Statistics in Medicine

Research Article

Received: 15 December 2015, Accepted: 3 July 2016 Published online 24 July 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7056

Integrating independent spatio-temporal replications to assess population trends in disease spread

John VanBuren,[a*†] Jacob J. Oleson,[a] Gideon K. D. Zamba[a] and Michael Wall[b]

Glaucoma is the second leading cause of blindness in the USA. A visual field test (perimetry) is used to sample and quantitate visual field function in preselected regions in the eye. These regions can be considered a spatial field with replications across independently measured individuals. At return visits, a new set of visual field measurements is obtained producing a subject specific spatio-temporal dataset. We develop a Bayesian hierarchical modeling framework to analyze these spatio-temporal datasets both for individual level spread and as aggregate population level trends. Our model extends previous research utilizing a dimension reduction matrix and individual specific latent variables. Human characteristics are incorporated into the model to help explain glaucoma progression. One beneficial product of our model is smoothed estimates for individuals. We also specify how progression rates are computed for monitoring purposes so that clinicians can track changes and predict forward in time. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: Bayesian; hierarchical models; dimension reduction; glaucoma

1. Introduction

Models that adequately estimate and forecast the spread of disease over both space and time are important in many disciplines including epidemiology, ecology, biostatistics, applied mathematics, and geography. Spatio-temporal modeling has been used to model many epidemics and have been increasing in popularity over the years [1]. Each passing year brings more mathematical tools being developed and computational efficiency increases, which allows for more model complexity. With the abundance of large data collected today, statistical approaches need to evolve to handle new complex scenarios. One such scenario is when diseases spread over time throughout a person's body or specific organ. In this paper, we demonstrate how a spatio-temporal model can be used to evaluate population level trends when diseases progress at different rates between independent subjects.

Three important methods for modeling spatio-temporal disease spread are stochastic compartmental models (e.g., SIR models), diffusion models, and dimension reduction models. Kermack and McKendrick [2] used deterministic systems of differential equations, which have served as the basis behind the development of stochastic compartmental models. These models excel when locations are clearly defined and objects or people can be categorized into specific compartments [3–8]. Ecological diffusion differential equations describe the spatio-temporal process of a disease allowing for drastic dissimilarities between neighbors [9]. Hooten, Garlick, and Powell [9] introduced multi-scale homogenization methods based on an ecological diffusion model that can approximate partial differential equations by smoothing the noise process. This is computationally advantageous over traditional diffusion models first introduced by Whittle [10] at the price of estimation accuracy, which is useful if the underlying process cannot be modeled due to its excessive size. In general, ecological diffusion models are strategic when data are collected over a large number of time points, which is often not the case in biomedical studies [11–13].

[a] *Department of Biostatistics, The University of Iowa, Iowa City, IA, U.S.A.*
[b] *Department of Ophthalmology and Visual Sciences, The University of Iowa, Iowa City, IA, U.S.A.*
**Correspondence to: John VanBuren, Department of Biostatistics, The University of Iowa, Iowa City, IA, U.S.A.*
[†] *E-mail: jvanburen88@gmail.com*

Copyright © 2016 John Wiley & Sons, Ltd.

Dimension reduction models utilize the specification of basis functions that are numerical vectors that describe a specific proportion of the data or process variability. There are numerous methods that can be employed to specify basis functions, and each technique has its own benefits. Moran's operator has been used to determine positive and negative spatial dependences in spatial generalized linear mixed models [14,15]. This utilizes the pre-specified adjacency matrix in determining the basis vectors. Oleson and Wikle [16] used a pre-specified number of empirical orthogonal functions (EOF) on the spatio-temporal dataset as their basis functions along with latent expansion coefficients to predict the data. EOFs are optimal in describing the most variability among any possible orthogonal sets of basis functions with the same number of vectors [1]. Regardless of the method used in the creation of a basis function, dimension reduction models utilize a subset of the potential basis vectors to minimize the parameters needed.

When examining disease spread, it is not typical to have a single occurrence of the spread. In other words, not all diseases are spread only once, as many are seasonal as evidenced by influenza or strep throat [17]. In those cases, a reinfection process may be included in the model that incorporates a direct relationship between the different periods [6]. It is also important to note that different spatial regions may have different disease transmission rates, but the regions are not truly independent of each other as living objects are not stationary and infectious hosts often interact with susceptible individuals across locations [5]. For our research, we are interested in modeling the independent replication of disease spread across locations where the location is the spread within the body or organ. We use a model to describe how disease progression is related to an overall population effect via subject level baseline covariates. Our model extends upon the work of Oleson and Wikle [16] in a Bayesian hierarchical format to include a dimension reduction matrix along with latent variables in order to capture the spatio-temporal population trend.

