest. A sample size that is too small may not be adequately representative of a healthy population.[11] A sample size that is too large may represent an unwise use of resources—time, personnel, and cost—where a smaller cohort may provide effectively the same result, and "big data" may confound analysis through the introduction of unexpected biases.[12]

In this paper, we test a number of hypotheses. We first determined if there was a smaller sample size of healthy subjects that is sufficient for determining normative distribution limits relative to the size of the complete cohort. This could then provide guidance for sample size selection for descriptive statistic parameters, such as mean and upper and lower distribution limits, and thus, identification of VF anomalies. There is debate regarding the characteristics of the underlying distribution of VF sensitivities, depending on whether naïve or experienced subjects are used, and whether outlier results are excluded from normative data, as these also impact upon the descriptive statistics of VF sensitivity.[10,13] Thus, we determined whether the presence of outliers affects the distribution of normal VF sensitivities. Finally, we considered whether the normative distribution limits obtained from both the original cohort and bootstrapped data sets returned similar numbers and depths of VF defects, 'events', in a cohort of patients with glaucoma.

Methods

The healthy subjects in this study consisted of two groups: a retrospective cohort and a prospective cohort. All healthy subjects had undergone extensive ophthalmic examination at the Centre for Eye Health, University of New South Wales, which included, but was not limited to the following: measurement of visual acuities, intraocular pressure using an applanation tonometer, slit-lamp examination, dilated fundus examination, standard automated perimetry (Humphrey Field Analyzer (HFA) 24-2 SITA-Standard; Carl Zeiss Meditec, Dublin, CA), color fundus photography, and optical coherence tomography. Inclusion criteria for all healthy subjects were as per previously reported and included, but was not limited to the following: best-corrected visual acuity of 20/20 for patients <55 years old and 20/25 for patients 55 years or older; intraocular pressures between 10 and 21 mm Hg; normal optic discs and fundus appearance; spherical equivalent refractive error less than 8.00 diopters (D) and astigmatism less than 3.00 D; and no personal medical history of diabetes.[4,13]

From these patients, VF data were extracted. Left eye results were converted to right eye results for consistency. Sensitivities (in dB) were obtained from the printout: 52 test locations and 74 test locations from locations within the 24-2 and 30-2 test grids, respectively, after excluding the fovea and the two locations adjacent to the physiologic blind spot.

Participants: Retrospective Cohort

The retrospective cohort consisted of 500 24-2 SITA-Standard VF results of 112 healthy subjects (mean age: 59.1 ± 8.1 years; 46 males) seen between January and July 2017 at Centre for Eye Health for ophthalmic examination. All VF results from each subject were extracted, such that each subject potentially contributed more than one VF result.

Studies have suggested pooling data from age-corrected cohorts spanning a range of ages to a uniform metric for one large normative database or to facilitate comparisons across a variety of age ranges.[3,4,10,13–15] We continue to use this method in the present study to pool data within our cohorts by converting all sensitivity results in a point-wise manner into a 50-year-old equivalent patient.

Although the SITA-Standard algorithm modulates resultant sensitivities based on the subject's age using the full threshold age-correction factors, the further effects of postprocessing are not precisely known, and so we determined whether or not there was a significant age-related effect on sensitivity in the present cohort.[14,16] Here, the null hypothesis was that there is no significant age-related effect, meaning that sensitivities may simply be grouped and pooled together. When sensitivities (in dB) were plotted as a function of age (in years), linear regression analysis found slopes that were significantly different to zero at all locations, indicating an age-related effect on sensitivities. Thus, to mitigate the effect of age, we age-corrected all sensitivities to that of a 50-year-old equivalent patient, as per previously published methods and using the slopes of change found using the above regression analysis (Supplementary Fig. S1).[3,10,13–15] These slopes were slightly higher (by 0.11 dB per decade, $t = 3.306$, df = 53, $P = 0.0017$) than

those found by Heijl et al.[10] using the full threshold procedure, and may reflect the extra modulation factor used in SITA.

Participants: Prospective Cohort

Although the retrospective cohort consisted of subjects with normal ophthalmic examination, these were still perimetrically naïve subjects, and hence there is the potential for variability attributable to learning effects.[17] Thus, we also examined the sensitivities of a prospectively recruited cohort of 100 healthy subjects (mean age: 36.5 ± 15.7 years; 46 males) who had undergone VF testing on the HFA using the full threshold paradigm and the 30-2 test grid. Each subject contributed only one test result, that is, a single average sensitivity at each of the 75 test locations (including the fovea and excluding the 2 points near the blind spot) within the 30-2 test grid) from one eye, following extensive prior perimetric testing experience. These results were also age-corrected into a 50-year-old equivalent subject for pooling and analysis.[13]

Modelling Normative Cohort Size Using Bootstrapping Resampling

A nonparametric bootstrap was used to resample with replacement sensitivity data from retrospective and prospective cohorts (Fig. 1).[18] From the original cohort, we resampled a subset of the data (set size x). Each resample could consist of the same subjects, as replacement was performed. This process was repeated k (number of resamples) times.

To determine the effect of cohort size on "normative distribution" parameters, we systematically varied the number of resample sensitivities from the total cohort (we termed this the set size, x). For example, $x = 6$ indicates a set of six sensitivity values resampled from the original total cohort (i.e., either the retrospective or prospective cohorts). Using the resampled sensitivities, we determined the mean (i.e., the central tendency), 95th percentile and 5th percentile (i.e., the distribution limits), and the standard deviation (SD). We tested a range of values for x (for the retrospective cohort: $x = 6$, 12, 24, 36, 48, 60, 75, 100, 150, 200, 250, 300, 350, 400, 450, and 500; for the prospective cohort: $x = 6$, 12, 18, 24, 30, 36, 48, 60, 72, 85, and 100), which was capped at the total number of subjects in the cohort. To determine the confidence limits for the descriptive parameters from these set sizes, we tested two levels of k (number of resamples), the number of resamples from the total

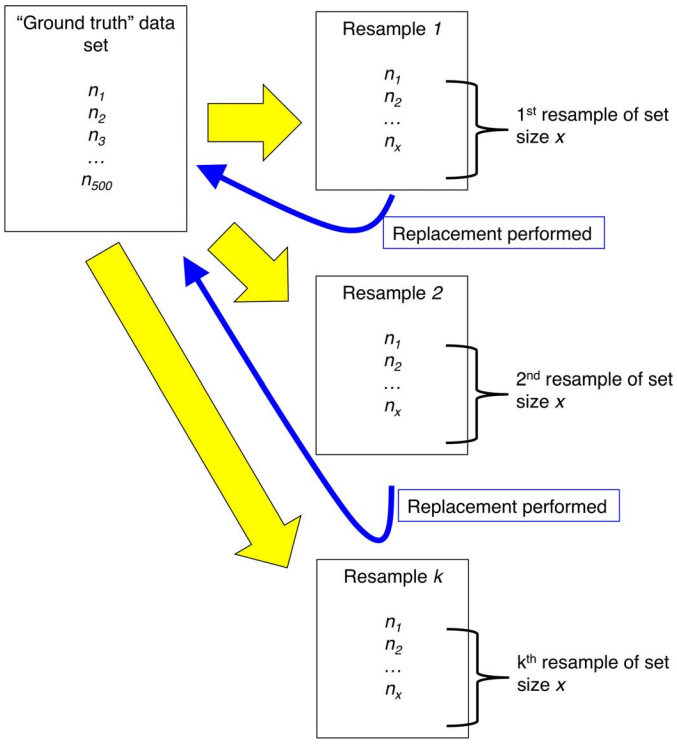


Figure 1. A flowchart describing the bootstrapping process used in the present study. The "ground truth" data set consisted of all of the subjects for each of the retrospective ($n = 500$) or prospective ($n = 100$) cohorts. A resample of set size x was performed, whereby a subset of data were extracted (*yellow block arrows*). The subjects were "replaced," such that subsequent resamples could include subjects in previous resamples (*blue arrows*). This resampling process was repeated k number of times.

cohort, which were $k = 100$ and $k = 200$. We also tested whether or not the value of k affected the resultant descriptive statistics independently of the set size. Thus, for set sizes of $x = 6$, 30, 60, and 500, we tested different levels of k: 1, 4, 8, 12, 16, 20, 24, 28, 36, 48, 60, 72, 84, 96, 120, 150, 200, 250, 300, 400, 500, and 750. This bootstrap procedure was performed on a custom written macro program using Visual Basic Editor in Microsoft Excel 2010 (Microsoft Corporation, Redmond, WA).

