

Reference intervals and diagnostic cut-offs for full-field electroretinography derived from real-world clinical data

Veronique Biggi Nicole Weiß

Friedrich-Alexander-Universität Erlangen-Nürnberg

Jan Kremers

Friedrich-Alexander-Universität Erlangen-Nürnberg

Cord Huchzermeyer

`cord.huchzermeyer@uk-erlangen.de`

Friedrich-Alexander-Universität Erlangen-Nürnberg

Research Article

Keywords: electrophysiology, real-world-data, reference intervals, full-field electroretinogram

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11351642/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Reference intervals and diagnostic cut-offs for full-field electroretinography derived from real-world clinical data

Veronique Biggi Nicole Weiß¹, Jan Kremers¹, Cord Huchzermeyer¹

¹Department of Ophthalmology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany

Corresponding author:

Veronique B.N. Weiß

Schwabachanlage 6

91054 Erlangen

veronique.weiss@me.com

Cord Huchzermeyer

Schwabachanlage 6

91054 Erlangen

Cord.Huchzermeyer@uk-erlangen.de

Orcid ID:

Veronique B.N. Weiß: 0000-0003-4116-950X

Cord Huchzermeyer: 0000-0002-0408-8981

Jan Kremers: 0000-0002-9293-7380

Key Words: electrophysiology; real-world-data; reference intervals; full-field electroretinogram

Abstract

Purpose:

Evaluate RefineR, a statistical method for indirect estimation of reference intervals (RIs) from real-world data, to derive RIs for full-field electroretinogram (ffERG) amplitudes and to compare these intervals with both RIs derived by direct measurements and cut-offs derived from receiver-operator-characteristic (ROC) analyses.

Methods:

Retrospective analysis of ISCEV-standard ffERGs obtained with Henkes contact lens electrodes at a tertiary referral center (January 2003 – January 2024; 3,812 examinations, 7,539 single-eye recordings, 3,116 patients). The central 95% RIs for five amplitude parameters were estimated using refineR. RefineR estimates RIs from statistical distribution using Box-Cox transformation without assuming a specific underlying distribution. For comparison, a clinical validation cohort (n=100) was randomly selected from 2019-2024, stratified by age and classified into six diagnostic groups based on their medical records. Diagnostic performance was assessed by ROC analysis and sensitivity and specificity were compared with three cut-offs: the lower-limits of the direct (1) and indirect (2) RIs, and the ROC-derived clinical decision cut-off.

Results:

Amplitude distributions were bimodal with the refineR derived lower-limits falling within the trough separating the two modes (68.8, 115.2, 235, 25.1 and 59.8 μ V for DA 0.01 b-, DA 3.0 a-/b- and LA 3.0 a-/b-waves, respectively). The estimated intervals were broader than the existing laboratory values (lower limits 148.0, 221.9, 400.0, 38.0 and 100.0 μ V, respectively). Exploratory ROC analyses of the clinically labelled subset showed that cut-offs enabling classification into clinically meaningful groups lay between direct and indirect RIs (Youden, cut-offs 47.4, 92.1, 349.9, 27.3 and 102.7 μ V, respectively).

Conclusion:

Indirect estimation from routine data could be feasible for ffERG amplitudes, but the proportion of normal ERGs was probably too small in our cohort. The value of RIs from well-characterized, highly selected normal cohort is uncontested, but our results show, that new, data-driven methods for classification of ffERGs may support clinical decision making in the future.

Word Count: 299 (max. 300)

58 Statement and Declarations

59 **Competing Interests:** The authors have no relevant financial or non-financial interests to disclose.

60

Introduction

Full-field electroretinography (ffERG) allows the objective assessment of global retinal function. Currently the International Society for Clinical Electrophysiology of Vision (ISCEV) standard for full-field clinical electroretinography defines six stimulus conditions, providing criteria for recording methodology, response nomenclature and quality control [1]. The ISCEV standard explicitly recommends that each laboratory establishes its own normative reference data derived from an appropriate local reference cohort [1–3], due to variability of ffERG amplitudes and implicit times between recording systems and populations.

ISCEV also gives practical and statistical guidance on how to create reference ranges [1]. How such reference ranges should be built has been reviewed in detail. For direct estimation, the accepted gold standard is a sample of at least 120 healthy subjects stratified for different age decades and a nonparametric analysis where the reference is defined as the interval covering the central 95 % of that reference population [2]. Assembling such a cohort is labor-intensive and sometimes not feasible in routine care. Such a reference interval characterizes a highly selected normal or even “super-normal” reference population.

Indirect methods provide an alternative by estimating reference limits from heterogeneous data that accumulate in everyday practice, without selecting a predefined reference group [4]. Instead, “normal” subjects are selected from a large clinical dataset with statistical techniques based on under the assumption that a sufficient number of normal subjects is measured. The R package *refineR* and the underlying technique for reference interval estimation is based on real-world data and uses an inverse Box–Cox transformation as well as likelihood-based optimization to isolate this “non-pathological subpopulation” hidden inside a mixed clinical distribution [5].

Reference intervals can be used for all clinical purposes but it may not always be the optimal cut-off for discrimination between presence and absence of disease in specific clinical situations. For example, clinicians use ffERG to detect retinal diseases and are not interested in “nonspecifically” reduced amplitudes due to lens opacification. Receiver operating characteristic (ROC) analysis can be used to derive optimal cut-off values for classification in a specific clinical context. For this, a training group with known values and a “gold-standard” classification (usually healthy vs. disease) is necessary. The area under the ROC curve (AUC) captures how well it discriminates overall, and the Youden index picks the cut-off that maximizes sensitivity and specificity.

The present study aims to (i) estimate laboratory-specific reference intervals for key ffERG amplitude parameters from a large real-world clinical dataset using the *refineR* algorithm; (ii) derive cut-offs based on ffERG amplitudes of 100 patients, randomly collected from the large dataset, for a pre-defined clinical classification using an ROC analyses; and (iii) compare the diagnostic performance of

95 the refineR-derived reference limits from (ii) with the laboratory reference values currently used in
96 clinical practice and the Youden-index optimal thresholds derived from the ROC analysis.
97

98 Methods

99 Study design and setting

100 For this retrospective observational study, ERG data from routine clinical examinations were
101 retrospectively analyzed. No additional examinations were performed. All ISCEV standard full-field
102 electroretinograms that were recorded with the RETIport system (Roland Consult, Brandenburg a.d.
103 Havel, Germany) and Henkes contact lens electrodes between January 2003 and January 2024 at the
104 Department of Ophthalmology, University Hospital Erlangen, Germany, a tertiary care center. Raw
105 data were exported as comma-separated values (CSV) files from the internal database of the RETIport
106 (n=5,069 examinations).

107 For the second part of the study, clinical data were collected from the health records for a subset of
108 100 patients. These patients were randomly picked from recordings between January 2019 and
109 December 2024 after stratification for age. This period was selected because complete clinical records
110 were available for this period. This allowed classifying this subset into four different clinical groups:
111 (1) healthy, (2) suspected functional visual impairment, (3) retinal disease or dystrophy; and (4) other
112 ocular pathology.