The method is developed in the context of glaucoma progression. Wall, Woodward, Doyle, and Artes [18] discussed research designed to study the mechanism of perimetric variability. In their study, visual field measurements were tested and retested covering the central 24° of the visual field by using the Humphrey Field Analyzer. The visual field is gridded on a two-dimensional scale into 54 different regions as shown in Figure 1. Patients are scheduled to return for repeat testing every 6 months for 4 years. The resulting data are a spatial field with measurements that change with each visit, yielding a within-person spatio-temporal dataset. On a population level, we can use these between-subject spatio-temporal replications of independent participants to determine how human characteristics and demographics can collectively affect glaucomatous progression. Clinically, this research can help determine population subgroups with a higher risk of glaucoma progression. In addition, the statistical model will help determine which visual field areas are prone to progression.

Previous research presented six distinct nerve fiber bundle zones within the central 24° of the visual field that are thought to be important in assessing glaucoma progression [19]. We will use a confirmatory factor analysis technique on spatio-temporal data to relate the observed data to those nerve fiber bundle regions. Finally, once the population model parameters are estimated, the model is used to predict visual field measurements for an individual. Rates of change in an individual at each location can be modeled through the smoothed temporal estimates to calculate individual glaucoma progression rates.

In this paper, we introduce the motivating example behind our research in Section 2. Subsequently, we discuss previous work by other researchers to build the foundation for our Bayesian hierarchical model in Section 3. This leads to simulation results in Section 4, and we apply the model to a dataset

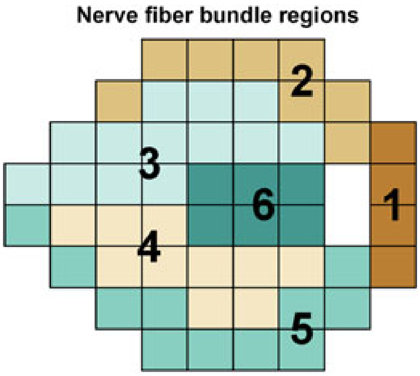

Figure 1. Nerve fiber bundle regions as defined by Garway-Heath et. al.

Copyright © 2016 John Wiley & Sons, Ltd.

of participants with glaucoma in Section 5. To conclude, we discuss the findings and future steps in Section 6.

2. Motivating study

The motivation behind this model comes from a dataset collected from the Mechanisms of Perimetric Variability grant studying glaucoma at the University of Iowa. Glaucoma is a condition of the eye that damages the optic nerve and is a leading cause of blindness [20]. Retinal ganglion cells are neurons that run through the optic nerve to connect to the brain, which provides vision to a person. In the eye, there are three different sources of pressure on the optic nerve: the intraocular pressure, the cerebrospinal fluid pressure, and the arterial blood pressure. For the optic nerve to remain healthy and allow the retinal ganglion cells to properly function, the interaction of these three pressures needs to be balanced. In many glaucoma cases, it appears that an increase in intraocular pressure damages is the main factor leading to optic nerve damage.

The study enrolled 120 participants with glaucoma and 60 participants with normal vision. Both eyes were assessed upon enrollment to determine which visual field(s) met inclusion criterion. If only one visual field met the criterion, it was measured at the remaining follow-up visits. If both visual fields met the criterion, one was randomly chosen and it was measured at all follow-up visits. The visual grid on the left eye is flipped to the form of the right eye so visual grid locations correspond to identical locations regardless of the eye measured. All data are presented in the right eye form. Visual field measurements were collected covering the central 24° of the visual field using the Humphrey Field Analyzer standard automated perimetry size III. This central 24° section of the eye is gridded on a two-dimensional scale into 54 locations. The visual field measurements can be measured at each of the 54 locations at a single time point per individual. The raw data represented a ratio of background light intensity to stimulus light intensity. An inverse $\log_{10}$ transformation was implemented to convert the raw data to a decibel measurement, the latter ranging from 0 to 40, which makes it easier for clinicians to process. Locations 26 and 35 on the grid mapped to physiological blind spots in the visual field so they were removed from all analyses.

Two baseline visits were completed within 8 weeks of each other and then were averaged together at each location to produce a more stable measurement. After baseline, follow-up measurements were collected every 6 months for 4 years, or eight total follow-up visits. When an individual is routinely checked, his or her data form a unique spatio-temporal dataset ranging from a baseline visit to the most recent checkup (time, T). Across all participants, population estimates can be calculated on explanatory variables using all the participants' data.