The differences in the mean and distribution limits between the true sample values (i.e., parameters from the original cohort data, which we refer to as the 'ground truth' parameters) and the bootstrapped values were determined for each combination of x and k. The difference (in dB) was plotted as a function of x to determine the limit at which there ceased to be a significant change (i.e., when the difference between ground truth and bootstrapped parameters was minimized). A positive value in the difference from the ground truth and bootstrapped parameters

indicates that the ground truth was higher than the bootstrap, while the converse was true for a negative difference.

Outlier Removal

Outliers were identified and removed using a combination of robust nonlinear regression and outlier removal (with Q = 10%; GraphPad Prism 7; GraphPad Inc., La Jolla, CA).[19] In this method of robust nonlinear regression, it is assumed that variation around the curve follows a Lorentzian distribution, rather than Gaussian, which has wide tails and is less affected by outliers in the fit. Unlike least squares fitting, which quantifies the variance around the curve using $S_{y.x}$ (the SD of the residuals), robust nonlinear regression determines the 68.27th percentile of the absolute values of the residuals (1 SD from the mean in a Gaussian distribution), and this is called the robust SD of the residuals (RSDR). Each residual is divided by the RSDR, and this ratio approximates a *t* distribution, from which a *P* value can be determined. Outliers that are greater than the value Q (controlling the false discovery rate) are then removed.

We compared the number of outliers removed by Q = 0.1% ($n = 14$, 0.05%), 1% ($n = 24$, 0.09%), and 10% ($n = 119$, 0.46%), and examined the sensitivities identified as outliers. As expected, a Q of 10% removed the greatest number of outliers at all locations; over 52 test locations, this equated to approximately 1.8 more values removed per location compared with the 1% level. Central tendency results were similar, but the variance was reduced when using Q = 10%. The Q = 10% condition removed points that there at least 3.3 SD away from the mean, equating to a *P* value of 0.05%, which is the lowest level of significance flagged on the HFA total deviation and pattern deviation maps. Thus, in order to obtain data with the most likely outliers removed, we continue to report results using Q = 10%.

Outlier removal was performed on both the retrospective and prospective cohorts to obtain a "cleaned" data set. The above procedures for obtaining the ground truth and bootstrapped means and distribution limits were applied to the cleaned data set, and were again compared across different levels of *x* and *k*. Furthermore, the mean and distribution limits were also compared between the complete data set (i.e., including those points deemed to be outliers) and the cleaned data set. Here, we tested the hypothesis that the cohort inclusive of all points and the cohort trimmer of outliers return different results in descriptive statistics and also have different levels of *k* at which no change from the ground truth parameter is obtained.

Empirical Testing of the Bootstrapped Results in Patients With Glaucoma

Finally, we empirically tested the capacity of the bootstrapped-derived parameters for various normative cohort sizes to detect SITA-Standard VFs deficits in a cohort of patients with open-angle glaucoma seen and/or treated at the Centre for Eye Health. Inclusion criteria for the glaucoma patients have been previously detailed, but in short, required characteristic glaucomatous optic nerve head changes (e.g., thinning of the neuroretinal rim, notching, increased vertical cup-disc ratio), retinal nerve fiber layer defects, open angles on gonioscopy (angle closure glaucoma patients were excluded), and with or without VF defects.[3,4] Intraocular pressures were not used as part of the diagnostic criteria. Inclusion criteria for analysis of their VF results were as per the healthy cohort described above.

The 5th percentile of the retrospective normative cohort was used as the lower limit of normality, (i.e., 'events').[3–5] This offered a practical method for assessing the normative cohort sizes required to result in the same level of performance in terms of defect identification as when using the ground truth parameters. We used 390 SITA-Standard VF results of 112 patients with open-angle glaucoma (mean age: 62.0 ± 12.6 years; 189 right eyes; 74 males; average mean deviation: −3.56 ± 3.89 dB), and determined the number and depth of 'events' (difference in sensitivity from the mean in dB) flagged when using the 5th percentiles obtained from the ground truth and from the different levels of *k*. We determined the level of *k* at which there was no longer a significant change from the ground truth value. This analysis was performed when using both the complete data set and when outliers were removed.

Statistical Analysis

Descriptive statistics and ANOVAs with Dunn's multiple comparisons were used to compare the mean sensitivities and distribution limits obtained from each condition. Although sensitivity values are logarithmic units (dB), we continued to report arithmetic means and SD, as there were no significant differences with geometric mean calculations, and this would be in line with use of mean deviation and pattern SD statistics reported in the HFA (Supple-

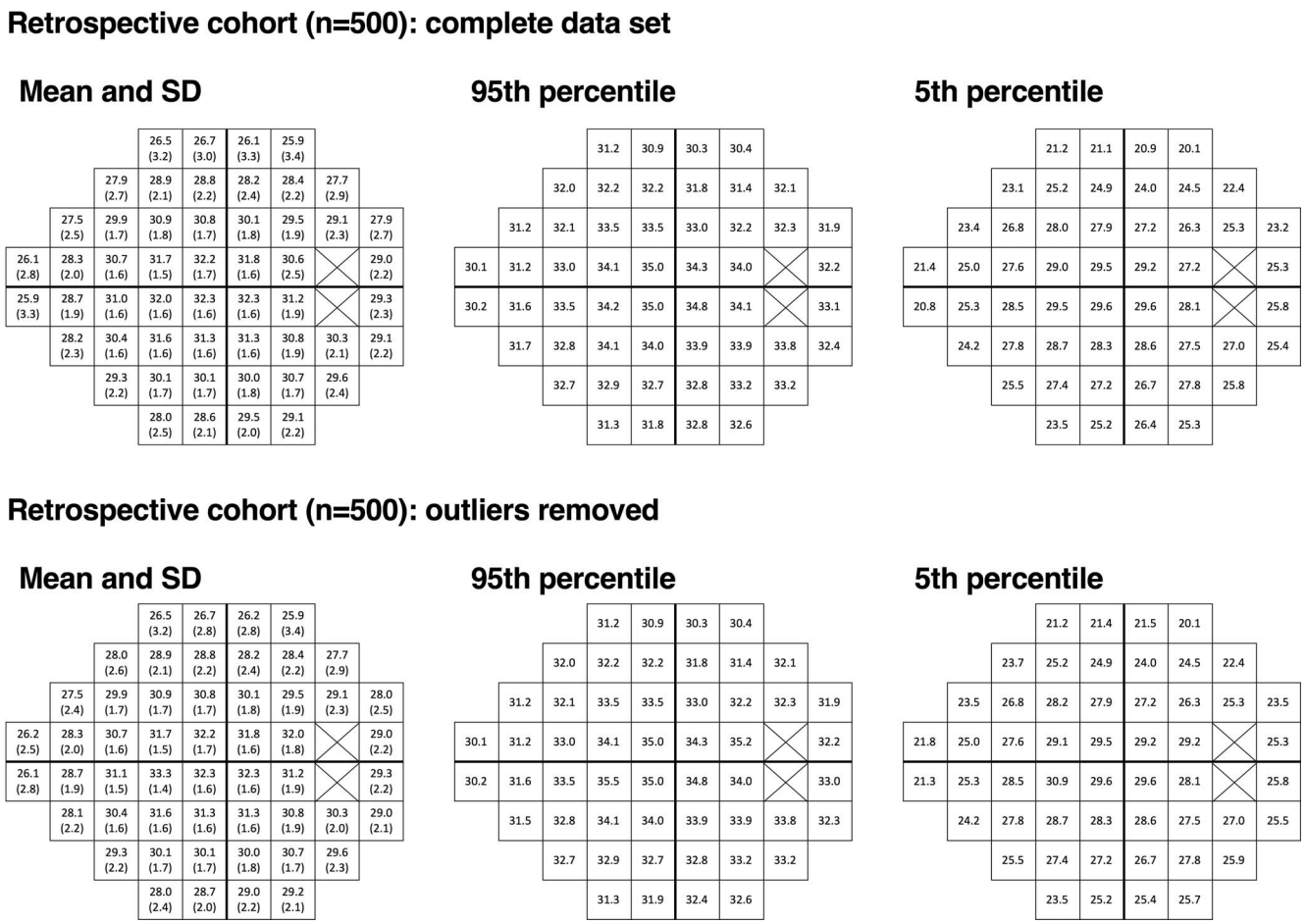


Figure 2. Mean and SD, 95th percentile and 5th percentile sensitivity values (in dB) for the retrospective cohort ($n = 500$ SITA-Standard VF results) when the complete data set was used (A) and when outliers were removed (B) for locations within the HFA 24-2 test grid. The *crossed cells* denote the two test locations near the physiological blind spot, excluded from analysis.

