Validating Variational Bayes Linear Regression Method With Multi-Central Datasets

Hiroshi Murata,[1] Linda M. Zangwill,[2] Yuri Fujino,[1] Masato Matsuura,[1] Atsuya Miki,[3] Kazunori Hirasawa,[4,5] Masaki Tanito,[6] Shiro Mizoue,[7] Kazuhiko Mori,[8] Katsuyoshi Suzuki,[9] Takehiro Yamashita,[10] Kenji Kashiwagi,[11] Nobuyuki Shoji,[5] and Ryo Asaoka[1]

[1]Department of Ophthalmology, University of Tokyo Graduate School of Medicine, Tokyo, Japan
[2]Shiley Eye Institute Hamilton Glaucoma Center, University of California, San Diego, La Jolla, California, United States
[3]Department of Ophthalmology, Osaka University Graduate School of Medicine, Osaka, Japan
[4]Moorfields Eye Hospital NHS Foundation Trust and University College London, Institute of Ophthalmology, London, United Kingdom
[5]Orthoptics and Visual Science, Department of Rehabilitation, School of Allied Health Sciences, Kitasato University, Kanagawa, Japan
[6]Department of Ophthalmology, Shimane University Faculty of Medicine, Shimane, Japan
[7]Department of Ophthalmology, Ehime University Graduate School of Medicine, Ehime, Japan
[8]Department of Ophthalmology, Kyoto Prefectural University of Medicine, Kyoto, Japan
[9]Department of Ophthalmology, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan
[10]Department of Ophthalmology, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima, Japan
[11]Department of Ophthalmology, University of Yamanashi Faculty of Medicine, Yamanashi, Japan

Correspondence: Ryo Asaoka, Department of Ophthalmology, University of Tokyo Graduate School of Medicine, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan;
rasaoka-tky@umin.ac.jp.

Submitted: August 31, 2017
Accepted: March 4, 2018

Citation: Murata H, Zangwill LM, Fujino Y, et al. Validating variational Bayes linear regression method with multi-central datasets. *Invest Ophthalmol Vis Sci*. 2018;59:1897–1904. https://doi.org/10.1167/iovs.17-22907

Purpose. To validate the prediction accuracy of variational Bayes linear regression (VBLR) with two datasets external to the training dataset.

Method. The training dataset consisted of 7268 eyes of 4278 subjects from the University of Tokyo Hospital. The Japanese Archive of Multicentral Databases in Glaucoma (JAMDIG) dataset consisted of 271 eyes of 177 patients, and the Diagnostic Innovations in Glaucoma Study (DIGS) dataset includes 248 eyes of 173 patients, which were used for validation.

Prediction accuracy was compared between the VBLR and ordinary least squared linear regression (OLSLR). First, OLSLR and VBLR were carried out using total deviation (TD) values at each of the 52 test points from the second to fourth visual fields (VFs) (VF2–4) to 2nd to 10th VF (VF2–10) of each patient in JAMDIG and DIGS datasets, and the TD values of the 11th VF test were predicted every time. The predictive accuracy of each method was compared through the root mean squared error (RMSE) statistic.

Results. OLSLR RMSEs with the JAMDIG and DIGS datasets were between 31 and 4.3 dB, and between 19.5 and 3.9 dB. On the other hand, VBLR RMSEs with JAMDIG and DIGS datasets were between 5.0 and 3.7, and between 4.6 and 3.6 dB. There was statistically significant difference between VBLR and OLSLR for both datasets at every series (VF2–4 to VF2–10) ($P < 0.01$ for all tests). However, there was no statistically significant difference in VBLR RMSEs between JAMDIG and DIGS datasets at any series of VFs (VF2–2 to VF2–10) ($P > 0.05$).

Conclusions. VBLR outperformed OLSLR to predict future VF progression, and the VBLR has a potential to be a helpful tool at clinical settings.

Keywords: visual field, glaucoma, progression, variational Bayes, JAMDIG, The Japanese Archive of Multicentral Databases in Glaucoma, DIGS, the Diagnostic Innovations in Glaucoma Study

We previously proposed a new statistical model (variational Bayes linear regression [VBLR])[1] for predicting future visual fields (VFs) in glaucomatous patients. It is of importance to estimate VF progression rate in clinical settings, because glaucomatous VF defect is progressive and irreversible. Therefore, accurate prediction of future VF decay would contribute to appropriate medical or surgical intervention. Given that glaucoma is the second leading cause of blindness in the world[2] and that it could deteriorate quality of life, it is worthwhile improving it.

In the previous study, we reported the prediction performance was far better than that of ordinary squared linear regression, which means that the model successfully avoided overfitting. Nevertheless, there were some limitations. Although there was no overlapping between the training and test data, the VFs in the two data were retrieved at the same institution: the University of Tokyo Hospital. It is one thing that a model avoids overfitting with a specific dataset, but it is quite another that one is sufficiently generalized to a dataset from a completely different population and institution. In the training process of VBLR, the patterns of VF defects and progression were clustered applying Gaussian mixture model, which is a type of clustering method. Hence, in the prediction process, the forecast was thought to be performed according to a similar VF group. Therefore, if the clustering worked as expected, then the model would also function well on external and heterogeneous data. It is therefore important to determine the generalizability of VBLR, and how it works when it is trained with data at a single institution and applied to external data from a different institution and population.

The main purpose of this study is to validate the prediction accuracy of VBLR by training it with the data from the University of Tokyo and applying it to two external datasets. First, we applied VBLR to the Japanese Archive of Multicentral Databases in Glaucoma (JAMDIG),[3] excluding the data from the University of Tokyo Hospital, in order to compute the prediction accuracy. Next, we also applied it to Diagnostic Innovations in Glaucoma Study (DIGS) dataset,[4] which was thought to be more challenging. The patients recruited at the University of Tokyo Hospital comprised Asians, especially Japanese, and consequently, normal tension glaucoma (NTG) was prevalent.[5] In contrast, the DIGS dataset consists of glaucoma patients of European and African descent with most patients having primary open angle glaucoma with elevated intraocular pressure (IOP) and few patients with NTG. Therefore, by comparing the result of previous study with JAMDIG and DIGS datasets, we could show whether and how much degree the model is generalized.