113 Data handling complied with institutional data governance policies and the General Data Protection
114 Regulation (“Datenschutzgrundverordnung”, DSGVO). All data were anonymized prior to analysis.
115 For a retrospective analysis of anonymized data, informed consent was not necessary, and a waiver of
116 consent was obtained from the ethics committee of the medical faculty of the Friedrich-Alexander-
117 University Erlangen-Nürnberg (Reference 25-195-Br).

118 Electrophysiological recordings

119 Between January 2003 and January 2024, full-field ERGs had been obtained during routine clinical
120 practice using the RETIport system in accordance with the ISCEV standard for full-field clinical
121 electroretinography applicable at the time of examination [1, 6]. Indication for testing was determined
122 by ophthalmologists. Three stimulus conditions were analyzed: dark-adapted $0.01 \text{ cd} \cdot \text{s} \cdot \text{m}^{-2}$ (rod
123 response, DA 0.01), dark-adapted $3.0 \text{ cd} \cdot \text{s} \cdot \text{m}^{-2}$ (combined rod–cone response, DA 3.0) including
124 oscillatory potentials, and light-adapted $3.0 \text{ cd} \cdot \text{s} \cdot \text{m}^{-2}$ (cone single-flash response, LA 3.0)

125 For each of these three conditions, a-wave and b-wave peak amplitudes and the corresponding
126 implicit times were extracted. Markers were automatically placed at the minima or maxima within
127 pre-defined time intervals by the manufacturer’s software. They were inspected by a trained
128 technician and corrected manually if necessary.

In clinical practice, reference intervals were used that had been derived from a selected reference cohort between June 2002 and December 2003. Demographics are still available for this cohort along with the reference intervals, but no further details had been documented.

Laboratory reference values used in clinical practice

The laboratory reference values used in clinical practice (for Henkes electrodes) had been derived from a selected cohort of healthy subjects (n=43) from June 2002 to December 2003. Participants were divided into three age groups (<30, 30–50 and >50 years) to reflect the age dependence of ffERG amplitudes. The cohort covered a wide age range between 9 and 82 (median 41 years) but most were female (72 %, Table 1).

Table 1. Demographic characteristics of the ERG reference cohort (Henke normative data), stratified by age group.

Age group, years	n	Male, n (%)	Female, n (%)	Age, median (range)
< 30	16	3 (19)	13 (81)	24 (9–30)
30–50	14	6 (43)	8 (57)	42.5 (32–47)
> 50	13	3 (23)	10 (77)	60 (51–82)
All	43	12 (28)	31 (72)	41 (9–82)

n, number of participants. Sex is given as absolute count with column percentage in parentheses; age is given in years as the median with range (minimum–maximum). Percentages are rounded and may not sum to 100.

Data source and preprocessing

The complete institutional ffERG database comprised 3,812 ffERG examinations (7,539 single-eye recordings) from 3,116 patients with Henkes contact lens electrode recordings acquired between January 2003 until January 2024. Each examination presented from one or both eyes. When the two eyes were recorded, they were treated as separate and independent measurements.

Electrophysiological raw data, filter settings, a- and b-wave amplitudes and implicit times, as well as basic demographic information were exported into csv-files using an export function implemented in the manufacturer's software for the RETIport system. These files were parsed with R (R Foundation for Statistical Computing, Vienna, Austria) for further processing and analysis.

Automated preprocessing was also implemented in R and comprised harmonization of variable names (which had changed over the years with different versions of the RETIport software), type conversion

and removal of exact technical duplicates. Manual review was limited to correcting structural inconsistencies introduced during database export, such as misallocated columns, mislabeled stimulus conditions and column shifts arising from spreadsheet parsing. No electrophysiological measurements (amplitudes or implicit times) were modified manually and no recordings were excluded on the basis of subjective waveform interpretation. Harmonization of the data accounted for changes in manufacturer nomenclature over time. Non-positive amplitude values, indicating failed marker placement, were excluded before reference interval estimation.

Reference interval estimation

Laboratory-specific reference intervals were estimated using an indirect method to derive reference limits from heterogenous real-world-clinical data with the refineR package in R (R 4.5.2, refineR 2.0.0) [5]. The refineR algorithm models the non-pathological subpopulation within a mixed clinical dataset using a Box-Cox transformation:

$$y(\lambda) = \frac{x^\lambda - 1}{\lambda} \text{ for } \lambda \neq 0 \text{ and } y = \ln(x) \text{ for } \lambda = 0$$

This allows skewed laboratory data to be approximated by a Gaussian distribution[5]. It then fits the parameters λ , μ , σ and a population scaling factor to the presumed physiological population while minimizing the influence of pathological values [5]. In contrast a conventional Gaussian approach fits to the entire dataset and typically defines the 95% reference interval as: $\mu \pm 1.96\sigma$.

This optimization estimates the model parameters and corresponding reference intervals [5]. Central 95 % reference limits were derived from the estimated non-pathological distribution for each parameter and stimulus condition. The resulting amplitude distributions with refineR-derived lower limits are shown in Figure 2.

Validation cohort and clinical classification

A random subset of 100 patients was drawn from the overall dataset (Figure 3) for clinical validation. For this part of the analysis, a single recording per patient was retained (the right eye was preferred and the left eye used only if no right-eye recording was available) to avoid bias by intra-subject correlation.

Each case underwent structured categorization based on the available routine clinical information and was assigned to one of four primary groups: (1) healthy, (2) suspected functional visual impairment (including malingering), (3) retinal disease or dystrophy; or (4) other ocular pathology (Figure 3). Retinal disease or dystrophy was further subdivided into generalized dystrophies, localized dystrophies and other retinal changes. Where applicable diagnostic certainty was specified as clinically or genetically confirmed.

For the ROC analysis cases were additionally grouped for further analysis. These subsets included retinal pathology vs non-retinal conditions (retinal disease/dystrophies (3) vs. all other categories (1, 2, and 4) Figure 3. dotted line) and healthy (1) vs. all ocular pathologies (2, 3 and 4).

Statistical analysis

Data handling was done with the *data.table* package (v 1.17.8), and ROC analyses were performed using the *pROC* package (v. 1.19.0.1)[7]. The diagnostic performance of individual ffERG parameters in distinguishing predefined clinical groups was assessed by ROC analysis. The area under the curve (AUC) served as the primary measure of diagnostic performance, with 95 % confidence intervals estimated by non-parametric methods. For the ROC analysis we predefined that the lower amplitudes and longer peak times were treated as indicating disease rather than the higher values. Allowing the software to select the direction would prevent the AUC from falling below 0.5 and would bias the estimates especially in small groups. Optimal cut-off values were determined using the Youden index [8] defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

This can be mathematically be equivalent to:

$$J = \text{Sensitivity} - (1 - \text{Specificity})$$

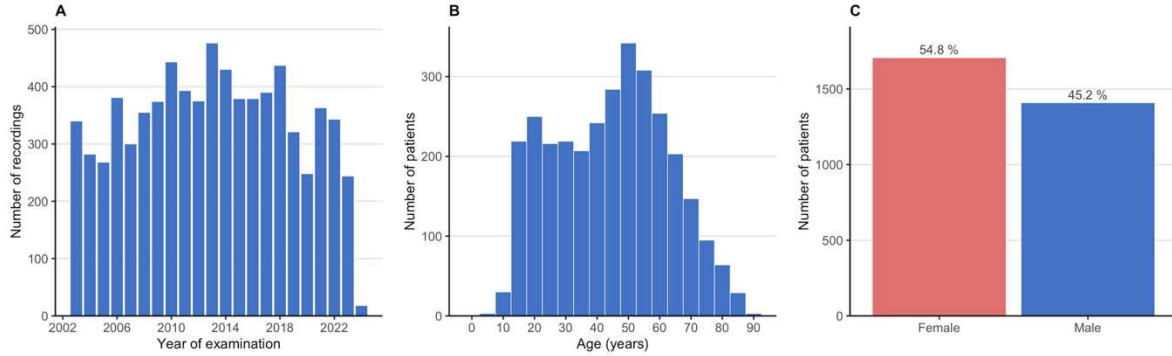
Where (1-Specificity) represents the false-positive rate. The Youden-index therefore identifies the ROC-curve point with the greatest vertical distance above the diagonal line of no discrimination, representing the cut-off that provides the best combined sensitivity and specificity [7, 8].