mentary Fig. S2). A $P < 0.05$ was considered significant. Difference between the values obtained using bootstrapping and the ground truth parameters were plotted as a function of set size (x) or number of resamples (k). The asymptotic point was determined by one-way ANOVA and multiple comparisons. When the multiple comparisons showed no more significant differences across adjacent and successive conditions (e.g., $x = 60$, $x = 72$, $x = 96$ are considered successive conditions), then the asymptotic point was reached.

The method of checking the approximate asymptotic point was also used to determine the set size at which no further change in VF defect detection in glaucoma patients. As mentioned, we compared the ability of the complete data set and the data set when outliers were removed to assess glaucomatous VF defects. To assess this difference, we determined the number of defects found using different percentile cut-off values for normality (lower 5th percentile to highest, i.e., 100th percentile value), that is, receiver operator characteristic (ROC) curves. One of the typically used criteria for a glaucomatous VF defect is three or more contiguous points depressed at the $P < 0.05$ level, at least one of which is reduced below the $P < 0.01$ level. However, as we were assessing different levels of percentile cut offs as a surrogate specificity value (e.g., the lower 5th percentile of the healthy cohort is effectively a 95% true negative rate, or specificity), for the purpose of this analysis, we regarded three or more points of sensitivity reduction below the cut-off as a criterion for a glaucomatous VF defect. This allowed determination of the true positive rate (i.e., test "sensitivity," which was defined as the number of patients meeting the VF cut-off criterion of 3 or more points divided by the total number of glaucoma patients) for different true

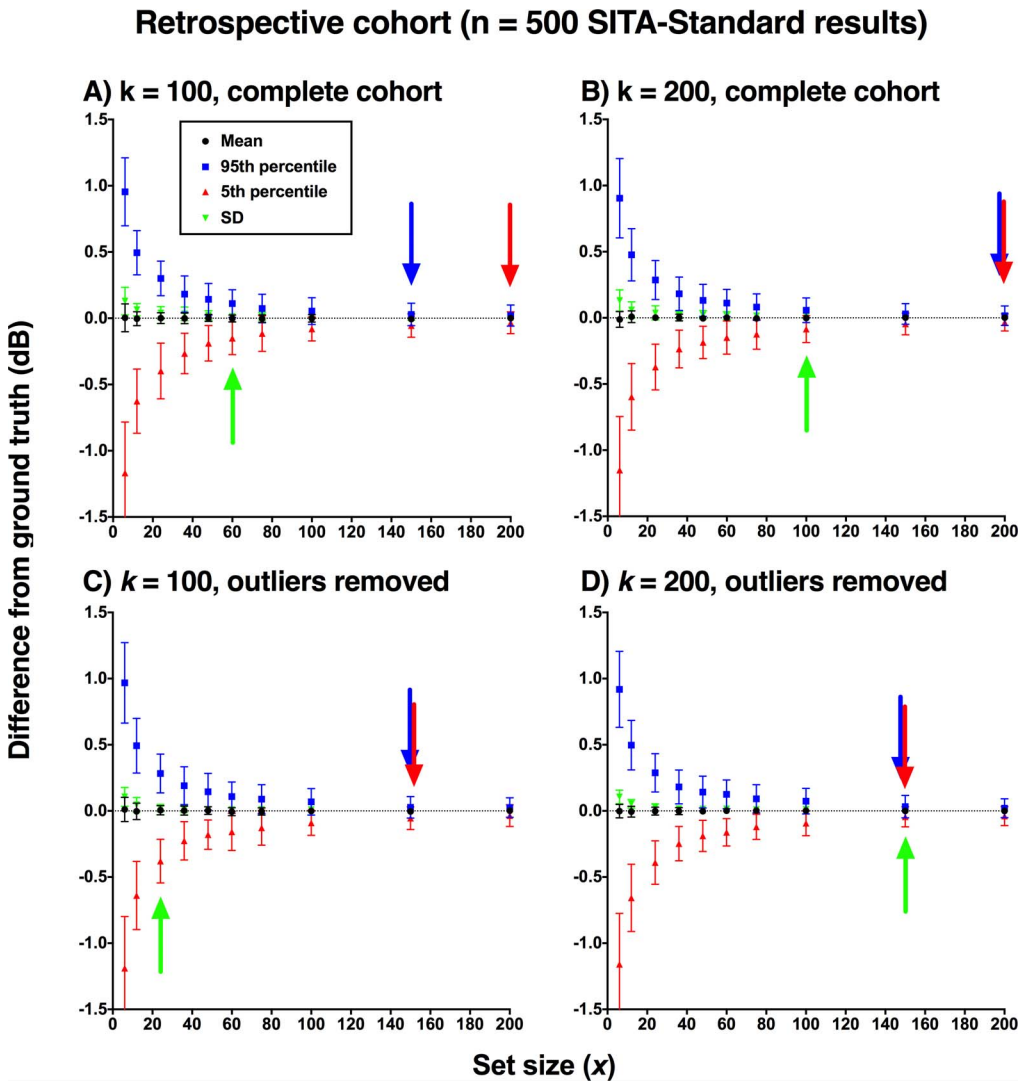


Figure 3. Difference from ground truth (dB) as a function of set size (*x*) for the retrospective cohort (SITA-Standard VF results, $n = 500$). Results when the complete cohort was used ([A] $k = 100$ and [B] $k = 200$ conditions) and when outliers were removed ([C] $k = 100$ and [D] $k = 200$ conditions) are shown separately. Mean (*black*), 95th percentile (*blue*), 5th percentile (*red*), and SD (*green*) are plotted for each set size condition. The *colored arrows* (corresponding to each statistic) indicate the approximate asymptotic point at which there is no longer a change in difference from the ground truth, as per Table 1 (mean not shown for clarity). *Error bars*: 1 SD. Although *x* was tested up to 500, values beyond $x = 250$ were excluded for clarity, as the results were identical as that of $x = 500$.

negative rates (i.e., "specificity," set at each percentile cut-off level), and thus, ROC curves.

Results

Retrospective Cohort: SITA-Standard VF Results

Mean, 95th percentiles, 5th percentiles, and SD values for each location within the 24-2 VF are shown in Figure 2. When outliers were removed, there was no significant difference across mean ($P = 0.2685$), 95th percentile ($P = 0.2971$), 5th percentile ($P = 0.2017$), or SD ($P = 0.9282$) compared with when the complete cohort was used (Fig. 2B).

The difference between the ground truth and bootstrapped mean, 95th and 5th percentiles, and SD values were determined at each test location. Although more centrally located points showed, on average, smaller differences between ground truth and bootstrapped parameters when compared with more peripherally located points, there did not appear to be a significant, systematic effect of eccentricity on the difference (one-way ANOVA, $F_{3.14,47.1} = 1.12$, $P =$

Table 1. Levels of x at Which Multiple Comparisons Were No Longer Significantly Different ($P < 0.05$) Across Adjacent Levels for the Retrospective Cohort and Prospective Cohort for the Bootstrapped 95th Percentile, 5th Percentile, and SD Parameters.

	95th Percentile	5th Percentile	SD
Retrospective cohort ($n = 500$)			
$k = 100$, complete cohort	150	200	60
$k = 200$, complete cohort	200	200	100
$k = 100$, outliers removed	150	150	24
$k = 200$, outliers removed	150	150	150
Prospective cohort ($n = 100$)			
$k = 100$, complete cohort	48	60	30
$k = 200$, complete cohort	60	60	30
$k = 100$, outliers removed	60	60	30
$k = 200$, outliers removed	72	72	60

The bootstrapped mean was not significantly different to the best-fit value across all conditions.