Methods

This retrospective study was approved by the review board of each institute. Written consent was given by patients for their information to be stored in the hospital database and used for research. As to patients in JAMDIG data whose written consent was not given, their data were used in accordance with the regulations of the Japanese Guidelines for Epidemiologic Study 2008 issued by the Japanese Government. The study protocols did not require that each patient provide written informed consent, and instead, the protocol was posted at the outpatient clinic to notify participants of the study. This study was performed according to the tenets of the Declaration of Helsinki. As to DIGS data, the University of California San Diego Institutional Review Board approved the study methodologies, and all methods adhered to the Declaration of Helsinki guidelines for research in human subjects and the Health Insurance Portability and Accountability Act.

The VFs in training data were recorded at the University of Tokyo Hospital (Tokyo dataset) between 2002 and 2016. The data consisted of 7268 eyes of 4278 subjects; all of them received at least five VF tests, and the first ones were excluded from training in order to avoid learning effect.[6,7] Reliability criteria applied for training data were: fixation losses (FL) ≤ 33 %, false-positive responses (FP) ≤ 33 % and false-negative rate (FN) $\leq 33\%$.

Patients in two datasets were used for validation: JAMDIG and DIGS datasets. JAMDIG includes 1348 eyes of 805 primary open-angle glaucoma patients with 10 VFs measured with 24-2 or 30-2 Humphrey Field Analyzer (HFA), collected in 10 institutes in Japan. In the current study, the JAMDIG dataset was filtered according to the criteria as follows: (1) glaucoma was the only disease causing VF damage; (2) each patient had at least 11 VF measurements with 24-2 or 30-2 HFA II (Carl Zeiss Meditec, Inc., Dublin, CA, USA); (3) patients at the University of Tokyo Hospital were excluded; (4) patients' first VFs were excluded; and (5) VFs with FL $\geq 20\%$ and FP $\geq 15\%$ were excluded.

Regarding JAMDIG data, primary open-angle glaucoma was defined as (1) presence of typical glaucomatous changes in the optic nerve head such as a rim notch with a rim width ≤ 0.1 disc diameters or a vertical cup-to-disc ratio of >0.7 and/or a retinal nerve fiber layer defect with its edge at the optic nerve head margin greater than a major retinal vessel, diverging in an arcuate or wedge shape; and (2) gonioscopically wide open angles of grade 3 or 4 based on the Shaffer classification. Exclusion criterion were age below 20 years and possible secondary ocular hypertension in either eye.

As to DIGS data, the methodological details were described previously.[4] In brief, for DIGS glaucoma subjects recruited at the University of California San Diego Shiley Eye Institute, inclusion criteria were 20/40 or better best-corrected visual acuity, spherical refraction within ±5.0 diopters (D), cylinder correction within ±3.0 D, open-angles on gonioscopy, and at least two consecutive and reliable standard automated perimetry (SAP) examinations with either a pattern standard deviation (PSD) or a glaucoma hemifield test (GHT) result outside the 99% normal limits. Exclusion criteria were eyes with coexisting retinal disease and eyes with nonglaucomatous optic neuropathy. This dataset originally had 3583 eyes of 1913 patients and the criteria same to JAMDIG dataset was applied: (1) each patient had at least 11 VF measurements with 24-2 or 30-2 HFA II; (2) patients' first VFs were excluded; and (3) VFs with FL $\geq 20\%$ and FP $\geq 15\%$ were excluded.

Finally, test data consisted of 271 eyes of 177 patients for JAMDIG and 248 eyes of 173 patients for DIGS dataset. Test locations on the blind spot were excluded from the analyses. When a VF was measured using the 30-2 test pattern, only the 52 test points overlapping with the 24-2 test pattern were used.

Statistical Modeling

The statistical model of VBLR was described in detail previously.[1] In brief, let $\mathrm{t}_n{}^{\mathrm{T}} = (t_n{}^1, t_n{}^2, \cdots, t_n{}^{D_t})$ represent the total deviation (TD) values of a patient's nth VF in their series; D_t is the dimension of the vector t_n and is equal to 52 in this study. Let n_m be the set of indices of data obtained from the mth eye, T_m denotes the set $\{\mathrm{t}_n\}_{n\in n_m}$, while w_m is the parameter vector of the mth eye (where the first half and latter half of this vector include the intercept and slope coefficients of all 52 test VF points, respectively). Next, let x_n denote the interval from the first VF test of the nth data, $\Phi(x_n)$ denotes a matrix defined as $\Phi(x_n) = \begin{pmatrix} 1 \\ x_n \end{pmatrix} \otimes \mathrm{I}_{D_t}$ where I_{D_t} is a 52-dimensional identity matrix and $\otimes$ denotes Kronecker product; D_w is then the dimension of vector w_m (equal to 104 in this study). Then, λ_{m} represents the scalar of the magnitude of reliability of VFs obtained from the mth eye. A less strict criteria (33% FL and FP) was employed for training data to increase the size of the dataset and to better represent what happens in clinical practice, and λ_{m} could contribute to exploit data with less reliability. We assumed the data, t_n, were independently drawn from a Gaussian distribution with mean vector $\Phi(x_n)^{\mathrm{T}}\mathrm{w}_m$ and inverse of covariance matrix $\lambda_m^{-1}\mathrm{L}_m^{-1}$ where L_m is a 52 by 52 matrix. It is worthwhile to mention that L_m is not a diagonal matrix that enables this model to incorporate spatial and temporal correlation among test points. The likelihood is given by

$$p(\mathrm{T}_m|\mathrm{w}_m, \lambda_m, \mathrm{L}_m) = \prod_{n\in n_m} \mathcal{N}\left(\mathrm{t}_n|\Phi(x_n)^{\mathrm{T}}\mathrm{w}_m, \lambda_m^{-1}\mathrm{L}_m^{-1}\right). \qquad (1)$$