Cut-offs derived from three approaches were compared: 1) the reference limits derived from a selected cohort that had been used in clinical practice, 2) the lower limits of the refineR-derived intervals and 3) the Youden-based ROC thresholds. Sensitivity and specificity in the subgroup of 100 patients were calculated for each cut-off. The following contrasts were evaluated: healthy versus generalized retinal dystrophies; healthy versus localized retinal dystrophies; retinal pathology versus non-retinal conditions; and healthy versus all ocular pathologies (Figure 3).

Results

Study population

For the present analysis, ffERG examinations performed between January 2003 and January 2024 were included. Of the included patients, 1,707 (54.8%) were female and 1,409 (45.2 %) were male. Mean patient age at the first examination was 44.3 ± 18.2 years (median 46.0, IQR 29.0–58.0, range 1–90 years) (Figure 1).



Henkes contact lens electrode recordings, January 2003 – January 2024; n = 7,539 recordings from 3,812 examinations in 3,116 patients

Figure 1. Distribution of ffERG recordings and patient demographics for the Henkes electrode cohort (n = 3,812 examinations from 7,539 single-eye recordings from 3,116 patients). (A) Number of recordings per year. (B) Age distribution of patients at time of examination. (C) Sex distribution of the cohort. IQR, interquartile range.

Reference intervals derived from real-world data

Each distribution was bimodal with a narrow low-amplitude mode separated from the main mode by a trough where the reineR-derived lower limits consistently fell within this trough (Figure 2). For the DA 3.0 b-wave, 102 recordings (1.4 %) showed amplitudes below 10 μV and 342 (4.6 %) below 20 μV ; these were distributed evenly across the observation period, which is consistent with severely reduced or extinguished responses in advanced generalised dystrophies rather than with technical failure. This reflects the heterogenous clinical composition of the dataset from which the limits were estimated. Reference intervals were also estimated from the corresponding peak times. Reference intervals were also estimated by refineR for all analyzed amplitude parameters (Table 2). For all stimulus conditions, refineR-derived lower limits were substantially below the laboratory-specific reference values, extending the non-pathological range towards lower amplitudes.

Table 2. Laboratory-specific reference intervals (central 95 %) for selected ffERG amplitude parameters. Henkes values are age-stratified; refineR-derived limits were estimated from the Henkes electrode subset.

ERG response	Parameter	Henkes <30 yrs (μV)	Henkes 30–50 yrs (μV)	Henkes >50 yrs (μV)	refineR (μV)
DA 0.01 (Rod)	b-wave	183–359	148–316	122–308	69–307

ERG response	Parameter	Henkes <30 yrs (μV)	Henkes 30–50 yrs (μV)	Henkes >50 yrs (μV)	refineR (μV)
DA 3.0 (Combined)	a-wave	265–436	221–420	173–337	115–402
	b-wave	473–723	400–720	388–649	236–690
LA 3.0 (Cone single flash)	a-wave	43–88	38–80	31–64	25–81
	b-wave	156–280	100–229	89–177	60–257

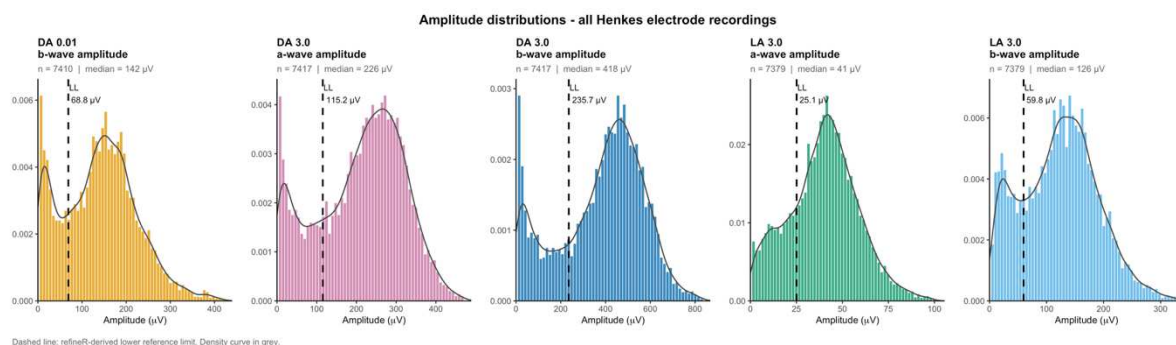


Figure 2. Amplitude distributions for all Henkes electrode recordings with refineR-derived lower reference limits. Distributions are shown for the five analysed amplitude parameters: DA 0.01 b-wave (A), DA 3.0 a-wave (B), DA 3.0 b-wave (C), LA 3.0 a-wave (D) and LA 3.0 b-wave (E). Dashed vertical lines indicate the refineR-derived lower reference limit (LL, in μV). Values below zero and extreme outliers (>99th percentile) are excluded for display. n, number of recordings with non-missing values for the respective parameter.

Subgroup for ROC analyses

A random subset of 100 patients was drawn from this cohort for clinical validation; demographic and clinical characteristics are summarised in Table 3. The cohort comprised 22 healthy controls and 53 patients with retinal disease or dystrophy (33 generalised, 8 localised, 12 other retinal changes), with the remaining 25 patients classified as suspected functional visual impairment (n = 8) or other ocular pathologies (n = 17) (Figure 3).