0.3521). Therefore, we grouped the results of all test locations together for further analysis for different levels of x and k.

For the $k = 100$ condition, one-way ANOVA showed no significant effect of x on the difference between ground truth and bootstrapped means ($P = 0.3521$), but showed a significant difference in the 95th percentile ($P < 0.0001$), 5th percentile ($P = 0.0001$), and SD ($P = 0.0049$) parameters (Fig. 3A). A similar tendency was found for $k = 200$ (Fig. 3B). When outliers were removed, $k = 100$ and $k = 200$ showed the same effects as when the complete cohort was used (Figs. 3C, 3D).

By inspection, the difference from the ground truth parameter reached an asymptote at 0 dB for all parameters at approximately $x = 150$. This was also examined using one-way ANOVAs and multiple comparisons, whereby the level of x at which there was no significant difference across all further adjacent levels (e.g., $x = 60$ to $x = 72$, and so forth having $P > 0.05$) was taken as the asymptotic point. These results were generally consistent with this estimation (Table 1).

Prospective Cohort: Full Threshold VF Results

In contrast to the retrospective cohort, there were significant differences in mean (mean difference: 0.04 ± 0.08, $P < 0.0001$), 5th percentile (mean difference: 0.13 ± 0.21, $P < 0.0001$), and SD (mean difference: −0.14 ± 0.19, $P < 0.0001$) values when comparing the complete cohort and the results when outlier sensitivities were removed (Fig. 4). The differences in the 95th percentile value were borderline in terms of statistical significance (mean difference −0.03 ± 0.11, $P = 0.0504$). Despite the statistically significant differences, these were unlikely to be of any clinical significance, as the mean differences were well within instrument test-retest variability.[20]

As for the retrospective cohort, we determined the normative cohort size at which there was no longer a significant difference between bootstrapped and ground truth parameters. Similar to the retrospective cohort, there was a significant effect of the set size x on the difference between 95th percentile ($P < 0.0001$), 5th percentile ($P < 0.0001$), and SD ($P < 0.0001$) values, but not on the mean ($P = 0.1400$) for $k = 100$ (Fig. 5A). A similar tendency was found for $k = 200$, and when outliers were removed (Figs. 5B–D).

By inspection, the difference between ground truth and bootstrapped conditions approached 0 dB with a cohort size of approximately $x = 60$ across all conditions. One-way ANOVA and multiple comparisons also generally agreed with this estimation (Table 1).

The Effect of the Number of Times Resampled (k)

We determined whether or not the number of times resampled affected the differences between ground truth and bootstrapped values. The differences between ground truth and bootstrapped values were plotted as a function of the number of resamples (k) for four set size conditions: $x = 6$, 30, 60, and 500 using data from the retrospective cohort (Fig. 6). There was no effect of k on mean (mean P value = 0.6086), 95th percentile (mean P value = 0.4488), 5th percentile (mean P value = 0.6697), or SD (mean P value = 0.6296), expect for SD at the $x = 30$ condition ($P = 0.0328$). These results suggest that set size x is more important than the number of resamples k when attempting to minimize the difference to the ground truth statistic.

The Effect of Using a Smaller Sized Initial Cohort

So far, the resamples have been drawn from the complete data set. With the retrospective and the prospective cohorts, only a small proportion of the

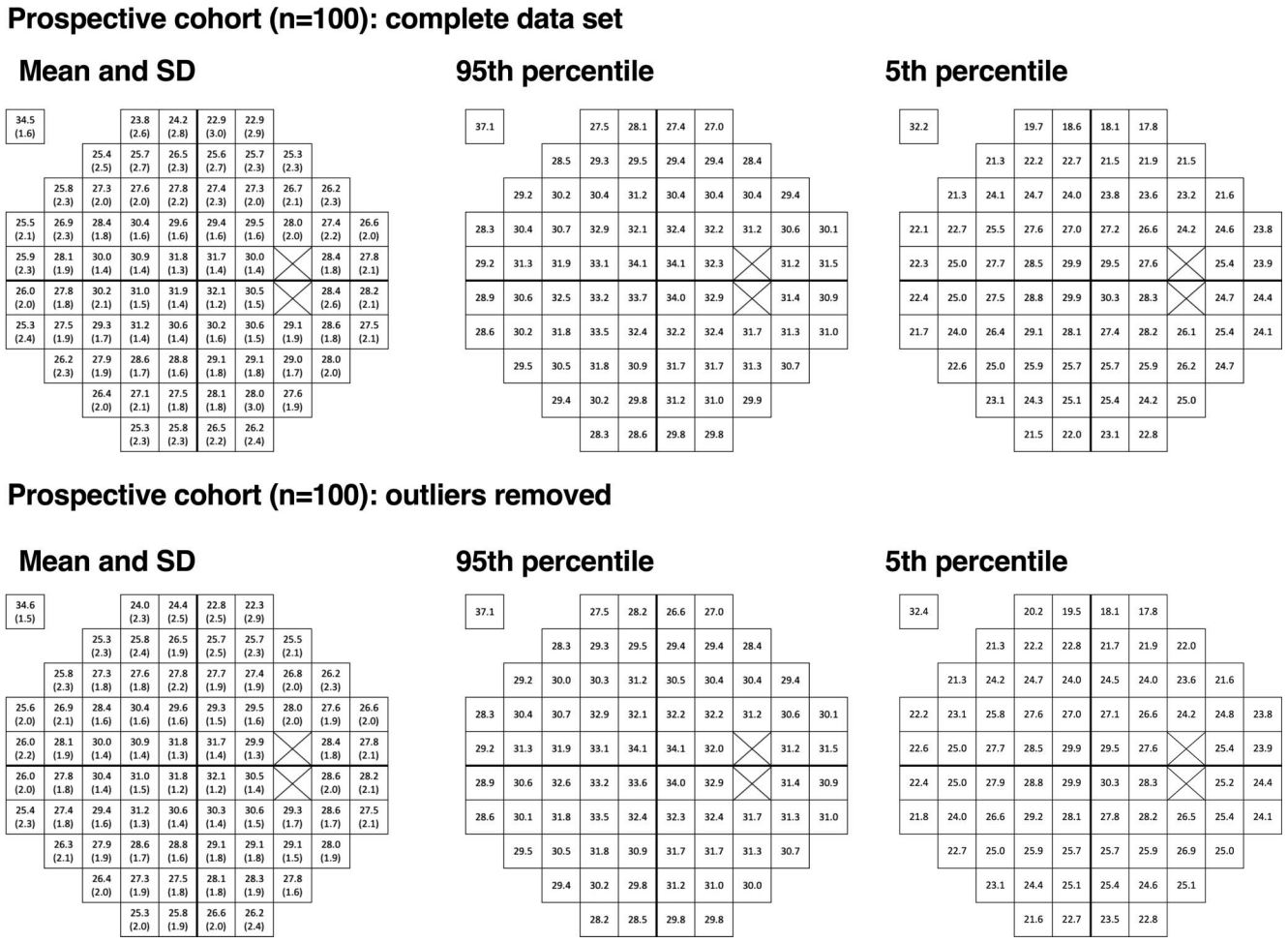


Figure 4. Mean and SD, 95th and 5th percentile sensitivity values (in dB) for the prospective cohort ($n = 100$ full threshold VF results) when the complete data set was used (A) and when outliers were removed (B) for locations within the HFA 30-2 test grid. The two test locations near the physiological blind stop have been crossed out, and the fovea has been offset to the upper left for clarity.

total data set was required to provide a similar estimate of mean and distribution limits: approximately 40% and 60% for retrospective and prospective, respectively. We therefore tested whether the size of the initial sample changed the number of subjects within each resample required to arrive at the same mean and distribution limits. If a smaller initial sample size results in a proportional reduction in the size of resample, it may suggest that a correspondingly large sample would be required to determine the characteristics of the entire general population.