Moreover, we assumed w_m, λ_m, and L_m were random variables that followed a Gaussian mixture distribution, a Gamma mixture distribution, and a Wishart mixture distribu-

Table 1. The Demographic Data of the Three Datasets, and Is Described in Mean ± SD

	TOKYO	JAMDIG	DIGS
Number of eyes	7268 Eyes of 4278 Subjects	271 Eyes of 177 Patients	248 Eyes of 173 Patients
MD at baseline, dB	−6.7 ± 6.5	−7.1 ± 6.7	−4.0 ± 4.4
Follow-up, y	6.5 ± 2.9	5.3 ± 1.0	7.0 ± 2.7

TOKYO, the dataset at the University of Tokyo Hospital.

tion, respectively. The likelihoods were given by

$$p(\mathrm{L}_m|\zeta_m) = \prod_{h=1}^{H} \{\mathcal{W}(\mathrm{L}_m|v_h, \mathrm{W}_h)\}^{\zeta_{mh}} \qquad (2)$$

$$p(\zeta_m) = \prod_{h=1}^{H} \eta_h^{\zeta_{mh}} \qquad (3)$$

$$p(\mathrm{w}_m|\mathrm{z}_m, \lambda_m) = \prod_{k=1}^{K} \{\mathcal{N}(\mathrm{w}_m|\mu_k, \lambda_m^{-1}\Lambda_k^{-1})\}^{z_{mk}} \qquad (4)$$

$$p(z_m) = \prod_{k=1}^{K} \pi_k^{z_{mk}} \qquad (5)$$

$$p(\lambda_m|\gamma_m) = \prod_{g=1}^{G} \{\mathcal{G}(\lambda_m|a_g, b_g)\}^{\gamma_{mg}} \qquad (6)$$

and

$$p(\gamma_m) = \prod_{g=1}^{G} \theta_g^{\gamma_{mg}} \qquad (7)$$

where $z_{mk} \in \{0,1\}$, $\sum_{k=1}^{K} z_{mk} = 1$, $\zeta_{mh} \in \{0,1\}$, $\sum_{h=1}^{H} \zeta_{mh} = 1$, $\gamma_{mg} \in \{0,1\}$, $\sum_{g=1}^{G} \gamma_{mg} = 1$, $\eta_{\mathrm{h}} \in [0,1]$, $\sum_{h=1}^{H} \eta_h = 1$, $\pi_{\mathrm{k}} \in [0,1]$, $\sum_{k=1}^{K} \pi_k = 1$, $\theta_{\mathrm{g}} \in [0,1]$, $\sum_{g=1}^{G} \theta_g = 1$; K, H, and G are the number of components in each mixture distribution, $\mathcal{W}(\cdot|\cdot)$ denotes the Wishart distribution, and $\mathcal{G}(\cdot|\cdot)$ denotes the gamma distribution. K, H, and G were set to 15, 2, and 2, respectively. Hyperparameters larger than 15, 2, and 2 for K, H, and G lead to unstable computational result.

The training process was conducted by maximizing the marginalized log-likelihood of the training data

$$\ln p(\mathrm{T}) = \ln \int p(\mathrm{T}, \mathrm{w}, \lambda, \mathrm{L}, \zeta, \mathrm{z}, \gamma)\ \mathrm{dw}\ \mathrm{d}\lambda\ \mathrm{dL}\ \mathrm{d}\zeta\ \mathrm{dz}\ \mathrm{d}\gamma. \qquad (8)$$

using variational approximation for computing expectations.

Prediction Accuracy

Prediction accuracy was compared between the VBLR approach and ordinary least squared linear regression (OLSLR). First, OLSLR was carried out using TD values at each of the 52 test points from the second to the fourth VFs (VF2–4) of each patient, and the TD values of the 11th VF test were predicted. The same procedure was carried out using the TD values in different series: VF2–5, VF2–6, VF2–7, VF2–8, VF2–9, and VF2–10, and the TD values of 11th VFs were predicted every time. Likewise, TD values of 11th VFs were predicted with the VBLR approach using series of VFs from VF2–2 (the second VF only) to VF2–10 (all previous VFs). The aforementioned procedure was carried out on both JAMDIG and DIGS datasets. The predictive accuracy of each method was compared through the root mean squared error (RMSE) statistic, defined as follows:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{52} \frac{\begin{pmatrix}\text{predicted TD value of the } i\text{th point}\\ -\text{actual TD value of the } i\text{th point}\end{pmatrix}^2}{52}}. \qquad (9)$$

Likewise, prediction accuracy for mean TD (mTD) was also investigated using series for VFs from VF2–2 to VF2–10, and it was evaluated through absolute errors, which is defined as |predicted mTD value − actual mTD value|.

Software

Data preparation and analyses were carried out using the statistical programming language R version 3.0.3 (provided in the public domain by the R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/), and VBLR program was written in C++ using Armadillo C++ linear algebra library.[8]

Results

As shown in Table 1, the mean ± SD of initial mean deviation (MD) in Tokyo, JAMDIG, and DIGS datasets were –6.7 ± 6.5, –7.1 ± 6.7, and –4.0 ± 4.4, and follow-up period (the second VFs to the last ones) were 6.5 ± 2.9 5.3 ± 1.0, and 7.0 ± 2.7 years, respectively.

RMSEs of OLSLR from series of VFs from VF2–4 to VF2–10 for DIGS dataset were 31 ± 63, 15 ± 26, 9.8 ± 14, 7.0 ± 5.0, 5.5 ± 3.0, 4.7 ± 2.3, and 4.2 ± 1.9 dB, and those for JAMDIG were 19.5 ± 12.9, 11.8 ± 7.0, 8.5 ± 5.1, 6.5 ± 3.2, 5.2 ± 2.3, 4.4 ± 2.0, and 3.9 ± 1.7 dB, respectively. RMSEs of VBLR from series of VFs from VF2–2 to VF2–10 for JAMDIG were 5.2 ± 3.2, 4.9 ± 3.0, 4.7 ± 2.8, 4.6 ± 2.7, 4.4 ± 2.5, 4.3 ± 2.5, 4.1 ± 2.3, 3.9 ± 2.2, and 3.8 ± 2.1 dB (Fig. 1), and those for DIGS were 4.6 ± 2.4, 4.4 ± 2.3, 4.2 ± 2.3, 4.1 ± 2.2, 4.0 ± 2.1, 3.9 ± 2.0, 3.8 ± 2.0, 3.7 ± 1.9, and 3.6 ± 1.8 dB (Fig. 2), respectively (Table 2).