Explanatory variables of interest included glaucoma type, gender, surgery status, mean deviation category, an indicator for optic disc hemorrhage, baseline age, and the baseline average retinal thickness. Glaucoma type was categorized into three groups: Primary Open-Angle Glaucoma, Normal-Tension Glaucoma, and Other (includes angle closure glaucoma, mixed-mechanism glaucoma, pigmentary glaucoma, pseudoexfoliative glaucoma, traumatic glaucoma, and uveitic glaucoma). Surgery status is an indicator of whether or not a person had surgery after being diagnosed with glaucoma. In most cases, trabeculectomy was the surgical procedure performed, but specific details of the surgery were not available. Mean deviation is a measure of average loss across the visual field with some weighting for the central locations. These values are categorized into mild and moderate. Age and baseline average retinal thickness were centered prior to performing the analyses. Baseline demographics of the participants are presented in Table III. There were 120 people enrolled in the study. Only 52 of them have complete data and were used in the analysis. The other 68 individuals were used to construct the dimension reduction matrix, which is described in Section 5.1.

3. Bayesian hierarchical model

3.1. Model introduction

The modeling framework of Hooten and Wikle [11] and Oleson and Wikle [16] serve as the basis for our proposed model. Our Dimension Reduction On Independent Infectious Disease Systems (DROIIDS) model is an extension of these models. Both previous models focus on a single occurrence of the disease spread rather than many independent replications of the process. Our problem of interest contains spatio-temporal processes over identical regions replicated on independent people. This will allow for

Copyright © 2016 John Wiley & Sons, Ltd.

population averaged interpretation to identify the underlying parameters associated with glaucoma progression. In the hierarchical Bayesian framework chosen here, the data model relates the observed data to a set of unobserved latent variables through a combination of the specified likelihood and a set of prior distributions. The process model describes how the latent variables interact through one or more sets of equations. Parameter models contain prior distributions which play a crucial role in the forming of full conditionals which are used in the estimation process. We describe the three subset models (i.e., Data, Process, and Parameter) in the DROIIDS model in the next three sections.

3.2.1. Model definition (data model). The outcome score, z_{itl}, is the assumed normally distributed decibel measurement described in Section 2 for person $i=1,\ldots,N$; at time point t, $t=1,\ldots,T$ and location $l=1,\ldots,L$. Let $\boldsymbol{z}_{i,t}$ be a vector containing all the locations observed values ($\boldsymbol{z}_{i,t}=(z_{i,t,1},\ldots,z_{i,t,l}\ldots,z_{i,t,L})^{'}$).The following data model is considered for outcome measurement $\boldsymbol{z}_{i,t}$.

$$\text{Data Model}: \qquad \boldsymbol{z}_{i,t} = \boldsymbol{\Phi a}_{i,t} + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{i,t} \tag{1}$$

Three components are included in this level: a dimension reduction portion ($\boldsymbol{\Phi a}_{i,t}$), a covariate portion ($\boldsymbol{X}_i\boldsymbol{\beta}$), and a vector of location-specific errors ($\boldsymbol{\varepsilon}_{i,t}$). The full data for an individual, $\boldsymbol{z}_i$, are structured in an $L\times T$ matrix. The rows within column t are the L-correlated spatial observations at time t. For each row, l, there are T successive and regularly spaced measurements collected across visits.

Propagator matrices, or transition matrices, contain loadings that relate an individual's latent variables to the observed data [1]. Latent variables are unobserved variables to which we will return shortly. In our model, the propagator matrix, $\boldsymbol{\Phi}$, allows for dimension reduction in the data through the specification of a subset of basis functions, letting us describe the variability through fewer parameters. Basis functions are mathematical representations of the observed data simplified into linear vectors. There are many different methods to create basis functions from the data such as orthogonal polynomials, generalized Moran, EOFs, and factor analyses (see Cressie and Wikle [1] for details). To produce EOF basis vectors and values, eigenvalue decompositions are performed on the spatial correlation matrix between the locations. To illustrate, let $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_N \end{bmatrix}$, where $\boldsymbol{z}_i$ is a $1\times L$ matrix of baseline spatial values. The spatial correlation matrix can be computed by taking the correlation between the locations of $\boldsymbol{z}$ (i.e., $cor(\boldsymbol{z})$), which results in an $L\times L$ matrix. EOF basis vectors have the characteristic that the first basis vector explains the most variability across the spatial grid and each successive basis vector decreases the amount of variability explained. They are optimal in explaining the most variability for a fixed number of basis vectors compared with any other set of bases and are recommended to be the standard for which all other basis functions are judged [1]. Suppose the first p basis vectors of the EOFs are kept. Higher loading magnitudes within a basis vector indicate the locations that are represented by that specific basis vector. Using the first p basis vectors, a Varimax rotation can be performed to create factor analysis basis vectors. This rotates the pre-specified number of basis vectors so that locations are predominately represented by a single basis vector instead of through a linear combination of every basis vector, while maintaining the orthogonality property of the principal components. Oftentimes, the rotated vectors are more interpretable than their corresponding original basis vectors. The number of basis vectors or columns retained in the propagator matrix, $\boldsymbol{\Phi}$, is equivalent to the number of latent variables needed for an individual at each time point. The propagator matrix has dimension $L\times p$ and is assumed to be identical across person and time. This matrix is designed to capture the spatial correlation in the data.