To test this hypothesis, we randomly selected 300 and 400 subjects from the original retrospective cohort to serve as new, complete cohorts. We performed the same analysis as above using set sizes ranging from $x = 6$ to 300 and 400, respectively, and with $k = 100$. We then compared the level of x at which the multiple comparisons did not significantly change further, as described in the above analysis. For both $n = 300$ and $n = 400$, we found a level of x that was similar to when $n = 500$ (Table 1) was used: $x = 150$ for 95th percentile, $x = 150$ for 5th percentile, and $x = 60$ for SD (Fig. 7). Thus, this suggests that differences are stable at this level of n, rather than being proportional to the base "population" size used.

The Application of Bootstrapped Results to Patients With Glaucoma: Comparison With Ground Truth Performance for Defect Detection

Finally, we wanted to test the effect of bootstrapped VF results on the detection of glaucomatous defects in a clinical cohort of patients. We determined the number of 'events' and their depth against the limits determined from the original retrospective data with and without the inclusion of outliers. Inclusion

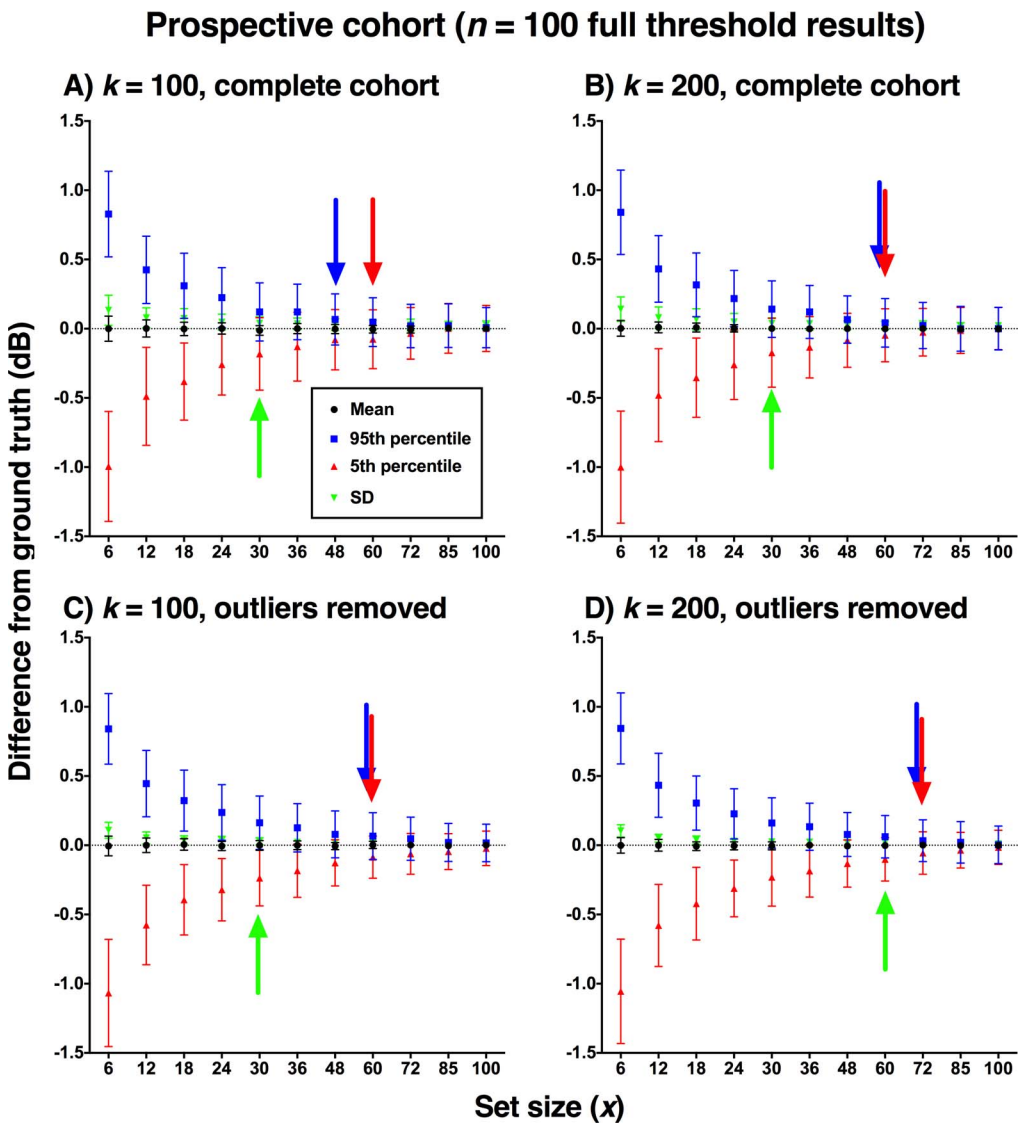


Figure 5. Difference from ground truth (dB) as a function of set size (x) for the prospective cohort (full threshold VF results). Results when the complete cohort was used ([A] $k = 100$ and [B] $k = 200$ conditions) and when outliers were removed ([C] $k = 100$ and [D] $k = 200$ conditions) are shown separately. Mean (*black*), 95th percentile (*blue*), 5th percentile (*red*), and SD (*green*) are plotted for each set size condition. The *colored arrows* (corresponding to each statistic) indicate the approximate asymptotic point at which there is no longer a change in difference from the ground truth, as per Table 1 (mean not shown for clarity). *Error bars*: 1 SD.

of the complete data set resulted in on average 0.51 (95% confidence interval 0.44–0.60) fewer 'events' detected compared with when outliers were excluded.

Next we assessed 'events' and depth against limits determined from the bootstrapped cohorts of different sizes (at $k = 200$). One-way ANOVA showed a significant effect of set size x ($P < 0.0001$ across all conditions) on the number of 'events' detected and depth of defect (Fig. 8). When using the 5th percentile from the original retrospective data as the 'ground truth', smaller set sizes tended to overestimate the number of 'events', and underestimated their depth, corresponding to higher 5th percentile and lower mean values (as per Fig. 3). Multiple comparisons showed significant differences across all bootstrapped conditions compared with the ground truth for 'events' detected when the complete data set including outliers was used ($P < 0.0001$), but showed an asymptote at $x = 250$ when outliers were removed. There were similar asymptotes at $x = 250$ and $x = 300$ for the depth of defect for the complete data set and when outliers were removed, respectively. When setting the criterion to be a difference of one 'event' flagged (as 'events' are reported in integer values), $x =$