To compare RMSEs of OLSLR and VBLR, linear mixed model analysis and paired *t*-test were performed on DIGS and JAMDIG results, respectively. There was statistically significant difference between VBLR and OLSLR for both datasets at every series (VF2–4 to VF2–10) except for VF2–10 in JAMDIG ($P <$ 2.2e-16, < 2.2e-16, < 2.2e-16, 8.0e-11, 1.1e-5, 0.02, 0.47 for JAMDIG, and $P <$ 6.5e-11, 2.6e-10, 2.3e-10, < 2.2e-16, < 2.2e-16, < 2.2e-16, and 5.4e-16 for DIGS). However, there was no statistically significant difference in prediction performance of VBLR between JAMDIG and DIGS data at any series of VFs (VF2–2 to VF2–10).

Absolute errors for mTD of OLSLR from series of VFs from VF2–4 to VF2–10 for JAMDIG dataset were 5.4 ± 9.0, 3.2 ± 4.8, 2.7 ± 3.1, 2.0 ± 2.0, 1.8 ± 1.7, 1.4 ± 1.3, and 1.2 ± 1.1 dB, and those for DIGS were 10 ± 20, 5.3 ± 11.1, 3.6 ± 6.0, 2.6 ± 2.7, 2.2 ± 2.0, 1.8 ± 1.7, and 1.6 ± 1.4 dB, respectively (Figs. 3, 4). Absolute errors for mTD of VBLR from series of VFs

Table 2. RMSEs for Predicting 11th VFs (Point-Wise)

Method	VF2–2	VF2–3	VF2–4	VF2–5	VF2–6	VF2–7	VF2–8	VF2–9	VF2–10
JAMDIG									
OLSLR			20 ± 13	12 ± 7.0	8.5 ± 5.1	6.5 ± 3.2	5.2 ± 2.3	4.4 ± 2.0	3.9 ± 1.7
VBLR	5.2 ± 3.2	4.9 ± 3.0	4.7 ± 2.8	4.6 ± 2.7	4.4 ± 2.5	4.3 ± 2.5	4.1 ± 2.3	3.9 ± 2.2	3.8 ± 2.1
P value			< 2.2e-16	< 2.2e-16	< 2.2e-16	8.0e-11	1.1e-5	0.02	0.47
DIGS									
OLSLR			31 ± 63	15 ± 26	9.8 ± 14	6.5 ± 3.27.0 ± 5.0	5.2 ± 2.35.5 ± 3.0	4.7 ± 2.3	4.2 ± 1.9
VBLR	4.6 ± 2.4	4.4 ± 2.3	4.2 ± 2.3	4.1 ± 2.2	4.0 ± 2.1	3.9 ± 2.0	3.8 ± 2.0	3.7 ± 1.9	3.6 ± 1.8
P value			6.5e-11	2.6e-10	2.3e-10	< 2.2e-16	< 2.2e-16	< 2.2e-16	5.4e-16

The RMSEs were described in mean ± SD. *P* values were obtained by comparing between OLSLR and VBLR.

from VF2–2 to VF2–10 for JAMDIG dataset were 2.3 ± 2.6, 2.1 ± 2.3, 1.9 ± 2.1, 1.8 ± 2.0, 1.7 ± 1.8, 1.6 ± 1.7, 1.5 ± 1.6, 1.4 ± 1.4, and 1.3 ± 1.3 dB, and those for DIGS were 2.3 ± 2.1, 2.1 ± 2.0, 2.0 ± 2.0, 1.9 ± 1.9, 1.8 ± 1.8, 1.7 ± 1.7, 1.6 ± 1.6, 1.5 ± 1.5, and 1.4 ± 1.4 dB, respectively (Table 3).

Figure 5 shows spatial pattern of predicting 11th VFs for JAMDIG and DIGS datasets, and Figure 6 shows the relation between changes of 11th VFs from initial VFs and RMSEs. In all series of VFs for DIGS, there were statistically significant correlations between changes of 11th VFs from initial VFs and RMSEs ($P < 0.05$), while correlations did not reach statistical significance for all series of JAMDIG ($P > 0.05$). Likewise, Figure 7 shows the relationship between initial mTD (the second VFs) and RMSEs that represent the association between severity of glaucoma and prediction performance, and in all series of VFs, there were statistically significant correlations ($P < 0.05$). However, as Figure 8 shows, there was no significant correlation between mean of raw error values that represents the discrepancy between real VFs and prediction, and initial mTD ($P > 0.05$) except for VF2–2 and VF2–3 in DIGS ($P = 0.01$ and 0.02). Furthermore, the regression lines between raw prediction error and initial mTD with VF2–2 and VF2–3 were near horizontal.