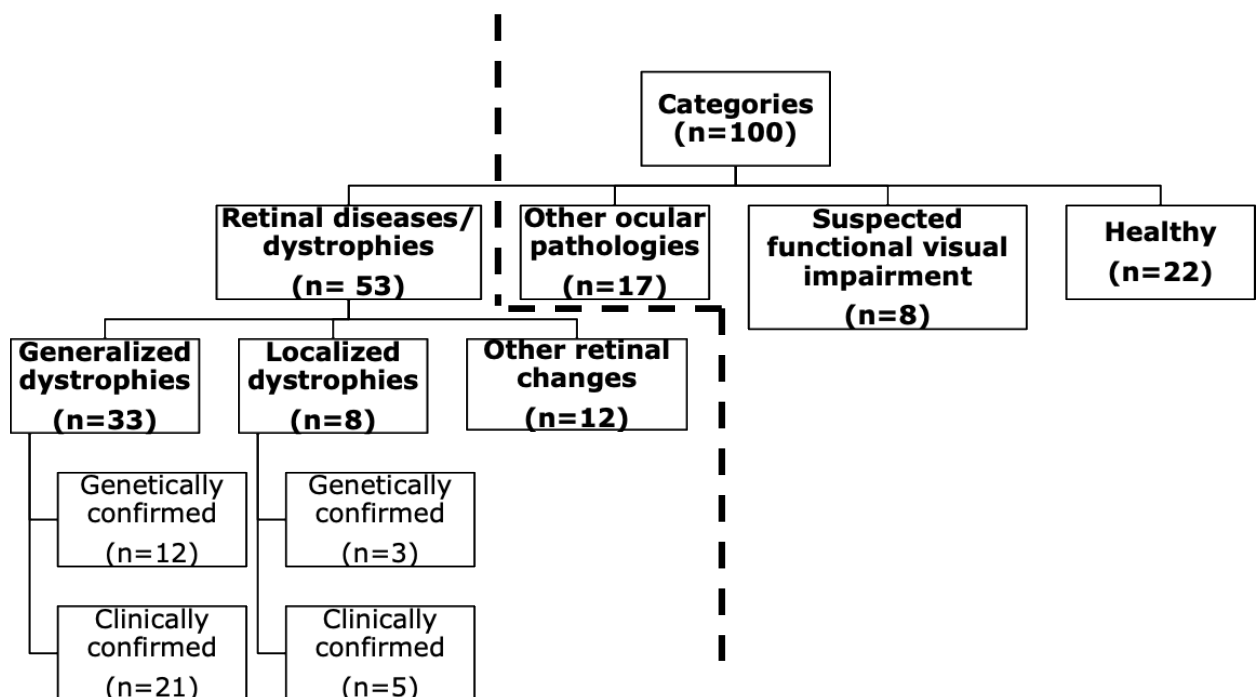


Figure 3. Clinical classification of the validation cohort (n = 100). Patients were assigned to one of four principal groups: healthy controls, suspected functional visual impairment, retinal disease or dystrophy (further subdivided into generalised dystrophies, localised dystrophies and other retinal changes), and other ocular pathologies. The dotted line indicates the binary grouping applied in the retinal versus non-retinal contrast for receiver operating characteristic analysis.

Table 3. Demographic and clinical characteristics of the validation cohort (n = 100).

Characteristic	n (%) or value
Age (years)	
Mean $\pm$ SD	48.3 $\pm$ 22.3
Median (IQR)	47.0 (28.0–69.3)
Age group	
0–33 years	33 (33)
34–66 years	35 (35)
>66 years	32 (32)
Sex	
Female	56 (56)

Male	44 (44)
Clinical classification	
Healthy	22 (22)
Suspected functional visual impairment	8 (8)
Other ocular pathology	17 (17)
Retinal disease or dystrophy	53 (53)
Generalised dystrophies	33 (33)
<i>Clinically confirmed</i>	21 (21)
<i>Genetically confirmed</i>	12 (12)
Localised dystrophies	8 (8)
<i>Clinically confirmed</i>	5 (5)
<i>Genetically confirmed</i>	3 (3)
Other retinal changes	12 (12)

IQR, interquartile range; SD, standard deviation. Continuous variables are given as mean $\pm$ SD and median (interquartile range); categorical variables as number of patients (percentage of cohort).

Diagnostic performance of ffERG amplitudes

ROC analyses showed good to excellent discrimination between healthy controls and generalized retinal dystrophies. The largest area under the curve (AUC) was obtained for the DA 3.0 b-wave amplitude (AUC 0.948, 95 % CI 0.888–1.000), followed by the LA 3.0 b-wave amplitude (0.925, 0.847–1.000), the DA 3.0 a-wave amplitude (0.893, 0.809–0.977) and the DA 0.01 b-wave amplitude (0.902, 0.815–0.989) (Figure 4A).

Discrimination between healthy controls and localized retinal dystrophies was moderate. The LA 3.0 b-wave amplitude reached the largest AUC (AUC 0.864, 95 % CI 0.681–1.000), followed by the DA 3.0 b-wave amplitude (0.830, 0.666–0.993), the DA 0.01 b-wave peak time (0.716, 0.508–0.924) and the DA 3.0 a-wave amplitude (0.773, 0.602–0.944) (Figure 4B).

For the broader contrast of retinal pathology versus non-retinal conditions, the DA 3.0 b-wave amplitude (AUC 0.799, 95 % CI 0.712–0.886), the LA 3.0 b-wave amplitude (0.805, 0.715–0.894), the LA 3.0 a-wave amplitude (0.749, 0.648–0.849) and the DA 0.01 b-wave amplitude (0.737, 0.639–0.836) ranked highest (Figure 4C).

Comparing healthy controls against all ocular pathologies combined, the DA 3.0 a-wave amplitude (AUC 0.815, 95 % CI 0.718–0.911), the DA 3.0 b-wave amplitude (0.819, 0.729–0.910), the DA 0.01 b-wave amplitude (0.786, 0.690–0.882) and the LA 3.0 b-wave amplitude (0.786, 0.694–0.877) were the most discriminative parameters (Figure 4D).

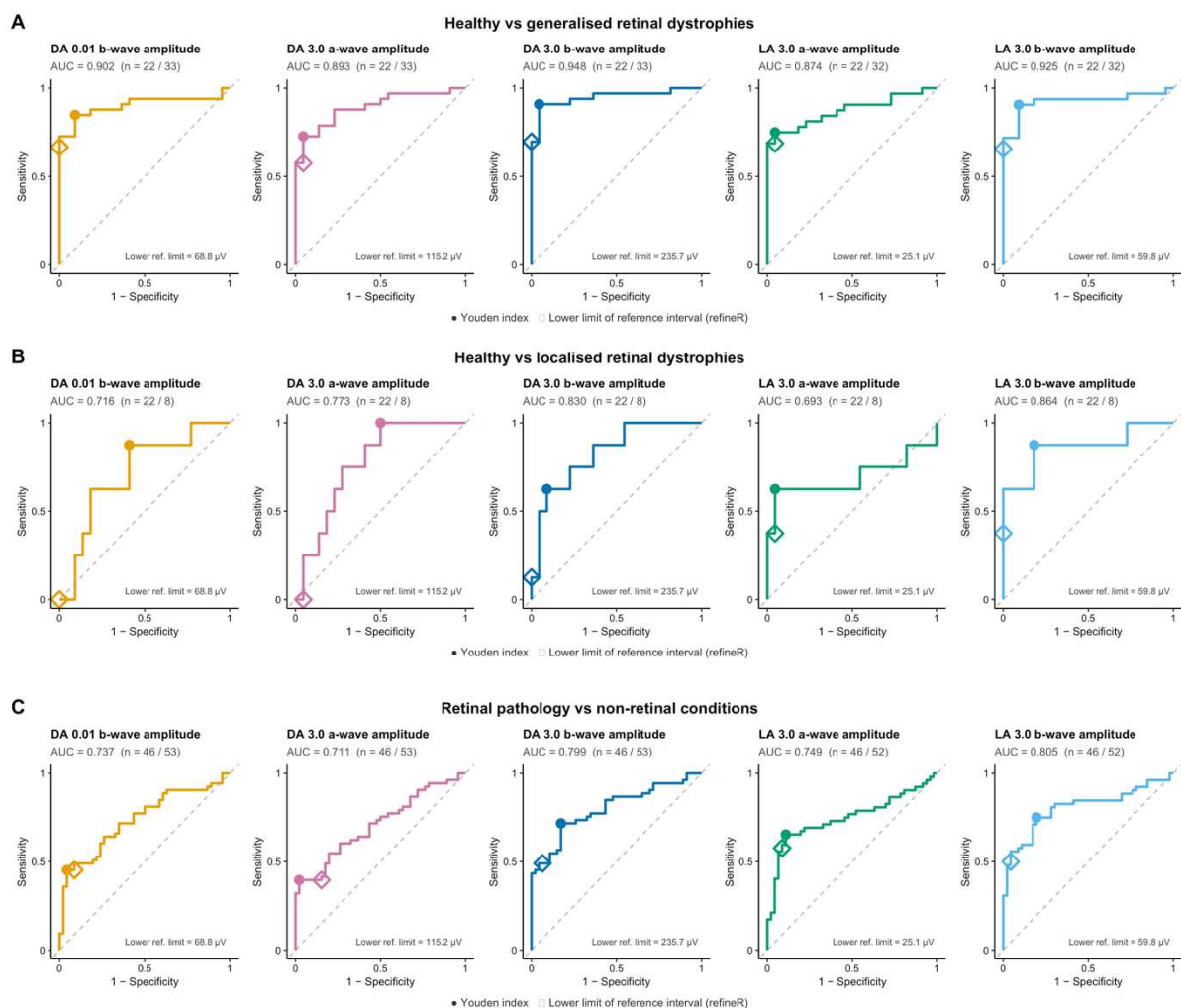
Comparison of different limits of the reference interval