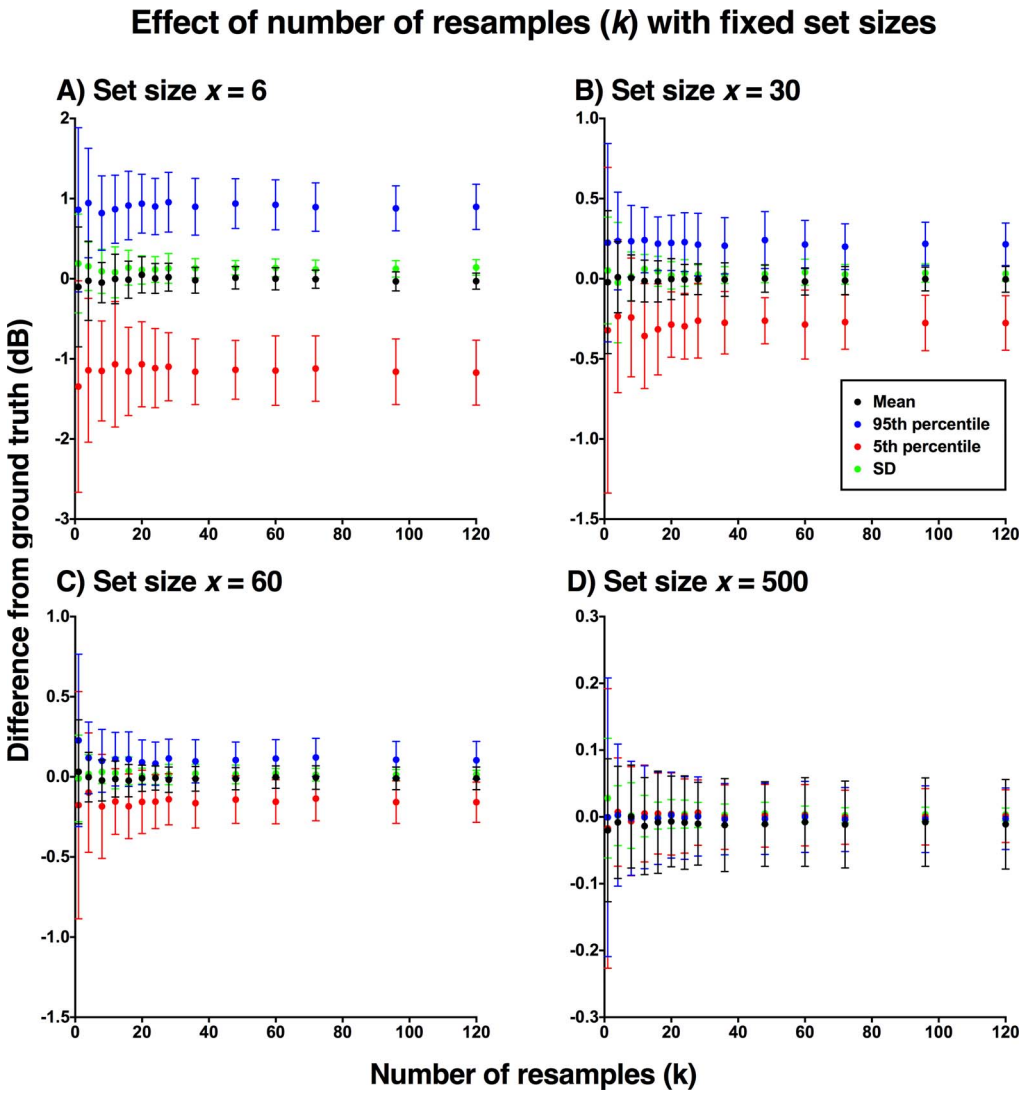


Figure 6. Difference from ground truth parameters (dB) as a function of number of resamples (*k*) for the retrospective cohort (SITA-Standard VF results) for predetermined set sizes of $x = 6$ (A), 30 (B), 60 (C), and 500 (D) when all points were included (no outliers removed). Mean (*black*), 95th percentile (*blue*), 5th percentile (*red*), and SD (*green*) are plotted for each set size condition. *Error bars*: 1 SD. Although *k* was tested up to 750, values beyond $k = 120$ were excluded for clarity, as the results were not different to that of $k = 750$.

60 (complete data set) and $x = 48$ (outliers removed) were the minimum set sizes for which the difference between ground truth and bootstrapped values was less than one 'event'.

As mentioned in the Methods, ROC curves were generated by modifying the normative percentile cut off for labeling a VF defect to obtain surrogate specificity values (e.g., a cut off of the lowest 10th percentile represents a specificity of 90%) to then determine sensitivity (true positive rate). Comparison of the area under the ROC (AUROC) curves found using the complete data set and when outliers were removed showed similar results ($F_{1,160} = 0.0258$, $P = 0.8727$; Fig. 9). As expected, the AUROC was slightly greater when using a smaller set size, $x = 6$, in comparison to the other conditions, as the resultant percentile cut-off values were higher under conditions of low specificity. However, there was no significant difference between ground truth and set sizes $x = 6$, 200 and 500 ($F_{3,160} = 0.1307$, $P = 0.9417$).

Discussion

In the present study, we examined the VF results of two cohorts of healthy subjects to determine whether or not there was a limiting number of resamples that could generate a value for the mean, SD, and upper

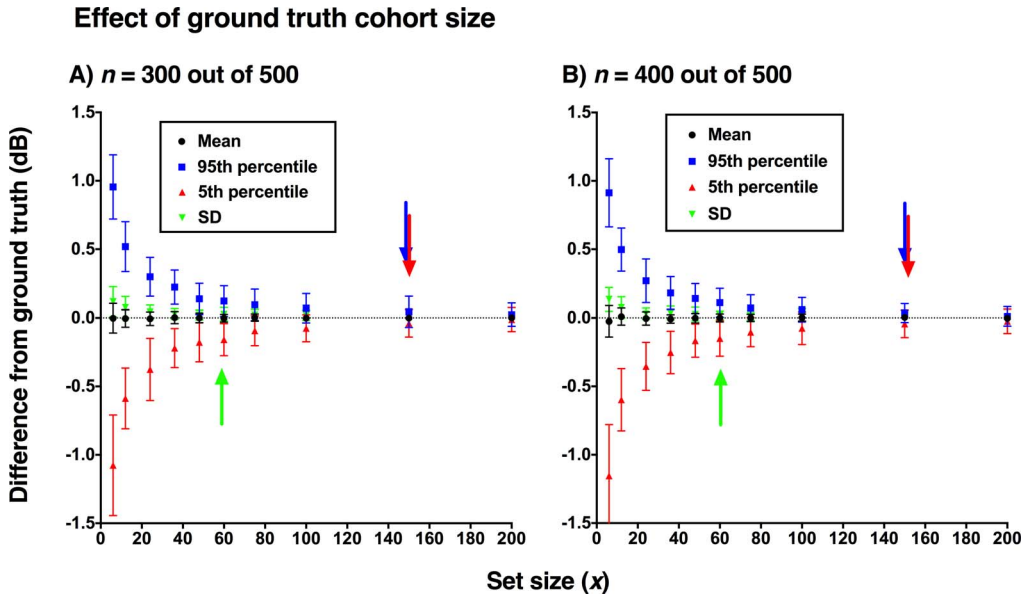


Figure 7. Difference from ground truth (dB) as a function of set size (x) for the retrospective cohort when using a smaller total population ($n = 300$ and $n = 400$, randomly sampled from the original cohort of $n = 500$), plotted as per Figure 3. Mean (*black*), 95th percentile (*blue*), 5th percentile (*red*), and SD (*green*) are plotted for each set size condition. *Error bars*: 1 SD.

and lower percentiles that was comparable to the original underlying data set, therein providing potential guidance for cohort sizes needed for the generation of robust VF normative databases. Here, the null hypothesis was that parameters from resampled smaller data sets would not be comparable to the original cohort's. However, we found that the difference from the complete cohort was no longer significant once a certain set size was reached (Table 1), suggesting that gains in recruiting healthy subjects beyond this number would be minimal.

Bootstrapping VF Results: Set Size or the Number of Resamples?

Bootstrapping was to test the precision of sample statistics from the original cohort and, by doing so, testing the validity of the model through generating random subsets.[21] This technique has been used widely in VF research.[22–25] There are two main variables in our approach: the set size (x) and the number of times a set is drawn (i.e., resampled [k]). Therefore, the number of resamples of k could potentially affect the bootstrapped statistics. For example, a large number of resamples of k could mask the differences found with low levels of x. However, we found no such tendency, and the results were similar across all levels of k, suggesting that x produces the differences seen between bootstrapped and ground truth values.

How Many Subjects are Required for Generation of Normative VF Data?

A previous study has provided guidance for sample sizes involving the demonstration of novel effects within a system without necessarily quantifying the parameter, and these may only require a small sample size.[26] For determination of the magnitude of difference between two groups, such as a treated and control group, conventional statistics and power analyses are available, but these do not necessarily provide guidance as to the number of samples required to generate normative data to serve as a comparison group.