Discussion

In the current study, prediction accuracy of VBLR was assessed with JAMDIG and DIGS data, compared with that of OLSLR. OLSLR is commonly used for assessment and estimation of glaucomatous VF progression, and that is the reason we compared VBLR and OLSLR. RMSEs in the previous study for VF2–2 to VF2–10 were 5.3 ± 2.8, 5.0 ± 2.7, 4.9 ± 2.6, 4.7 ± 2.5, 4.5 ± 2.4, 4.4 ± 2.3, 4.2 ± 2.2, 4.1 ± 2.1, and 3.9 ± 2.1 dB. Therefore, prediction accuracy in this study was better (smaller) than what we reported previously. Though one of the reasons could be ascribed to the size of the training data used in the two studies, it is of importance that prediction accuracy in this study was computed using external datasets. Consequently, VBLR could perform well on heterogeneous dataset as well, because VBLR was trained with the data retrieved only from the University of Tokyo Hospital. However, it should be addressed that there is a risk to predict future VFs by extrapolation outside the range of explanatory variable used to build the model, for example using less than 11 VFs to predict 11th VF, though this extrapolation is adopted in clinical settings, such as in Humphrey Guided Progression Analysis software. A possible caveat of the VBLR is that it assumes linear progression of VF damage, similarly to OLSLR. A previous study suggested the application of nonlinear regression, such as exponential regression, in particular when floor effect is concerned,[9] however, the merit would be only marginal, if any, because our previous study showed linear regression models outperformed nonlinear models, in terms of prediction accuracy.[10] It should be noted that the statistical significance of progression cannot be calculated with nonlinear regression, which limits the clinical usefulness of nonlinear regressions.

Though there was no statistically significant difference in RMSEs in JAMDIG and DIGS data, it was of surprise that the performance using DIGS data was better than that of JAMDIG data, because JAMDIG data consisted of mostly Asians and very similar to Tokyo data, while DIGS data mainly consisted of non-Asians. Moreover, the mean follow-up period of DIGS data was longer than that of JAMDIG data, which could have led to worse performance, but the reverse was true. In contrast, OLSLR performed better on JAMDIG data. The discrepancy was confounding, but it suggests, at least, VBLR model is generalizable to both external and heterogeneous data. On the other hand, as shown in Figure 6, prediction performance deteriorated in cases with a large difference between the initial and final VFs, as in the DIGS dataset. This finding was not observed in the JAMDIG dataset. These contradicting results would presumably be ascribed to the different dataset populations; the TOKYO training set and JAMDIG test datasets are obtained in Japan, whereas DIGS test dataset was collected outside the country. Performance can be expected to be better when the training and test datasets are similar. A further study would be needed to investigate whether similar results are obtained when VBLR is trained using heterogeneous data from different countries.

Table 3. Absolute Errors for Predicting 11th VFs (Mean TD)

Method	VF2–2	VF2–3	VF2–4	VF2–5	VF2–6	VF2–7	VF2–8	VF2–9	VF2–10
JAMDIG									
OLSLR			5.4 ± 9.0	3.2 ± 4.8	2.7 ± 3.1	2.0 ± 2.0	1.8 ± 1.7	1.4 ± 1.3	1.2 ± 1.1
VBLR	2.3 ± 2.6	2.1 ± 2.3	1.9 ± 2.2	1.8 ± 2.0	1.7 ± 1.8	1.6 ± 1.7	1.5 ± 1.6	1.4 ± 1.4	1.3 ± 1.3
DIGS									
OLSLR			10 ± 20	5.3 ± 11.1	3.6 ± 6.0	2.6 ± 2.7	2.2 ± 2.0	1.8 ± 1.7	1.6 ± 1.4
VBLR	2.0 ± 2.4	1.9 ± 2.2	1.8 ± 2.0	1.7 ± 1.9	1.6 ± 1.8	1.5 ± 1.7	1.4 ± 1.5	1.3 ± 1.4	1.3 ± 1.3

The absolute errors were described in mean ± SD.

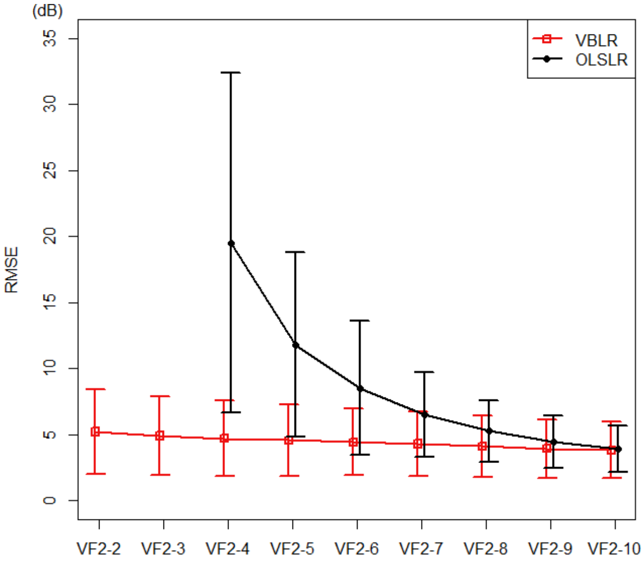

Figure 1. RMSEs for predicting 11th VFs (point-wise, JAMDIG). RMSEs for JAMDIG dataset. *Red line*: VBLR; *black line*: OLSLR. *Error bars* show standard deviations.

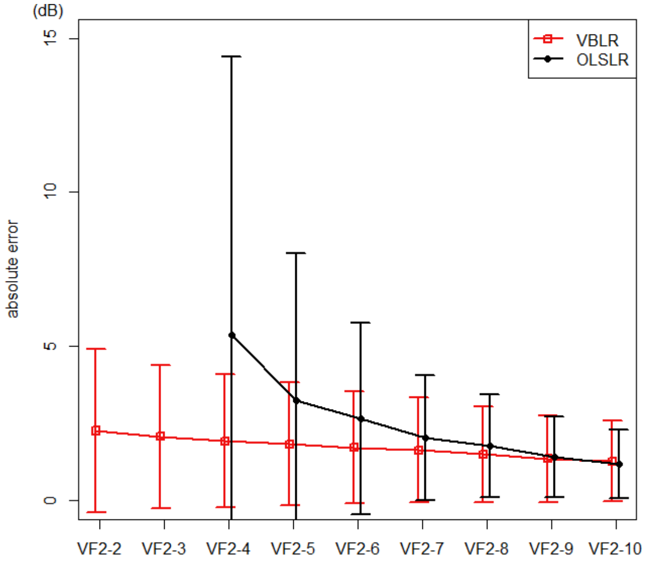

Figure 3. Absolute errors for predicting 11th VFs (mean TD, JAMDIG). Absolute errors for JAMDIG dataset predicting 11th VFs. *Red line*: VBLR; *black line*: OLSLR. *Error bars* show standard deviations.