Latent variables are an essential part of hierarchical models, and they allow the dimension reduction technique to be utilized. The latent variables for an individual person at a specific time point are the multiplicative factors that describe how the basis vectors should be combined to estimate every location across the spatial grid. In this model, the latent variables, $\boldsymbol{a}_{i,t}$, for an individual has dimension $p\times(T+1)$. Column two through $T+1$ corresponds to time points one through T, respectively. The first column in the latent matrix serves as the initial starting point for the latent process ($a_{t=0}$), which is necessary from a modeling standpoint.

Individual level baseline covariates (B variables) are characterized in the $\boldsymbol{X}$ matrix within Eqn (1), which has dimension $1\times B$, $\boldsymbol{X}_i=(X_1,\ldots,X_B)$. The respective population parameter coefficients are characterized in the $B\times 1$ vector, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_B)^{'}$. Specifying the covariates within the data model allows the variables to contribute to the overall effect across all locations conditional on $\boldsymbol{\Phi a}_{i,t}$. This requires the assumption that covariates have an overall positive or negative impact and that there is no contrasting relationship for the covariates between the different locations in the visual field after accounting for $\boldsymbol{\Phi a}_{i,t}$ (i.e., a covariate does not have a protective effect in one part of the visual field and a deteriorating effect in a different part). For each individual, the matrix multiplication $\boldsymbol{X}_i\boldsymbol{\beta}$ produces a single value, which is then included in the overall visual field measurement estimate for every location. While it is possible to include $\boldsymbol{X}_i\boldsymbol{\beta}$ in the process model (Section 3.2.2), which would give region specific covariate effects, we are most interested in the overall effect.

After adjusting for the data reduction portion and covariate portion, the remaining variability is assumed to be normally distributed mean zero with independent variances for each location. That is, $\boldsymbol{\varepsilon}_{i,t} \sim MVN_L(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = diag\left[\sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,L}\right]$. Even though we assume that the basis vectors account for all spatial correlation, the different error terms for each location allow us to separately model any information not captured. This is important if some locations are not specifically modeled by the basis vectors. If we were to find posterior estimates for every element in the covariance matrix, we would be estimating $\frac{L(L+1)}{2}$ parameters. The number of parameters in this situation would grow faster than the added data through the increase of locations. If covariances were desired, a structure would need to be implemented to reduce the number of estimates. Because there is no clear way of defining such a structure in our motivating example, we simplified the covariance matrix to a diagonal matrix.

3.2.2. Latent process (process model). The process model (Eqn (2)) allows for two different components: a transition from the previous time point's latent variables and a random error process. Within an individual, each specific time point will have a $p \times 1$ vector of latent values that is applied to the dimension reduction matrix.

$$\text{Process Model}: \qquad \boldsymbol{a}_{i,t} = \boldsymbol{M}\boldsymbol{a}_{i,t-1} + \boldsymbol{\eta}_{i,t} \tag{2}$$

The transition matrix $\boldsymbol{M}$ describes the one unit evolution of the latent variables in time. In this model, the matrix ($\boldsymbol{M}$) will be $p \times p$ with the diagonals of the matrix indicating the multiplicative fluctuation between time points for their specific vector. The off-diagonals indicate how the latent variables are related. For example, if we had a 4×4 transition matrix (i.e., ncol($\boldsymbol{\Phi}$) = 4), a positive value on the second row and third column indicates that the third latent variable has a positive influence in the change of the second latent variable, assuming that the third latent variable is positive. The transition matrix $\boldsymbol{M}$ is not required to be symmetric, which means pairwise relationships between the columns are directionally unique. In the simplest case with the $\boldsymbol{M}$ matrix being the identity, the latent process simplifies to a vector autoregressive random walk.

After adjusting for the latent transition variables, the latent variables are assumed to have independent normally distributed error terms. That is $\boldsymbol{\eta}_{i,t} \sim MVN_p(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = diag\left[\sigma^2_{\eta,1}, \ldots, \sigma^2_{\eta,p}\right]$.

3.2.3. Prior distributions (parameter model). The parameter model contains the specified prior distributions for the parameters we are estimating. For convenience, conjugate priors, as specified in Table I, are utilized for $\boldsymbol{a}_{i,t=0}$, $\boldsymbol{\beta}$, $\boldsymbol{M}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, allowing for increased computational speed through Gibbs sampling. The prior for the latent variables ($\boldsymbol{a}$) at time 0 follows a multivariate normal distribution with a $p \times 1$ mean matrix $\boldsymbol{\mu}_a$ and a $p \times p$ variance matrix $\boldsymbol{\Sigma}_a$. This prior is identical across individuals. The covariates have a multivariate normal distribution prior, with a $B \times 1$ mean matrix $\boldsymbol{\mu}_\beta$ and a $B \times B$ variance matrix $\boldsymbol{\Sigma}_\beta$. The transition matrix for the latent variables, $\boldsymbol{M}$, uses a multivariate normal prior with mean $\boldsymbol{\mu}_m$ and variance matrix $\boldsymbol{\Sigma}_m$. The variances for both the locations ($\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$) and the latent variables ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$) utilize

Table I. Prior distributions for model parameters.