Normative distribution limits are often generated empirically. In the case of VF studies the 5th percentile is often used as the cut off for an 'event', but therein lies a problem: in a normal cohort of 20 subjects, the 5th percentile represents only one individual's result.[3–5] The addition of 20 subjects at a time would only add one more subject with which to define the 5th percentile. The results of the present study suggest that beyond approximately 150 to 200 subjects for SITA-Standard and 60 subjects for full-threshold VF results, estimates of the distribution limits are similar to that of the ground truth, and the addition of more subjects do not provide further information. This number may vary slightly depending on the composition of the cohort (e.g., perimetri-

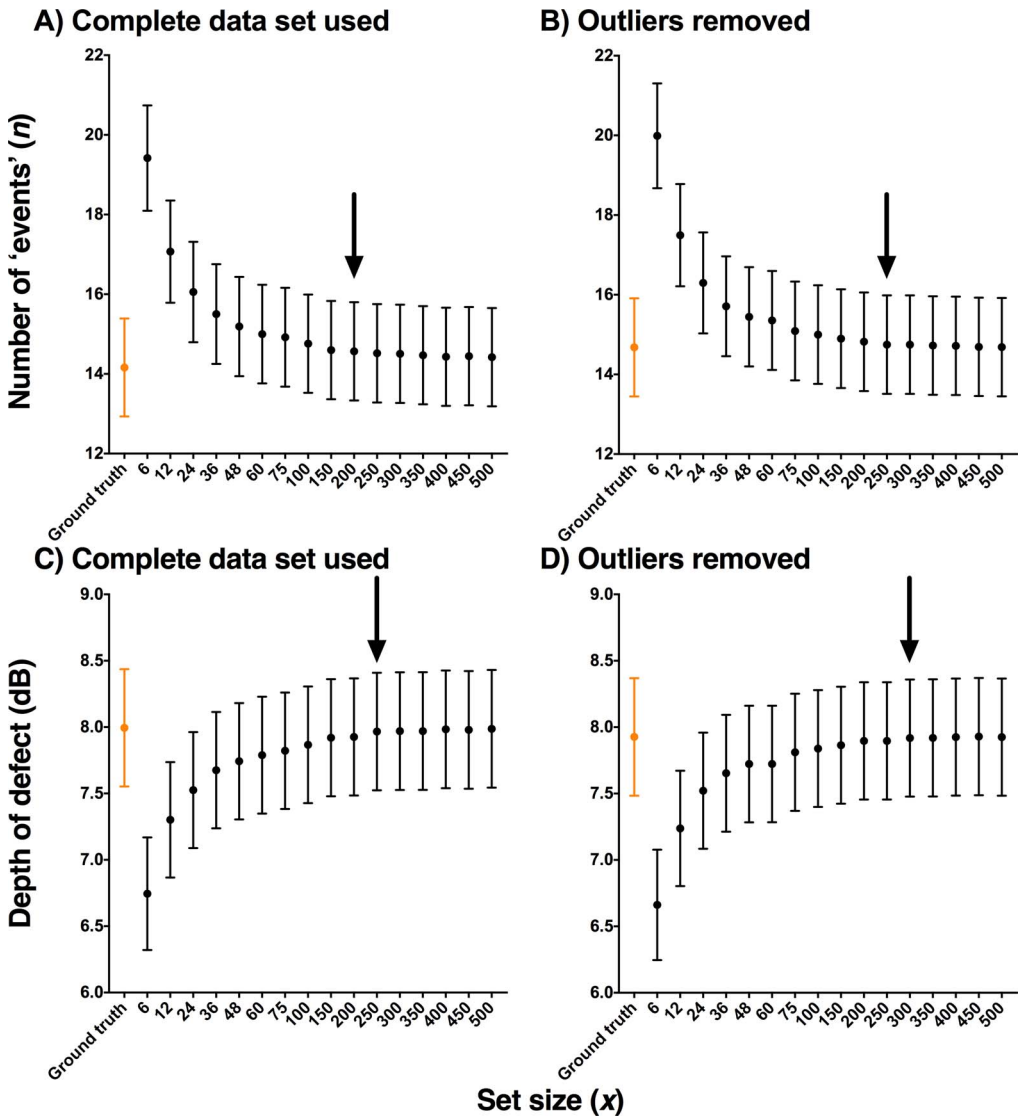


Figure 8. Number of 'events' (*n*, [A, B]) and depth of defect (dB, [C, D]) identified in the cohort of glaucoma patients as a function of set size ($x = 6$–500) when using the complete data set (*left*) and when outliers were removed from the normative data (*right*) for $k = 200$. Each datum point represents the average across all glaucoma patients for each level of *x*, and the *error bars* indicate the 95% confidence interval. The *orange point* indicates the values obtained when using the ground truth. The *black downward arrow* indicates the approximately asymptotic location for each condition at which there was no longer a significant change in number of 'events' detected of depth of defect between the bootstrapped value and the ground truth data.

cally naïve or experienced observers) but show that only a smaller group of subjects may be required.

Outlier Removal: Implications for Defect Detection

Aside from the initial screening process, one other method for tackling the effects of unreliable results is the removal of outliers. Heijl et al.[10] reported significant skew in the sensitivity data of healthy subjects; in comparison, recent work by Phu et al.,[13] having considered outliers, instead reported normally distributed sensitivity values. It is expected that outlier removal would, in particular, affect the lower limit of the normative distribution, but the question is by how much does this affect the result of normative comparison with a group of patients with disease? Our results show that although there is a statistically significant difference in the number of 'events' found when using normative data with and without outliers included, this difference was unlikely to be clinically significant as it was smaller than the lowest integer value. This was also consistent with the minimal

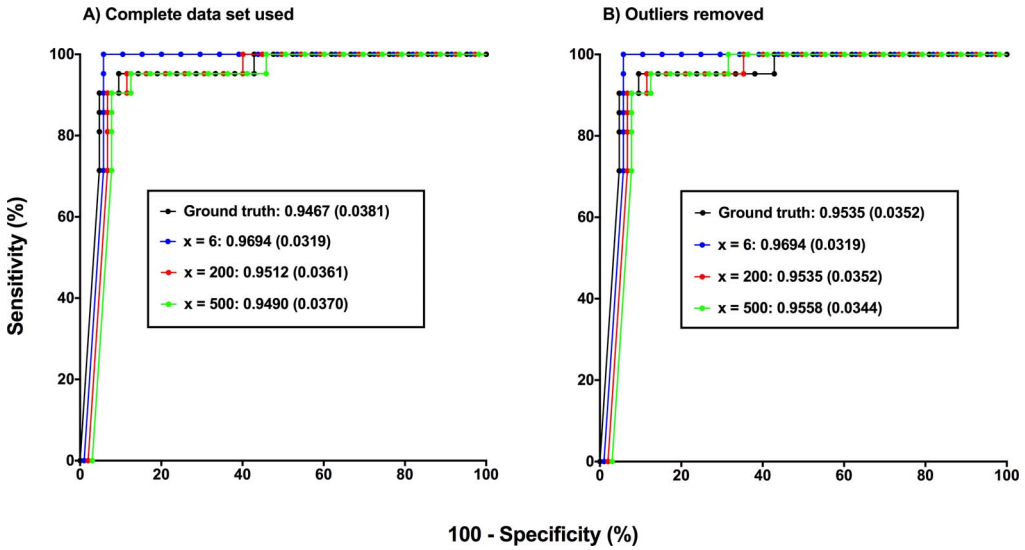


Figure 9. ROC curves plotting sensitivity (%) as a function of 100 − specificity (%) for complete data set was used (A) and when outliers were removed (B). Four conditions are plotted: ground truth (*black*) and bootstrapped data sets with set sizes $x = 6$ (*blue*), 200 (*red*), and 500 (*green*), each offset slightly for clarity. AUROC curves are shown next to each condition (standard error of the mean in *brackets*).

change in set size required to minimize the difference between ground truth and bootstrapped values. Therefore, the removal of outliers facilitates the use of conventional, Gaussian statistics, while not affecting the rate of VF defect detection.

Limitations

This study has a number of limitations. Multiple SITA-Standard VF results were used from each individual patient. Although this is a method practiced in previous papers and in existing normative databases, there may be confounding factors due to the relative contributions of each subject.[9,10] However, the prospective phase of the study, where only one VF result from each subject was used, arrived at similar conclusions. Although we sourced a large number of VF results, the true population norms are not known. The sensitivity values were also age-corrected to a 50-year-old equivalent, which may be subtly different to deriving a normative database with age-matched healthy subjects. However, the change in sensitivity per decade at each test location was small, and the amount of age correction performed on our normative cohort was within the instrument's test-retest variability range.[20]

Our conclusions rest on a ground truth where the normal representative "population" consisted of 500 SITA-Standard and 100 full-threshold VF results found using standard parameters (achromatic Goldmann size III target, presented for 200 ms upon an achromatic background). One of the analyses performed in the present study was the examination of whether the point where no further change occurred in 95th, 5th, and SD values when using different baseline cohort sizes (comparing $n = 300$ and $n = 400$ with the total "population" of $n = 500$). While we showed that the set size x was relatively unchanged at approximately 150 (instead of being a proportion relative to the total cohort size), this was still based on the assumption that $n = 500$ provided a reasonable estimate of the total normal population. True norms across a range of ages would require a much larger study with a more diverse representation of the population.[27] Population characteristics would also be specific to the research question being asked. However, we also compared the VF sensitivity parameters of our cohorts with that of other published studies from different geographic locations, and found no significant difference (Table 2, one-way ANOVA, excluding the sets where SD was not reported: $F_{1,1} = 0.646$, $P = 0.5690$).[5,28–30] This indicated that our cohort was likely to be robust and representative of a general, diverse population.