For the contrast of retinal pathology versus non-retinal conditions, sensitivities and specificities differed systematically between the three cut-off approaches (Table 4). Laboratory-specific reference values yielded moderate sensitivities (0.62–0.75) at moderate specificities (0.57–0.80). The lower limits of the refineR-derived intervals produced the highest specificities (0.85–0.96) but at reduced sensitivity (0.40–0.58). Youden-derived spanned a wider range with sensitivities of 0.40–0.75 and specificities of 0.80–0.98 across the analysed amplitude parameters. This contrast was selected for detailed cut-off comparison as it represents the primary clinical decision context in which ffERG findings guide the differentiation between retinal and non-retinal conditions.

Table 4. Comparison of cut-off strategies for the contrast retinal pathology versus non-retinal conditions. Sensitivity (Sens.) and specificity (Spec.) were calculated at the cut-off derived from laboratory-specific reference values, refineR-derived lower limits, and the Youden-index optimal threshold, respectively.

Parameter	Lab. reference values (Henkes electrodes)			refineR			Youden index		
	Cutoff (μ V)	Sens.	Spec.	Cutoff (μ V)	Sens.	Spec.	Cutoff (μ V)	Sens.	Spec.
DA 0.01 b-wave amplitude	148.0	0.72	0.65	68.8	0.45	0.91	47.4	0.45	0.96
DA 3.0 a-wave amplitude	221.0	0.62	0.63	115.2	0.40	0.85	92.1	0.40	0.98
DA 3.0 b-wave amplitude	400.0	0.75	0.65	235.7	0.49	0.93	349.9	0.72	0.83

Parameter	Lab. reference values (Henkes electrodes)			refineR			Youden index		
	Cutoff (μV)	Sens.	Spec.	Cutoff (μV)	Sens.	Spec.	Cutoff (μV)	Sens.	Spec.
LA 3.0 a-wave amplitude	38.0	0.73	0.57	25.1	0.58	0.91	27.3	0.65	0.89
LA 3.0 b-wave amplitude	100.0	0.73	0.80	59.8	0.50	0.96	102.7	0.75	0.80



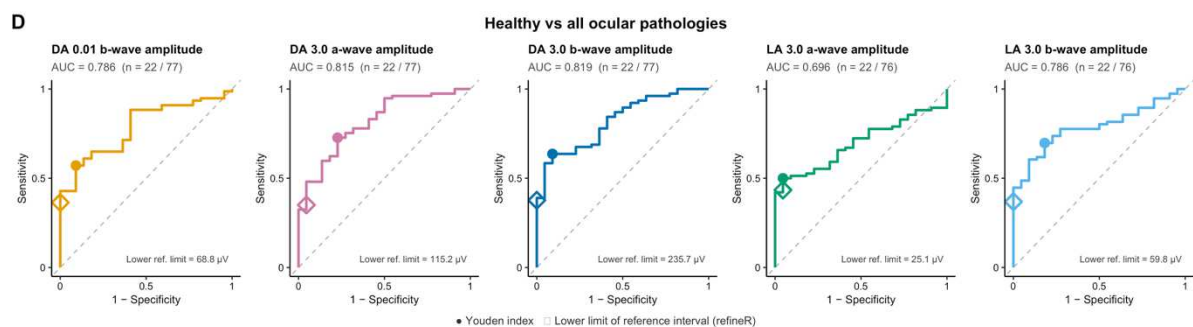


Figure 4 Receiver operating characteristic (ROC) analysis for five ffERG amplitude parameters across four diagnostic contrasts. Each panel shows five individual ROC curves side by side, one per parameter. Panel (A): healthy controls versus generalised retinal dystrophies. Panel (B): healthy controls versus localised retinal dystrophies. Panel (C): retinal pathology versus non-retinal conditions. Panel (D): healthy controls versus all ocular pathologies combined. Colour coding is consistent across all panels: orange, DA 0.01 b-wave amplitude; purple, DA 3.0 a-wave amplitude; dark blue, DA 3.0 b-wave amplitude; green, LA 3.0 a-wave amplitude; light blue, LA 3.0 b-wave amplitude. Filled circles mark the Youden-index optimal cut-off; open diamonds mark the refineR-derived lower reference limit. The diagonal dashed line represents chance-level discrimination (AUC = 0.5).

Discussion

This study estimated laboratory-specific reference intervals for ffERG amplitude parameters from a large, heterogeneous real-world clinical dataset using the indirect refineR algorithm and evaluated the diagnostic performance of these parameters through ROC analysis in a randomly selected validation cohort of 100 patients. The principal findings were: (i) refineR-derived intervals were consistently broader than the existing Henkes reference intervals particularly at the lower amplitude boundary; (ii) b-wave amplitudes under dark-adapted and light-adapted conditions showed good to excellent discrimination between healthy controls and generalized retinal dystrophies, with AUC values reaching 0.915 and (iii) Youden-index thresholds generally fell between the lower refineR and Henkes limits, reflecting a data-derived compromise between sensitivity and specificity.

Reference interval estimation

The R package refineR provides an indirect approach for estimating reference intervals from routine laboratory data in which healthy and pathological results may be mixed [5]. In the RI benchmark (developed by the same group that published the algorithm), refineR achieved the lowest overall deviation among the indirect methods compared [9]. This benchmark showed that accuracy depends on the fraction of pathological values more than on sample size with the results comparable to a direct study up to a pathological fraction of approximately 20% and substantial deviation at 50% or above [9]. The proportion of these pathological recordings in a tertiary electrophysiology laboratory like seen in this study is likely to exceed this range which may contributed to the lower limits.