Parameter	Prior distribution
$\boldsymbol{a}_{i,t=0}$	$MVN_p(\boldsymbol{\mu}_{a.\,t0}, \boldsymbol{\Sigma}_{a.\,t0})$
$\boldsymbol{\beta}$	$MVN_B(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$
$\mathbf{M}$	$\mathrm{MVN}_\mathrm{p}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$
$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$	$IW_L(\boldsymbol{V}_\varepsilon, n_\varepsilon)$
$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$	$IW_p(\boldsymbol{V}_\eta, n_\eta)$

Copyright © 2016 John Wiley & Sons, Ltd.

conjugate inverse Wishart distribution priors. The statistical code and package for model implementation are available upon request.

4. Simulation studies

4.1. Computational background

Simulations were conducted to assess the model performance under varying situations. All analyses were performed in RStudio version 0.98.1091, employing the coda, MASS, and parallel packages [21]. Gibbs sampling was implemented using the full conditionals for each set of parameters first estimating the latent variables ($\boldsymbol{a}$) followed by the covariate estimates ($\boldsymbol{\beta}$) and concluding with the variances ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, $\boldsymbol{\Sigma}_{\varepsilon}$). The transition matrix, $\boldsymbol{M}$, was set to be the identity simplifying the process model to a vector autoregressive random walk.

Vague priors were used for all parameters. We assumed multivariate normal priors for the latent variables for a single person at $t=0$, $MVN_p(0,\ 1000\boldsymbol{I}_p)$ as well as for the explanatory variables, MVN_B $(0,\ 1000\boldsymbol{I}_B)$. The variances were set to be independent with 0 value on the off-diagonals. The priors of each parameter in both $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\varepsilon}$ were assumed to follow inverse gamma priors $IG(0.001,\ 0.001)$. Three chains were used for each dataset, and convergence diagnostics were assessed using Gelman and Rubin's [22] univariate potential scale reduction factors (PSRF). It is suggested that PSRF 0.975 quantiles less than 1.20 indicate convergence.

Due to the large amount of parameters being estimated per chain, it is computationally overkill to output and import the iterations without previously thinning and removing non-converged areas. With the amount of simulations performed, manual examination of burn-in and thinning periods was not possible. Therefore, an alternative computing approach was implemented. The algorithm for estimating the posteriors works as follows. MCMC iterations were performed using Gibbs sampling for 1000 iterations without thinning after a burn-in of 100 iterations. Data were imported for the location variances ($\boldsymbol{\Sigma}_{\varepsilon}$), latent variable variances ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$), and the explanatory variables ($\boldsymbol{\beta}$) from a single chain. Autocorrelations were calculated within and between these variables, and the smallest lag that had less than 0.5 autocorrelation was chosen as the thinning value, capped at 100. Using the chosen thinning value, an iterative approach was implemented. MCMC iterations were run using the last observed estimates for all variables as starting values across the three chains thinning appropriately until 200 iterations were output. Upon completion, these iterations across the chains were read in and convergence was assessed for all variables. If at least one variable did not converge, the process was repeated using the last observed iteration values as starting values with the same thinning value. This iterative cycle was completed until there was convergence upon all three chains or until it looped through a maximum of five times, whichever happened first.

In all simulations, every element of the propagator matrix ($\Phi_{l,k}$, for $k=1,\ldots,p$) was generated from a normal distribution ($\Phi_{l,k} \sim N(0,\ 2.5)$) and was assumed known and fixed in the posterior estimation process. Two covariates were used in each simulation: one associated with a continuous explanatory variable and one associated with a dichotomous explanatory variable. These covariates were added linearly and independently from other parameters in the model.

Simulations were performed to validate the model under a complete set of observations. The number of people was set to be 50, the number of time points to be 10, the number of basis vectors to be 6, and the number of locations to be 50. There were 150 datasets generated and analyzed for these situations. These settings were chosen to reflect the attributes of our motivational example. In addition to the 150 simulations performed, 30 additional simulations were conducted, where exactly one of the parameters (e.g., number of time points) was altered to help understand how the model performs under varying circumstances. For example, 30 simulations in addition to the original 150 simulations were performed, where the number of time points was set to be 15 instead of 10 while maintaining the 50 people, 6 basis vectors, and 50 locations. Analysis time and thinning values were recorded for all simulations.