One of the advantages of Bayes statistics is that it can exploit information of existing data by computing posterior distribution based on prior distribution obtained beforehand. In clinical settings, glaucoma specialists determine medical strategy based on their experiences, which are analogous to prior information in Bayes statistics. Therefore, what Bayes statistics do is very similar to what clinicians do. We proposed VBLR previously, and one of the state of the art methods reported by other groups based on Bayes statistics is Analysis with Non-Stationary Weibull Error Regression and Spatial Enhancement (ANSWERS).[11]

However, in Bayes statistics, since the posterior distribution is computed according to prior distribution, the performance is thought to be influenced by training data. Hence, applying it to extraneous data could detract from prediction performance. Indeed, primary open-angle glaucoma with normal and elevated IOP has different patterns of VF defects.[12–14] In the previous and this study, VBLR was trained with the data only at the University of Tokyo Hospital, which means that the data mostly consisted of Asians and the prevalent type of glaucoma was NTG,[5] nonetheless the diagnostic/predicting performance in an external DIGS dataset obtained in United States was at least no worse than that in JAMDIG dataset collected in Japan.

In VBLR, mixture of Gaussian model is incorporated, and thence, spatial and temporal patterns of VF defects are clustered in the training phase. Hence, future VFs are predicted using similar groups based on spatial and temporal characteristics. Hypothetically, clustering would contribute to improvement of performance even on external data, and this was the main motivation of this study.

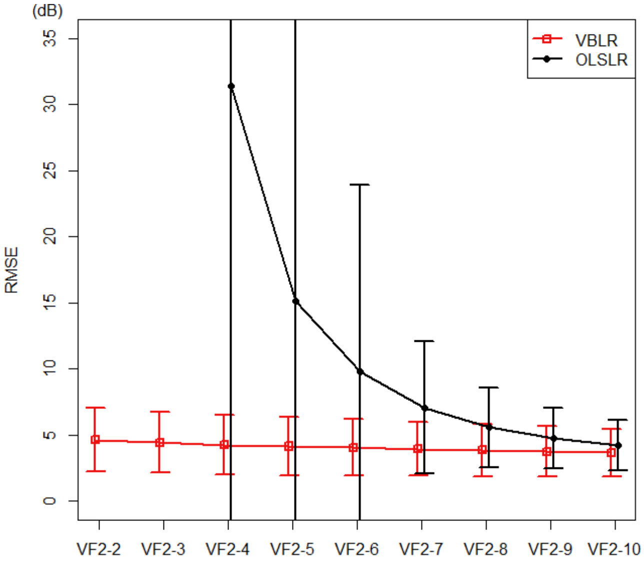

Figure 2. RMSEs for predicting 11th VFs (point-wise, DIGS). RMSEs for DIGS dataset. *Red line*: VBLR; *black line*: OLSLR. *Error bars* show standard deviations.

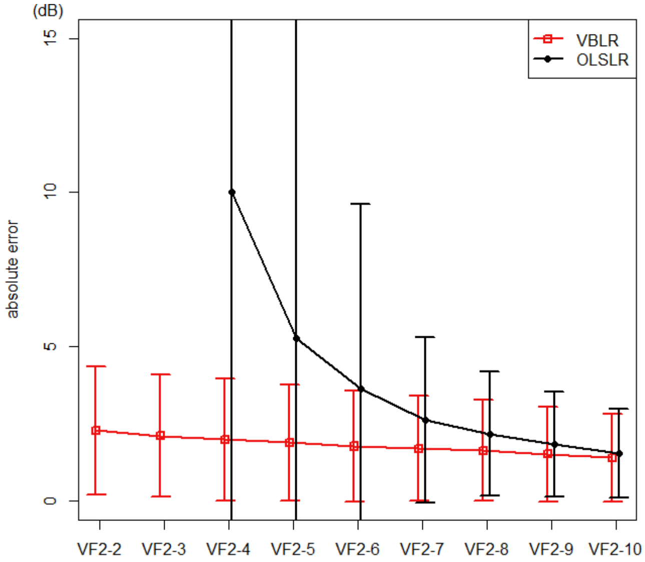

Figure 4. Absolute errors for predicting 11th VFs (mean TD, DIGS). Absolute errors for DIGS dataset predicting 11th VFs. *Red line*: VBLR; *black line*: OLSLR. *Error bars* show standard deviations.