Conclusions

Our results suggest that set sizes of approximately 150 normal SITA-Standard VF results and 60 normal full-threshold VF results provide a close estimate of the ground truth statistics. For similar ability to detect defects in glaucoma when using the SITA-Standard paradigm, a sample size of roughly 200 normal VFs closely approximates the performance of

Table 2. Mean (SD, Where Available) Sensitivities (Averaged Across the VF, in dB) for SITA-Standard and Full Threshold VFs of the Patients from the Present Study Compared With Those Reported by Other Studies

	Present Study ($n = 500$)	Bengtsson & Heijl[28] ($n = 335$)[a]	Shirato et al.[29] ($n = 19$)[b]	Wild et al.[30] ($n = 50$)	Wall et al.[5c]
SITA-standard	30.0 (1.66)	29.5	29.8 (1.2), 30.2 (1.5)	29.27 (1.82)	29.5, 29.6
Full threshold	28.0 (1.53)	28.3	28.0 (1.4), 28.5 (1.5)	28.47 (1.80)	n/a

[a] SD not available from the paper.
[b] Paper reported two mean sensitivity values for each test.
[c] Paper reported median and interquartile range; n/a, not available.

a ground truth of 500 healthy subjects. Thus, a smaller and possibly more practically sized cohort of 150 to 200 healthy subjects may be used to derive descriptive statistics that reflect the underlying population, if the cross section is representative of the population of interest. These results provide further guidance toward the construction of future VF normative databases.

Acknowledgments

The authors thank Janelle Tong and Henrietta Wang for technical assistance.

Supported by grants from a PhD scholarship provided by Guide Dogs NSW/ACT and an Australian Government Research Training Program PhD scholarship (JP); an Australian Research Council Future Fellowship (BVB; FT130100338); and this work was supported by the National Health and Medical Research Council of Australia (NHMRC #1033224). Guide Dogs NSW/ACT are partners in the NHMRC grant.

Disclosure: **J. Phu**, None; **B.V. Bui**, None; **M. Kalloniatis**, None; **S.K. Khuu**, None

References

1. Heijl A. The Humphrey Field Analyzer, construction and concepts. In: Heijl A, Greve EL, ed. *Proceedings of the 6th International Perimetric Society Meeting*. Dordrecht: Junk Publishers; 1985:77–84.
2. Mills RP, Budenz DL, Lee PP, et al. Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease. *Am J Ophthalmol*. 2006;141:24–30.
3. Choi AYJ, Nivison-Smith L, Khuu SK, Kalloniatis M. Determining spatial summation and its effect on contrast sensitivity across the central 20 degrees of visual field. *PloS One*. 2016;11: e0158263.
4. Phu J, Khuu SK, Zangerl B, Kalloniatis M. A comparison of Goldmann III, V and spatially equated test stimuli in visual field testing: the importance of complete and partial spatial summation. *Ophthalmic Physiol Opt*. 2017;37: 160–176.
5. Wall M, Doyle CK, Zamba KD, Artes P, Johnson CA. The repeatability of mean defect with size III and size V standard automated perimetry. *Invest Ophthalmol Vis Sci*. 2013;54: 1345–1351.
6. Wall M, Kutzko KE, Chauhan BC. Variability in patients with glaucomatous visual field damage is reduced using size V stimuli. *Invest Ophthalmol Vis Sci*. 1997;38:426–435.
7. Wall M, Woodward KR, Doyle CK, Zamba G. The effective dynamic ranges of standard automated perimetry sizes III and V and motion and matrix perimetry. *Arch Ophthalmol*. 2010;128: 570–576.
8. Zalta AH, Burchfield JC. Detecting early glaucomatous field defects with the size I stimulus and Statpac. *Br J Ophthalmol*. 1990;74:289–293.
9. Asman P, Heijl A. Glaucoma Hemifield Test. Automated visual field evaluation. *Arch Ophthalmol*. 1992;110:812–819.
10. Heijl A, Lindgren G, Olsson J. Normal variability of static perimetric threshold values across the central visual field. *Arch Ophthalmol*. 1987;105: 1544–1549.
11. Fitzner K, Heckinger E. Sample size calculation and power analysis: a quick review. *Diabetes Educ*. 2010;36:701–707.
12. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci*. 2014;7:342–346.

13. Phu J, Khuu SK, Nivison-Smith L, et al. Pattern recognition analysis reveals unique contrast sensitivity isocontours using static perimetry thresholds across the visual field. *Invest Ophthalmol Vis Sci*. 2017;58:4863–4876.
14. Bengtsson B, Olsson J, Heijl A, Rootzen H. A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand*. 1997;75:368–375.
15. Garway-Heath DF, Caprioli J, Fitzke FW, Hitchings RA. Scaling the hill of vision: the physiological relationship between light sensitivity and ganglion cell numbers. *Invest Ophthalmol Vis Sci*. 2000;41:1774–1782.
16. Bengtsson B, Heijl A, Olsson J. Evaluation of a new threshold visual field strategy, SITA, in normal subjects. Swedish Interactive Thresholding Algorithm. *Acta Ophthalmol Scand*. 1998;76: 165–169.
17. Yenice O, Temel A. Evaluation of two Humphrey perimetry programs: full threshold and SITA standard testing strategy for learning effect. *Eur J Ophthalmol*. 2005;15:209–212.
18. Cherink MR. Bootstrap Methods: A Practitioner's Guide. New York: Wiley; 1999.
19. Motulsky HJ, Brown RE. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics*. 2006;7:123.
20. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from Full Threshold, SITA Standard, and SITA Fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43:2654–2659.
21. Armstrong RA, Davies LN, Dunne MC, Gilmartin B. Statistical guidelines for clinical studies of human vision. *Ophthalmic Physiol Opt*. 2011;31: 123–136.
22. Gardiner SK, Demirel S, Goren D, Mansberger SL, Swanson WH. The effect of stimulus size on the reliable stimulus range of perimetry. *Transl Vis Sci Technol*. 2015;4(2):10.
23. Swanson WH, Horner DG, Dul MW, Malinovsky VE. Choice of stimulus range and size can reduce test-retest variability in glaucomatous visual field defects. *Transl Vis Sci Technol*. 2014; 3(5):6.
24. Wang Y, Henson DB. Diagnostic performance of visual field test using subsets of the 24-2 test pattern for early glaucomatous field loss. *Invest Ophthalmol Vis Sci*. 2013;54:756–761.
25. Asaoka R. Mapping glaucoma patients' 30-2 and 10-2 visual fields reveals clusters of test points damaged in the 10-2 grid that are not sampled in the sparse 30-2 grid. *PLoS One*. 2014;9:e98525.
26. Anderson AJ, Vingrys AJ. Small samples: does size matter? *Invest Ophthalmol Vis Sci*. 2001;42: 1411–1413.
27. Williamson HA Jr, Williamson MT. The Beck Depression Inventory: normative data and problems with generalizability. *Fam Med*. 1989;21:58–60.
28. Bengtsson B, Heijl A. Inter-subject variability and normal limits of the SITA Standard, SITA Fast, and the Humphrey Full Threshold computerized perimetry strategies, SITA STATPAC. *Acta Ophthalmol Scand*. 1999;77:125–129.
29. Shirato S, Inoue R, Fukushima K, Suzuki Y. Clinical evaluation of SITA: a new family of perimetric testing strategies. *Graefes Arch Clin Exp Ophthalmol*. 1999;237:29–34.
30. Wild JM, Pacey IE, Hancock SA, Cunliffe IA. Between-algorithm, between-individual differences in normal perimetric sensitivity: full threshold, FASTPAC, and SITA. Swedish Interactive Threshold algorithm. *Invest Ophthalmol Vis Sci*. 1999;40:1152–1161.