4.2. Simulation results

For ease of discussion, the summaries of the 150 datasets containing 50 people, 50 locations, 10 time points, and 6 basis vectors are presented here. Within a simulation, the coverage probability was calculated for each set of variables ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, $\boldsymbol{\Sigma}_{\varepsilon}$, $\boldsymbol{\beta}, \boldsymbol{a}$). The average and standard deviation of the coverage probabilities of the four sets of variables values are presented in Table II. The standard deviations reflect how much variability was observed between the 150 coverage probabilities. The median run time for these

Copyright © 2016 John Wiley & Sons, Ltd.

Table II. Coverage probabilities for variable sets.

Variable set	Average coverage probability (SE)
$\boldsymbol{\Sigma}_{\varepsilon}$	94.97 (2.86)
$\boldsymbol{\Sigma}_{\eta}$	94.67 (8.92)
$\boldsymbol{a}$	94.65 (0.43)
$\boldsymbol{\beta}$	95.67 (16.32)

sets of simulations was 9.2 min. Across all four variable sets, the average credible intervals covered the true simulated variables close to the targeted 95%. The parameters in all simulations converged according to the Gelman and Rubin [22] univariate PSRF.

Allowing the simulations to vary across the four different sets of variables confirmed our intuition. Increasing the number of people, the number of time points, and the number of locations, while reducing the number of basis vectors, resulted in a reduction in average credible interval widths. The coverage percentage stayed stable around 95%. This is intuitive due to the extra data provided for people, location, and time or due to the fewer parameters estimated for the number of basis vectors.

5. Application

5.1. Introduction to $\boldsymbol{\Phi}$ *and a confirmatory analysis to nerve fiber bundles*

Previous research by Garway-Heath, Poinoosawmy, Fitzke, and Hitchings [19] has established six nerve fiber bundle regions within the central 24° of the visual field (Figure 1). These regions are thought to have different rates of glaucoma progression.

Empirical orthogonal functions (EOF) basis vectors of the data were created and analyzed using the baseline values of the 68 individuals who did not have a complete set of data (i.e., those individuals not used in the analysis). This allows us to attain an estimate of the spatial correlation while minimizing the possibility of overfitting to our complete set of data. Because the number of locations and individuals with complete data are fixed, only the number of basis vectors can vary. Based on the simulation results, caution should be taken for choosing a large number of basis vectors in order to preserve accurate and precise estimates. In dimension reduction models, there have been several suggested methods for determining the number of basis vectors to use depending on the method used in formulating the bases. Hughes and Haran [14] suggest using about 0.10*(number of locations) or fewer basis vectors in analyses when the bases are created using a generalized Moran operator. Others suggest using only basis vectors associated with eigenvalues greater than one [23]. More simply, because EOF basis vectors have the property of decreasing variability explained for each consecutive eigenvector, we could specify a targeted amount of variability we wish to explain in the data (e.g., 70%) and use the first p eivenvectors until that variability is reached.

The first six EOF basis vectors explain 79% of the original variance and produced similar results to those observed by Garway-Heath, Poinoosawmy, Fitzke, and Hitchings [19]. Therefore, p was set to be 6 and is displayed in Figure 2. Six vectors were needed to match the clinically defined regions, and the seventh EOF basis vector explained less than three percent of the spatial correlation in the data so the authors felt it was not worth the estimation cost in the model. Positive vector loadings in the basis vector are displayed in blue with darker shades representing larger magnitude values. Conversely, negative vector loadings are displayed in red. EOF basis vectors have the property of decreased variability explained with each successive vector. The first basis vector explains 36% of the variability in the data and focuses predominately on the upper part of the visual field representing high correlations in this region. The second EOF basis vector has negative loadings associated with it in the upper part of the visual field and large positive loadings associated with the lower part of the visual field indicating contrasting visual measurements between these regions. The third basis vector provides a contrast with negative values on the left side of the visual field against positive values on the right side of the visual field. The remaining basis vectors can be interpreted similarly, but intuitive interpretation of the regions gets more difficult. When we compare these first six EOF basis vectors created from the participants with glaucoma to the predefined regions as shown in Figure 1, we note consistent patterns.

Due to the similarities observed between the EOF basis vectors and the predefined clinical spatial fields, a factor analysis using a Varimax criterion was performed on these six EOF basis vectors. The

Copyright © 2016 John Wiley & Sons, Ltd.