The proportion of non-pathological results in a clinical dataset is not a fixed proportion but a property of the setting in which the data is acquired. This is what distinguishes visual electrophysiology from routine biochemistry for which the indirect methods were originally developed. Analyses of sodium or haemoglobin contents are typically requested in large panels and are frequently ordered in patients with a low probability of an abnormal result making the non-pathological fraction of a laboratory base typically high. ffERG recordings are, however, time consuming, required dark adaptation, mydriasis and corneal electrodes. This is generally reserved for patients in whom a specific retinal diagnosis is already suspected [1, 6]. Referral to a tertiary electrophysiology unit therefore acts as a preselection step that raises the pre-test probability of disease, an effect which has been long recognized as a determinant of apparent test performance in referral populations [10]. The pathological fraction in the present dataset is expected to lie above 20% up to which refineR has shown to approximate a direct study and within the range in which substantial deviation of the estimated limits has been reported [9].

The composition of the referral population therefore also determines the context on “non-pathological”. The conditions for which ffERG is principally requested (generalized photoreceptor dystrophies such as retinitis pigmentosa, or disorders of inner retinal transmission producing

electronegative waveform) are characterized by pronounced and in some cases near-pathognomic changes [11–13]. Conversely many patients undergo ffERG to exclude retinal diseases having conditions that may only cause modest amplitude reduction. These include media opacity which can affect ISCEV-standard responses even with mild cataract and full pupillary-dilation [14] and diabetes mellitus, where reduced dark- and light- adapted amplitudes have been reported before a clinical apparent retinopathy develops [15, 16]. More generalized amplitude reductions have also been described in neurodegenerative disorders such as Parkinsons and Alzheimers disease, although less consistently than in primary retinal disease [17, 18]. Those recordings are neither healthy nor abnormal in the sense intended by the referring clinician. The subpopulation isolated by refineR is therefore more accurately described as non-pathological with respect to the clinical question than as healthy (a distinction made explicit in the laboratory-medicine literature on indirect methods [4]) and this offers a substantive rather than a purely statistical explanation for a lower reference limit than that obtained by a direct sampled healthy cohort. This is compatible with the idea that laboratory-specific reference intervals established from appropriate local populations is recommended as the preferred approach to normative data in clinical electrophysiology [1, 2]. The present analysis pooled recordings across the full age range for refineR estimation. Age-related changes in ffERG parameters have been documented in healthy subject especially in dark-adapted responses [3, 19] where it is recommended to pre-filter by age when applying indirect methods [4]. Age-stratified estimation should be examined in a larger dataset with sufficient observations within each age group.

Diagnostic performance and cut-off comparison

ROC analysis demonstrated good discrimination between healthy controls and generalised retinal dystrophies (AUC range 0.874–0.948). This is consistent with the expectation that diffuse photoreceptor and inner nuclear layer involvement produces substantial, multi-parameter amplitude reduction which is detectable across all three standard stimulus conditions [11]. Performance was lower for localized dystrophies and for the broader, diagnostically heterogenous comparison groups. The three cut-off strategies showed markedly different sensitivity–specificity profiles. First, Henkes lower reference limits achieved moderate sensitivity (0.62 – 0.75) and specificity (0.57– 0.80) reflecting their derivation from a small, age-stratified healthy reference cohort. Second, the lower limits of the refineR-derived intervals, set at the 2.5th percentile of the modelled non-pathological distribution, were conservative and produced high specificities (0.85 – 0.96) at reduced sensitivity (0.40–0.58). Finally, the Youden-index thresholds, derived directly from the ROC curve of the validation cohort, provided the most balanced trade-off (sensitivity 0.40–0.75, specificity 0.80–0.98). The position of the Youden threshold relative to the reference limits was inconsistent across the parameters as for two parameters (DA 3.0 b wave, LA 3.0 a-wave) they fell between the Henkes and refineR limits and for other two paramters (DA 0.01 b-wave, DA 3.0 a wave) they were lower than

the refineR limit and the LA 3.0 b-wave was slightly above the Henkes limit. For the DA 0.01 b-wave, several thresholds achieved similar Youden indices despite markedly different sensitivities and specificities. Thus, equal weighting of sensitivity and specificity could not identify a clinically preferable threshold. The choice ultimately depends on the relative consequences of missed disease and unnecessary referral. Neither reference limits nor a purely statistical optimum should therefore be interpreted as an optimal diagnostic cut-off.

Furthermore, ROC derived measures are not intrinsic properties of a test but depend on the composition of the population in which they are estimated. The same amplitude parameters reached different AUCs depending on the comparison (e.g. LA 3.0b wave: 0.925 for generalised dystrophies on vs. 0.786 for all ocular pathologies combined). This reflects differences in disease severity and diagnostic spectrum rather than differences in test performance. The validation cohort drawn from our tertiary referral centre is likely enriched for advanced retinal disease. Sensitivities, specificities and Youden thresholds may therefore not transfer to populations with milder or earlier disease such as seen in screening settings. In addition, predictive values would further depend on disease prevalence which differs between clinical settings. Cut-offs derived here should therefore be validated in independent cohorts representative of the intended clinical use.

Comparison with related work

Previous work from our center has also examined the diagnostic information contained in electrophysiological recordings. Kulyabin et al. showed in multifocal ERG that classification performance is not restricted to isolated amplitude measurements but can also exploit waveform information [20]. Although the present analysis was limited to ffERG amplitudes, the moderate discrimination observed for several broader clinical comparisons suggests that amplitude alone does not fully capture the diagnostic information available in the recording.

These observations line up with what has been seen in clinical laboratory medicine, where indirect estimation from routine data tends to give wider reference limits than direct sampling of healthy volunteers [4, 21]. Similar datasets from laboratory information system found systematic differences between directly and indirectly derived reference intervals for standard biochemical markers [22] as well as for the indirect algorithms themselves. Closer to our own field, reference data for a range of ISCEV stimulus paradigms were reported in a recent study from ffERG recordings in 73 healthy subjects aged 14 to 73 with age-dependency modelled by regression rather than by age-stratification where the most pronounced age-related changes were observed in the dark-adapted responses [19]. Our indirect estimates from routine data complement these direct references by capturing the wider amplitude spread of an unselected clinical population.

419

420 Limitations

421 First, this was a retrospective single-center study restricted to recordings obtained with Henkes
422 electrodes on the RETIport system. The Henkes electrode subset spans approximately 21 years
423 (January 2003 to January 2024). Thereafter, Henkes electrodes were replaced and the ISCEV standard
424 for full-field ERG underwent several revisions [1]. Earlier recordings may therefore not fully conform
425 to the 2022 standard, introducing temporal heterogeneity in recording conditions that could affect
426 both reference interval estimates and diagnostic performance metrics.

427 Second, reference limits were estimated from all available recordings thus including both eyes of the
428 same individual as well as repeated examinations over the study period. Exclusion of multiple
429 measurement per subjects is recommended as pre-filtering when indirect methods are applied to
430 routine data [9] which would make the effective sample size smaller than the 7,539 single-eye
431 recordings included in the present study.