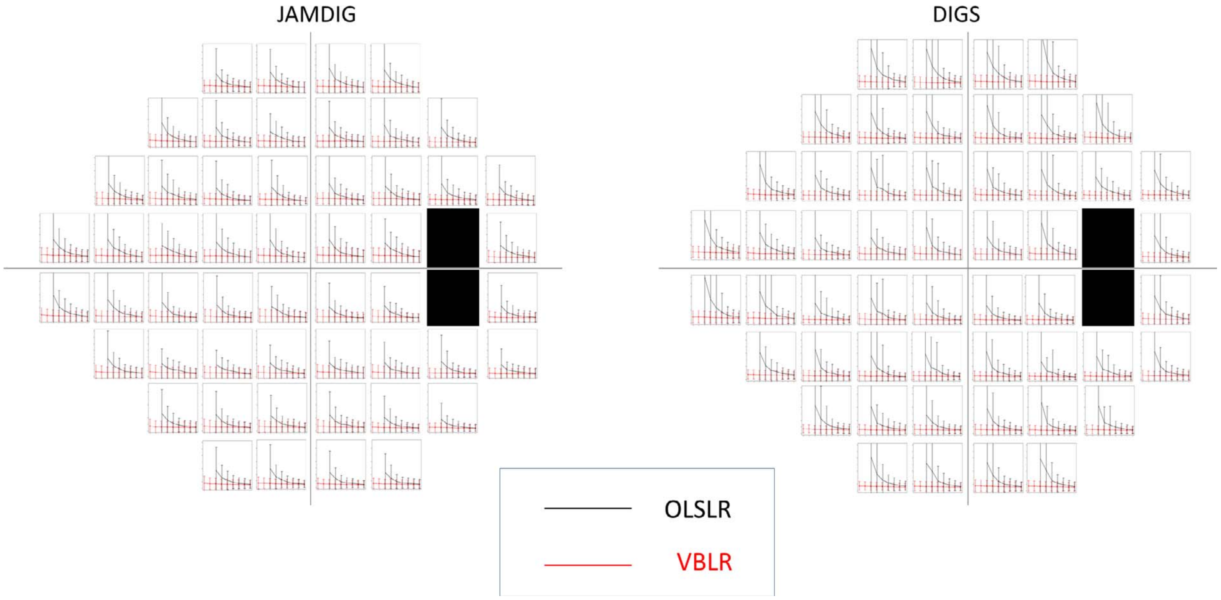

FIGURE 5. Spatial patterns of prediction errors. Spatial patterns of absolute errors for JAMDIG and DIGS dataset predicting 11th VFs. Similar to Figures 3 and 4, *x* axis denotes series of VFs used for predicting 11th VFs. The test grid for left eyes was flipped to show both eyes in the same diagram.

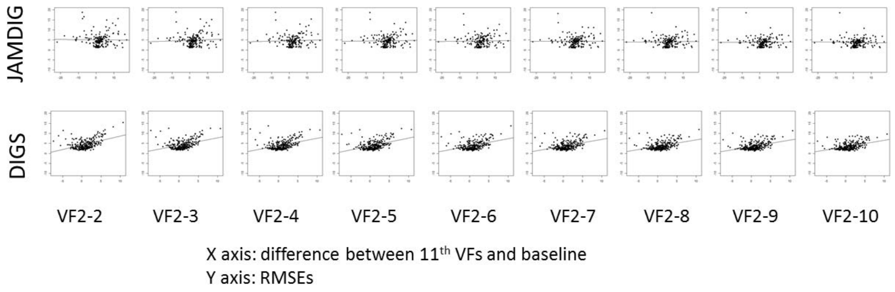

FIGURE 6. The relation between changes from the initial VFs and RMSEs. The relation between changes of 11th VFs from the initial VFs and prediction errors (VF2–2 to VF2–10). There were statistically significant correlations in all series in DIGS ($P < 0.05$), while there was not in all series in JAMDIG ($P > 0.05$). *Black lines* are regression lines.

One of the caveats of this study is that all the patients in the training data had more than five VFs, and accordingly, they were relatively stable in terms of glaucomatous VF progression. For example, if a patient has extremely high IOP and ends up blind in a short period of time, they would not be included in training data. Fortunately, this situation is relatively rare. Furthermore, patients in the test datasets had at least 11 VFs without surgical intervention, so they may be also relatively stable ones. Therefore, it is worthwhile mentioning that prediction accuracy for extreme cases has not been well investigated in this study. In addition, because Bayes methods update prior information with posterior ones, the predicted result with relatively small amount of VFs would be based on average patients. As shown in Figure 7, RMSEs of VBLR

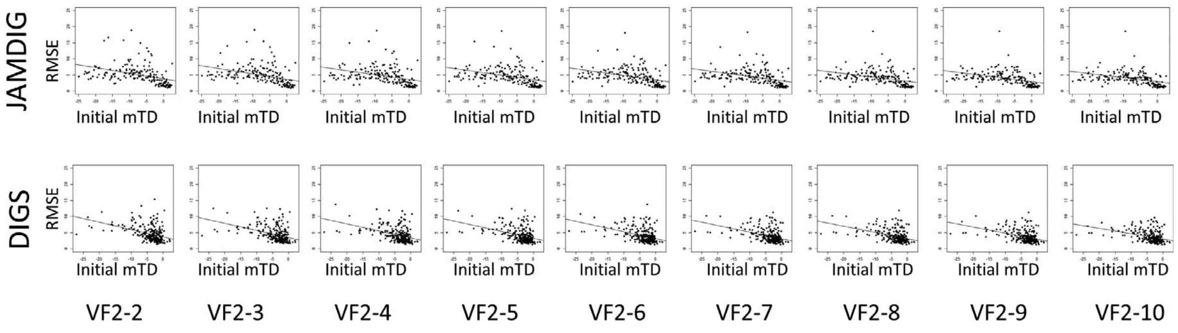

FIGURE 7. The association between initial mTD and RMSEs. The association between initial mTD (the second VFs) and RMSEs for VF2–2 to VF2–10. There were statistically significant correlations between mTD slope and prediction errors ($P < 0.05$) for both datasets. *Black lines* are regression lines.

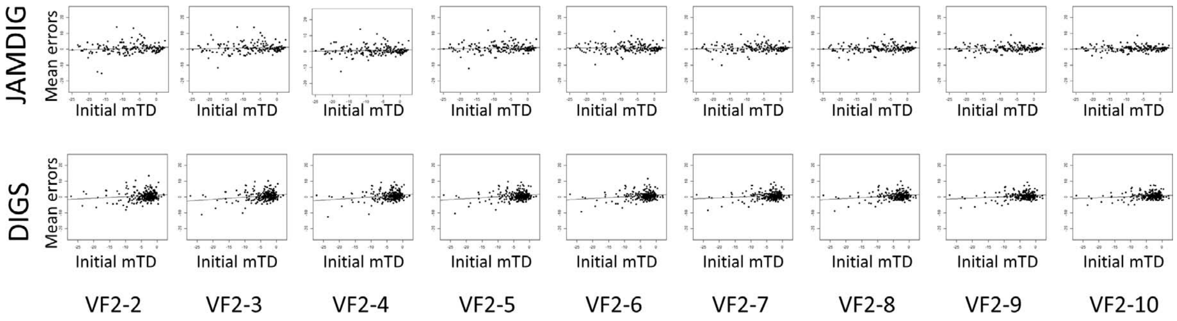

FIGURE 8. The relation between initial mTD and mean errors. The relation between initial mTD and mean errors (the difference between prediction and actual values). There was no statistically significant correlation except for VF2–2 ($P = 0.01$) and VF2–3 ($P = 0.02$) in DIGS. *Black lines* are regression lines.