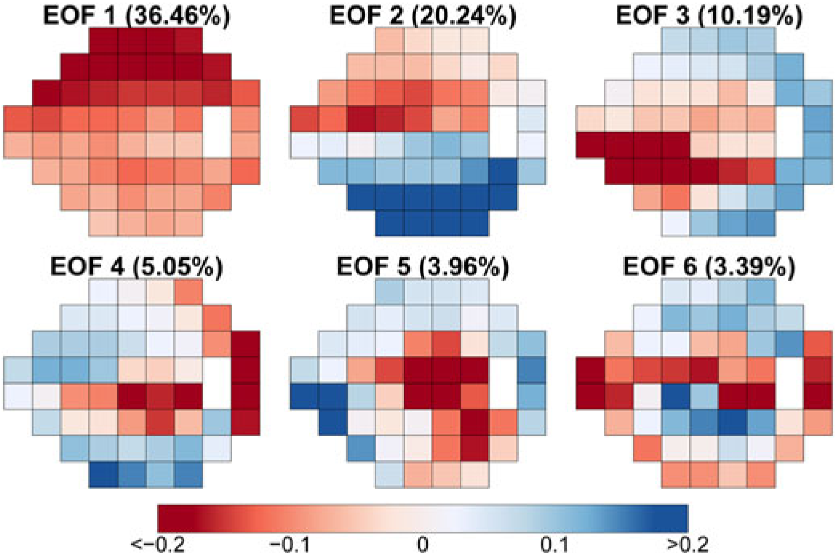

Figure 2. Visual representation of first six empirical orthogonal function basis vectors of baseline data for 68 participants with glaucoma. Values in parentheses are the proportions of variance explained by that EOF.

six rotated basis vectors are displayed in Figure 3. With the rotation, the loading values became mostly positive and the majority of locations can be largely represented by a single factor basis vector. Nerve fiber bundle region 2 appears to be covered in basis vector 3, while nerve fiber bundle region 3 is covered through basis vector 1. Nerve fiber bundle regions 1, 4, and 5 are represented by basis vectors 5, 4, and 2, respectively. Also, nerve fiber bundle region 6 can be observed in all the vectors. Thus, there is likely not to be a large amount of spatial variability in nerve fiber bundle regions 6, due to the fact that it is not significantly represented in the factor basis vectors. Because the modeling in this specific problem serves as a confirmatory analysis with this data, the Varimax rotation of the first six EOF basis vectors is used to create $\boldsymbol{\Phi}$. These basis vectors have more clinical interpretability than the EOF basis vectors and help confirm the method used in the creation of $\boldsymbol{\Phi}$.

5.2. *MCMC specifics*

Prior distributions were set to those discussed in the simulation in Section 4.1 for each set of variables. The data were thinned every 100 iterations as established by the autocorrelation part using the iterative

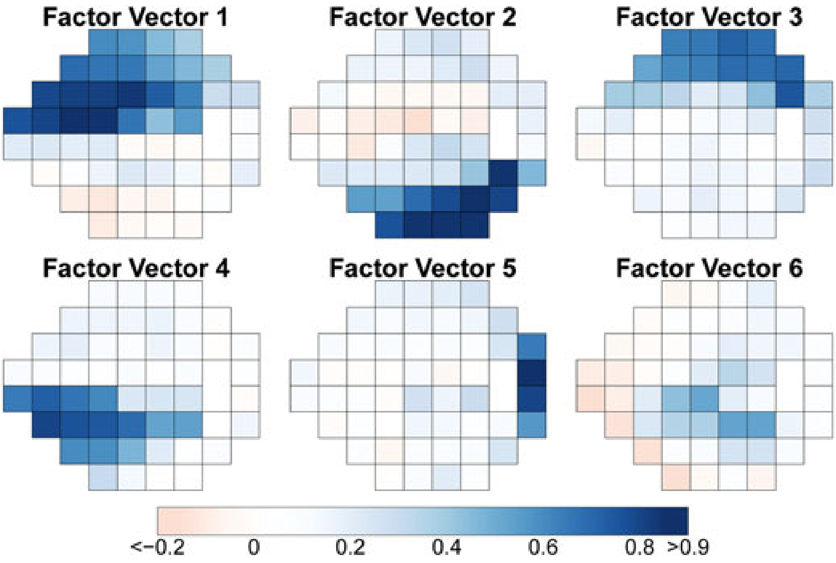

Figure 3. Visual representation of first six empirical orthogonal function basis vectors after rotation using a Varimax criterion of baseline data for 68 participants with glaucoma.

estimation procedure discussed in Section 4.1, and 350 iterations were output for each loop. Convergence occurred immediately (1000 burn-in; 35 000 iterations; thinned 100). The entire analysis took 83 min.

For this analysis, there were 24 336 observations (52 people × 52 locations × 9 time points). Sensitivity analyses were performed increasing the strength of the priors for the location and latent variances. Across the analyses, the means, standard deviations, and 95% credible intervals do not vary much, regardless of the prior strength (data not shown). We notice an increase in estimate and credible interval for the last two latent variances with increased prior strength, but not to a great magnitude. With the large amount of outcome information, hyperparameters and hyperpriors on the variance terms do not play a large influence in the posterior distribution unless they are specified to be strongly informative (e.g., prior for variance to be $IG(10, 10)$).

5.3. Covariate results

Prior distributions, posterior means and standard deviations, and the 95% credible intervals for selected variables as shown in Table III are presented in Table IV. The parameter estimates can be interpreted as how the covariates relate to the remaining variability after accounting for the dimension reduction and

Table III. Baseline demographic for participants with glaucoma.