432 In contrast, the validation cohort comprised one eye per patient so-that the ROC analysis was based
433 on independent observations. However, the 100-patient validation cohort was drawn from the same
434 overarching clinical database used to fit the refineR model. Although refineR relies exclusively on the
435 amplitude distribution and makes no use of clinical group labels, which were assigned independently
436 and used only for ROC analysis, this partial overlap between estimation and validation data prevents a
437 fully independent assessment of the reference intervals. Furthermore, the Youden-index cut-offs were
438 both derived and evaluated within the same 100-patient cohort. Since the threshold is chosen to
439 maximise the combined sensitivity and specificity in this particular sample it partly adapts to random
440 variation in the data resulting in performance estimates that are likely to be optimistic. Independent
441 prospective validation is therefore warranted to assess the generalizability of these thresholds. Also,
442 the validation cohort of 100 patients yields wide confidence intervals for AUC estimates. The
443 localized dystrophy subgroup contained only eight patients, and upper confidence limits for several
444 AUC values in 4B reached 1.000, results for this contrast must be interpreted with caution and treated
445 as exploratory. For these reasons, the reported AUC values and Youden indices should be considered
446 exploratory. They help acknowledge the diagnostic consequences of the derived reference intervals,
447 but they are not generalizable.

448 Finally, refineR estimation was performed on the pooled age distribution; age-stratified indirect
449 reference limits, as advocated for clinical electrophysiology normative data [23], were not feasible
450 given the available metadata structure and represent an important priority for future work. The
451 Youden-derived thresholds are furthermore exploratory as in a cohort of this size several local optima

of the Youden index lie close together so-that minor changes in the underlying data shift the selected threshold considerably.

We did not investigate reference values for implicit times, because it is not clear what the implicit time is in traces with low signal-to-noise ratio.

Outlook

The large, anonymized dataset generated by this study, comprising 7,539 single-eye ffERG recordings from 3,812 examinations in 3,116 patients collected over approximately 21 years may also support analyses beyond the present study. In particular, age-stratified estimation and validation in a larger independent cohort would address important limitations of the current analysis. The dataset was curated in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) [24] but data sharing with initiatives such as the ISCEV Task Force on Reference Values could additionally allow comparison of laboratory-specific reference intervals across centers and populations.

466

467 Conclusions

468 Comparison with reference intervals established directly in a sufficiently large and well-characterized
469 reference cohort represents the current standard for the interpretation of ERG parameters. In our
470 study, however, the lower reference limits of the indirect reference intervals (iRIs), estimated from
471 real-world clinical data using the refineR algorithm, were substantially lower than those of the directly
472 established reference intervals (dRIs) and encompassed a relevant proportion of ERGs that would
473 conventionally be considered subnormal. This finding indicates that, in our setting, indirect estimation
474 from unselected real-world data cannot substitute for directly established reference intervals.

475 Importantly, reference limits should be distinguished from clinical decision thresholds. ROC-derived
476 thresholds obtained for specific clinical classification tasks in a subset of our clinical population fell
477 between the lower limits of the dRIs and iRIs. This suggests that the boundary of a directly
478 established reference interval does not necessarily represent the optimal threshold for a specific
479 clinical decision. Accordingly, ERGs falling outside a reference interval may still have limited
480 discriminatory value for a particular clinical question. In clinical practice, such mildly abnormal
481 ERGs without a clear disease-specific pattern are often described as “unspecific” or “nonspecific.”

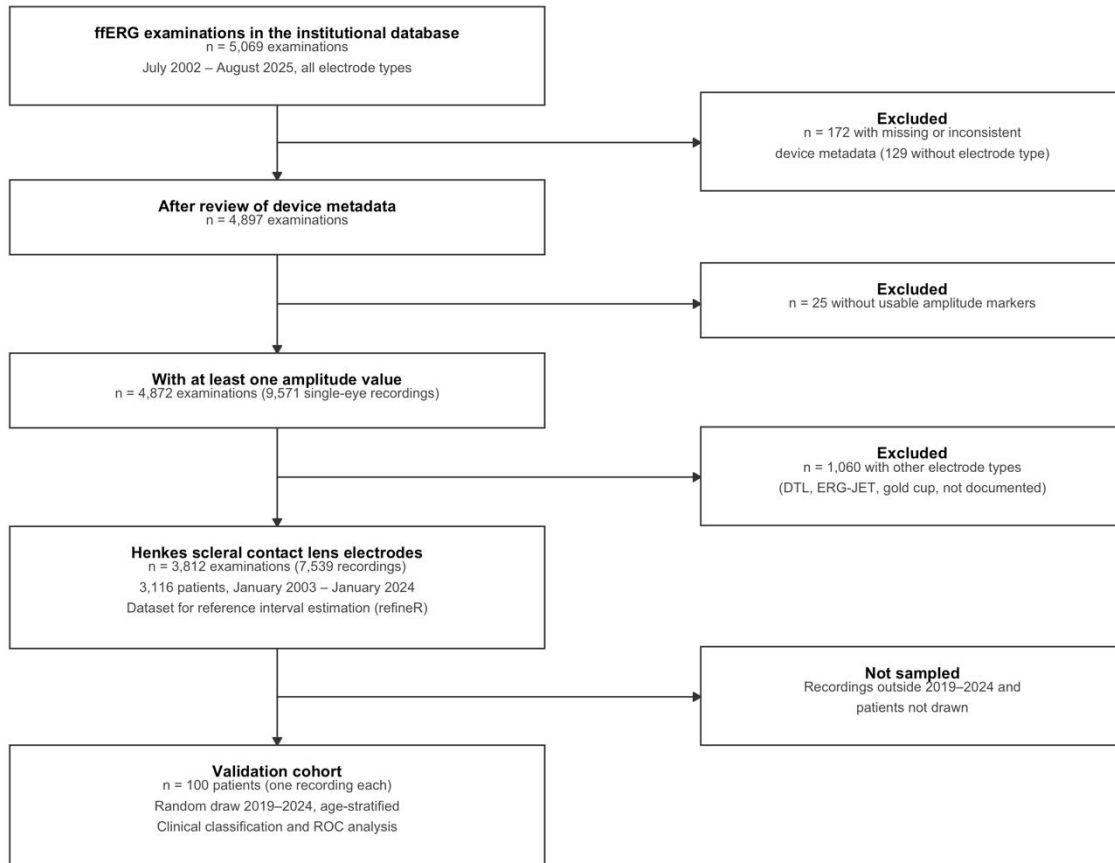
482 These observations support further investigation of data-driven approaches to ERG interpretation.
483 Potential improvements include more robust methods for estimating reference distributions from real-
484 world data, clinically annotated ERG datasets incorporating information from electronic health
485 records, multidimensional analysis integrating multiple ERG parameters, and automated classification
486 approaches based on transformed ERG representations and deep-learning methods. Such techniques
487 could be developed with the help of this large dataset.

488

489 Word Count: 5242

490

Supplementary Figures



Supplementary Figure S1. Flow of ffERG recordings and patients through the study. Of 5,069 examinations in the institutional database, 4,897 remained after manual review and 4,872 contained at least one amplitude measurement. Restriction to Henkes scleral contact lens electrodes and the study period yielded 3,812 examinations (7,539 single-eye recordings) from 3,116 patients. A stratified random sample of 100 patients from 2019–2024 was selected for clinical validation and ROC analysis, with one recording per patient.