depends on the stage of glaucoma, while there was little bias (under- or overestimation) depending on the severity of disease as shown in Figure 8. This may suggest the reason of the discrepancy between RMSE, which represents prediction errors including test variability (Fig. 7), and mean error (Fig. 8) that represents the discrepancy between real VFs and prediction is just the high variability of VF tests where glaucomatous damage is advanced.[15]

There is no doubt elevated IOP is a risk factor for the development and progression of glaucoma, as suggested by numerous previous papers.[16] However, in our recent study with the JAMDIG data, it was indicated that mean IOP was not associated with progression of VF damage,[17] probably because most of the patients in the JAMDIG dataset were already medically intervened and the mean IOP was within an appropriate and tight range. Indeed, we have recently proposed a novel method of regressing VF against IOP integrated time, instead of time, using the JAMDIG data.[18] As a result, significant improvement of prediction accuracy was observed, but the magnitude of the improvement was small and its impact on the real clinical settings was almost negligible. Thus, achieving improvement of VF progression prediction by applying VBLR, although it cannot reflect IOP status, will be a clinically useful approach when assessing VF progression of glaucoma patients.

In conclusion, the performance of VBLR was far better than that of OLSLR. Though there are some limitations in VBLR, VBLR would have potential to be a helpful tool for clinical settings compared with OLSLR based method, such as PROGRESSOR (Medisoft Ltd., Leeds, UK).[19]

Acknowledgments

Supported by Grants 17K11418 and 17K16959 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, Japan Science, and Technology Agency (JST), CREST JPMJCR1304, and Daiichi Sankyo Foundation of Life Science. DIGS was funded by National Eye Institute Grants EY011008, EY14267, EY019869, EY201818, and Core Grant P30EY022589.

Disclosure: **H. Murata**, None; **L.M. Zangwill**, None; **Y. Fujino**, None; **M. Matsuura**, None; **A. Miki**, None; **K. Hirasawa**, None; **M. Tanito**, None; **S. Mizoue**, None; **K. Mori**, None; **K. Suzuki**, None; **T. Yamashita**, None; **K. Kashiwagi**, None; **N. Shoji**, None; **R. Asaoka**, None

References

1. Murata H, Araie M, Asaoka R. A new approach to measure visual field progression in glaucoma patients using variational Bayes linear regression. *Invest Ophthalmol Vis Sci*. 2014;55: 8386–8392.
2. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90:262–267.
3. Fujino Y, Asaoka R, Murata H, et al. Evaluation of glaucoma progression in large-scale clinical data: the Japanese Archive of Multicentral Databases in Glaucoma (JAMDIG). *Invest Ophthalmol Vis Sci*. 2016;57:2012–2020.
4. Sample PA, Girkin CA, Zangwill LM, et al. The African Descent and Glaucoma Evaluation Study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127:1136–1145.
5. Iwase A, Suzuki Y, Araie M, et al. The prevalence of primary open-angle glaucoma in Japanese: the Tajimi Study. *Ophthalmology*. 2004;111:1641–1648.
6. Werner EB, Krupin T, Adelson A, Feitl ME. Effect of patient experience on the results of automated perimetry in glaucoma suspect patients. *Ophthalmology*. 1990;97:44–48.
7. Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol*. 1996;114:19–22.
8. Armadillo SC. *An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments*. Sydney, Australia: National Insurance Crime Training Academy (NICTA); 2010.
9. Caprioli J, Mock D, Bitrian E, et al. A method to measure and predict rates of regional visual field decay in glaucoma. *Invest Ophthalmol Vis Sci*. 2011;52:4765–4773.
10. Taketani Y, Murata H, Fujino Y, Mayama C, Asaoka R. How many visual fields are required to precisely predict future test results in glaucoma patients when using different trend analyses? *Invest Ophthalmol Vis Sci*. 2015;56:4076–4082.
11. Zhu H, Russell RA, Saunders LJ, Ceccon S, Garway-Heath DF, Crabb DP. Detecting changes in retinal function: analysis with non-stationary Weibull error regression and spatial enhancement (ANSWERS). *PLoS One*. 2014;9:e85654.
12. Chauhan BC, Drance SM, Douglas GR, Johnson CA. Visual field damage in normal-tension and high-tension glaucoma. *Am J Ophthalmol*. 1989;108:636–642.
13. Caprioli J, Sears M, Spaeth GL. Comparison of visual field defects in normal-tension glaucoma and high-tension glaucoma. *Am J Ophthalmol*. 1986;102:402–404.
14. Drance SM, Douglas GR, Airaksinen PJ, Schulzer M, Hitchings RA. Diffuse visual field loss in chronic open-angle and low-tension glaucoma. *Am J Ophthalmol*. 1987;104:577–580.
15. Henson DB, Chaudry S, Artes PH, Faragher EB, Ansons A. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Invest Ophthalmol Vis Sci*. 2000;41:417–421.

16. Garway-Heath DF, Crabb DP, Bunce C, et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. *Lancet*. 2015;385:1295–1304.

17. Fujino Y, Asaoka R, Murata H, et al. Evaluation of glaucoma progression in large-scale clinical data: The Japanese Archive of Multicentral Databases in Glaucoma (JAMDIG). *Invest Ophthalmol Vis Sci*. 2016;57:2012–2020.

18. The Japanese Archive of Multicentral Database in Glaucoma (JAMDIG) Construction Group. A novel method to predict visual field progression more accurately, using intraocular pressure measurements in glaucoma patients. *Sci Rep*. 2016; 6:31728.

19. Fitzke FW, Hitchings RA, Poinoosawmy D, McNaught AI, Crabb DP. Analysis of visual field progression in glaucoma. *Br J Ophthalmol*. 1996;80:40–48.