501

502 Acknowledgements

503 This present work was performed by the first author, V. B. N. Weiß, in fulfillment of the requirements
504 for the degree “Dr. med.” at the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

505 The conception of this thesis, the data analysis, the interpretation of the results and the writing were
506 carried out independently by the author. Large language-model-based tools (Claude, Anthropic,
507 version Claude (2026) Anthropic. Model version Claude Sonnet 4.5) were used only in a supporting
508 manner for language refinement of individual passages and for occasional assistance in debugging and
509 refining R code. These tools were not used to generate scientific context to analyse or interpret data or
510 to draw conclusions. All suggestions were critically reviewed and verified by the author who takes
511 full responsibility for the content of this work.

512

513

514 Authors contributions

515 V.B.N. Weiß, and C. Huchzermeyer, contributed to study design, data collection and methodology. C.
516 Huchzermeyer supervised the research project. V. B.N. Weiß performed the statistical analysis, data
517 interpretation and wrote the original manuscript draft. J. Kremers critically reviewed and revised the
518 manuscript. All authors read and approved the final manuscript.

519 References

- 520 1. Robson AG, Frishman LJ, Grigg J, et al (2022) ISCEV Standard for full-field clinical
521 electroretinography (2022 update). *Doc Ophthalmol* 144:165–177.
522 <https://doi.org/10.1007/s10633-022-09872-0>
- 523 2. Davis CQ, Hamilton R (2021) Reference ranges for clinical electrophysiology of vision. *Doc*
524 *Ophthalmol* 143:155–170. <https://doi.org/10.1007/s10633-021-09831-1>
- 525 3. Birch DG, Anderson J (1992) Standardized Full-Field Electroretinography: Normal Values and
526 Their Variation With Age. *Arch Ophthalmol* 110:1571.
527 <https://doi.org/10.1001/archopht.1992.01080230071024>
- 528 4. Jones GRD, Haeckel R, Loh TP, et al (2018) Indirect methods for reference interval
529 determination – review and recommendations. *Clinical Chemistry and Laboratory Medicine*
530 (CCLM) 57:20–29. <https://doi.org/10.1515/cclm-2018-0073>
- 531 5. Ammer T, Schützenmeister A, Prokosch H-U, et al (2021) refineR: A Novel Algorithm for
532 Reference Interval Estimation from Real-World Data. *Sci Rep* 11:16023.
533 <https://doi.org/10.1038/s41598-021-95301-2>
- 534 6. Robson AG, Nilsson J, Li S, et al (2018) ISCEV guide to visual electrodiagnostic procedures.
535 *Doc Ophthalmol* 136:1–26. <https://doi.org/10.1007/s10633-017-9621-y>
- 536 7. Robin X, Turck N, Hainard A, et al (2011) pROC: an open-source package for R and S+ to
537 analyze and compare ROC curves. *BMC Bioinformatics* 12:77. [https://doi.org/10.1186/1471-](https://doi.org/10.1186/1471-2105-12-77)
538 [2105-12-77](https://doi.org/10.1186/1471-2105-12-77)
- 539 8. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3:32–35.
540 [https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3)
- 541 9. Ammer T, Schützenmeister A, Prokosch H-U, et al (2022) RIBench: A Proposed Benchmark for
542 the Standardized Evaluation of Indirect Methods for Reference Interval Estimation. *Clinical*
543 *Chemistry* 68:1410–1424. <https://doi.org/10.1093/clinchem/hvac142>
- 544 10. Ransohoff DF, Feinstein AR (1978) Problems of Spectrum and Bias in Evaluating the Efficacy
545 of Diagnostic Tests. *N Engl J Med* 299:926–930.
546 <https://doi.org/10.1056/NEJM197810262991705>
- 547 11. Verbakel SK, Van Huet RAC, Boon CJF, et al (2018) Non-syndromic retinitis pigmentosa.
548 *Progress in Retinal and Eye Research* 66:157–186.
549 <https://doi.org/10.1016/j.preteyeres.2018.03.005>

12. Koh AHC, Hogg CR, Holder GE The incidence of negative ERG in clinical practice
13. Audo I, Robson AG, Holder GE, Moore AT (2008) The Negative ERG: Clinical Phenotypes and Disease Mechanisms of Inner Retinal Dysfunction. *Survey of Ophthalmology* 53:16–40. <https://doi.org/10.1016/j.survophthal.2007.10.010>
14. Tanikawa A, Suzuki K, Nomura R, et al (2021) The influence of mild cataract on ISCEV standard electroretinogram recorded from mydriatic eyes. *Doc Ophthalmol* 142:177–183. <https://doi.org/10.1007/s10633-020-09791-y>
15. McAnany JJ, Persidina OS, Park JC (2022) Clinical electroretinography in diabetic retinopathy: a review. *Survey of Ophthalmology* 67:712–722. <https://doi.org/10.1016/j.survophthal.2021.08.011>
16. Arias-Alvarez M, Tomas-Grasa C, Sopena-Pinilla M, et al (2024) Electrophysiological findings in long-term type 1 diabetes patients without diabetic retinopathy using different ERG recording systems. *Sci Rep* 14:3520. <https://doi.org/10.1038/s41598-024-54099-5>
17. Polat OA, Gultekin M, Sener H, et al (2023) Retinal dysfunction in Parkinson’s disease—results of the extended protocol for photopic negative response (PHNR) full-field electroretinogram (ERG). *Doc Ophthalmol* 147:89–98. <https://doi.org/10.1007/s10633-023-09945-8>
18. Constable PA, Lim JKH, Thompson DA (2023) Retinal electrophysiology in central nervous system disorders. A review of human and mouse studies. *Front Neurosci* 17:1215097. <https://doi.org/10.3389/fnins.2023.1215097>
19. Jung R, Kempf M, Righetti G, et al (2024) Age-dependencies of the electroretinogram in healthy subjects. *Doc Ophthalmol* 149:99–113. <https://doi.org/10.1007/s10633-024-09991-w>
20. Kulyabin M, Kremers J, Holbach V, et al (2024) Artificial intelligence for detection of retinal toxicity in chloroquine and hydroxychloroquine therapy using multifocal electroretinogram waveforms. *Sci Rep* 14:24853. <https://doi.org/10.1038/s41598-024-76943-4>
21. Ma S, Yu J, Qin X, Liu J (2023) Current status and challenges in establishing reference intervals based on real-world data. *Critical Reviews in Clinical Laboratory Sciences* 60:427–441. <https://doi.org/10.1080/10408363.2023.2195496>
22. Ozarda Y, Ichihara K, Jones G, et al (2021) Comparison of reference intervals derived by direct and indirect methods based on compatible datasets obtained in Turkey. *Clinica Chimica Acta* 520:186–195. <https://doi.org/10.1016/j.cca.2021.05.030>
23. Ammer T, Schützenmeister A, Prokosch H-U, et al (2023) A pipeline for the fully automated estimation of continuous reference intervals using real-world data. *Sci Rep* 13:13440. <https://doi.org/10.1038/s41598-023-40561-3>
24. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18>