Variable	Glaucoma baseline demographics	
Mean average retinal thickness (SD)	67.2 (14.7)	
Mean age (SD)	65.4 (9.5)	
Glaucoma type (%)	NTG	(34.0%)
	POAG	(41.5%)
	OTHER	(24.5%)
Gender (%)	Male	(37.7%)
	Female	(62.3%)
Optic disc hemorrhage (%)	Yes	(18.9%)
	No	(81.1%)
Surgery (%)	Yes	(30.2%)
	No	(69.8%)
Mean deviation category (%)	Mild	(54.7%)
	Moderate	(45.3%)

Table IV. Prior distributions and parameter estimates for analysis of glaucoma progression.

Parameter	Prior distribution	Posterior mean (SD)	95% credible interval
$\sigma^2_{\varepsilon,1}$	IG (0.001, 0.001)	15.35 (1.11)	(13.36, 17.60)
$\sigma^2_{\varepsilon,11}$	IG (0.001, 0.001)	36.35 (2.59)	(31.54, 41.74)
$\sigma^2_{\eta,1}$	IG (0.001, 0.001)	1.45 (0.33)	(0.90, 2.15)
$\sigma^2_{\eta,2}$	IG (0.001, 0.001)	1.33 (0.24)	(0.91, 1.85)
$\sigma^2_{\eta,3}$	IG (0.001, 0.001)	2.20 (0.52)	(1.36, 3.27)
$\sigma^2_{\eta,4}$	IG (0.001, 0.001)	1.27 (0.26)	(0.86, 1.83)
$\sigma^2_{\eta,5}$	IG (0.001, 0.001)	0.46 (0.18)	(0.03, 0.86)
$\sigma^2_{\eta,6}$	IG (0.001, 0.001)	0.48 (0.26)	(0.04, 1.03)
β_1 (Glaucoma POAG)	Normal (0, 1000)	11.92 (0.41)	(11.19, 12.65)
β_2 (Glaucoma NTG)	Normal (0, 1000)	17.80 (0.41)	(16.91, 18.73)
β_3 (Males)	Normal (0, 1000)	5.30 (0.36)	(4.59, 6.00)
β_4 (Surgery)	Normal (0, 1000)	5.94 (0.45)	(5.01, 6.81)
β_5 (Mean deviation mild)	Normal (0, 1000)	11.88 (0.37)	(11.15, 12.63)
β_6 (Optic disc hemorrhage)	Normal (0, 1000)	−1.24 (0.53)	(−2.25, −0.19)
β_7 (Age)	Normal (0, 1000)	−0.46 (0.02)	(−0.50, −0.42)
β_8 (Average retinal thickness)	Normal (0, 1000)	−0.23 (0.02)	(−0.26, −0.20)
$a_{1,1,1}$	Normal (0, 1000)	3.12 (1.90)	(−0.77, 6.61)
$a_{1,6,1}$	Normal (0, 1000)	1.17 (1.56)	(−1.90, 4.08)

Copyright © 2016 John Wiley & Sons, Ltd.

latent variable portion. A positive parameter estimate indicates that higher observed explanatory values are associated with overall higher outcome values, on average. On the contrary, negative parameter estimates indicate that higher observed explanatory values are associated with overall lower observed value in individuals. All of the parameter estimates significantly contribute to the remaining unexplained variability after accounting for the dimension reduction portion. Participants with Primary Open-Angle Glaucoma or Normal-Tension Glaucoma glaucoma, being male, having surgery, and those considered to have a mild mean deviation had higher overall average decibel measurements, while those with an optic disc hemorrhage, who were older, and who had a higher average retinal thickness had lower overall average decibel measurements, on average. The covariates are on the individual level, so inferences are made about an overall increase or decrease in the visual field instead of a location-specific interpretation. Incorporation of covariates in the DROIIDS model helps account for differences in average visual field decibel measurements between people and reduces the remaining variability.

5.4. *Rates of change analyses*

One beneficial product of this spatio-temporal model is the smoothed estimates of the overall process. The estimated latent variables and population parameter estimates can be utilized to find posterior predicted mean estimates within a person for each time point across locations. Spatio-temporal modeling helps capture the smoothed underlying trend, while measurement variability are captured in Φ, the covariates, and in the error terms.

The raw and smoothed decibel measurements visual representation plots at each time are displayed in Figure 4. Figures 4a and 4c represent the raw and estimated visual field measurements for an individual in the study, while Figures 4b and 4d represent the raw and estimated visual field measurements for a different individual in the study. For the first individual, we can see lower visual field measurements in the lower left part of his or her visual field (Figure 4a). We notice fluctuation over the nine visits in this region, but we do observe an overall deterioration in visual field measurements over time. Through

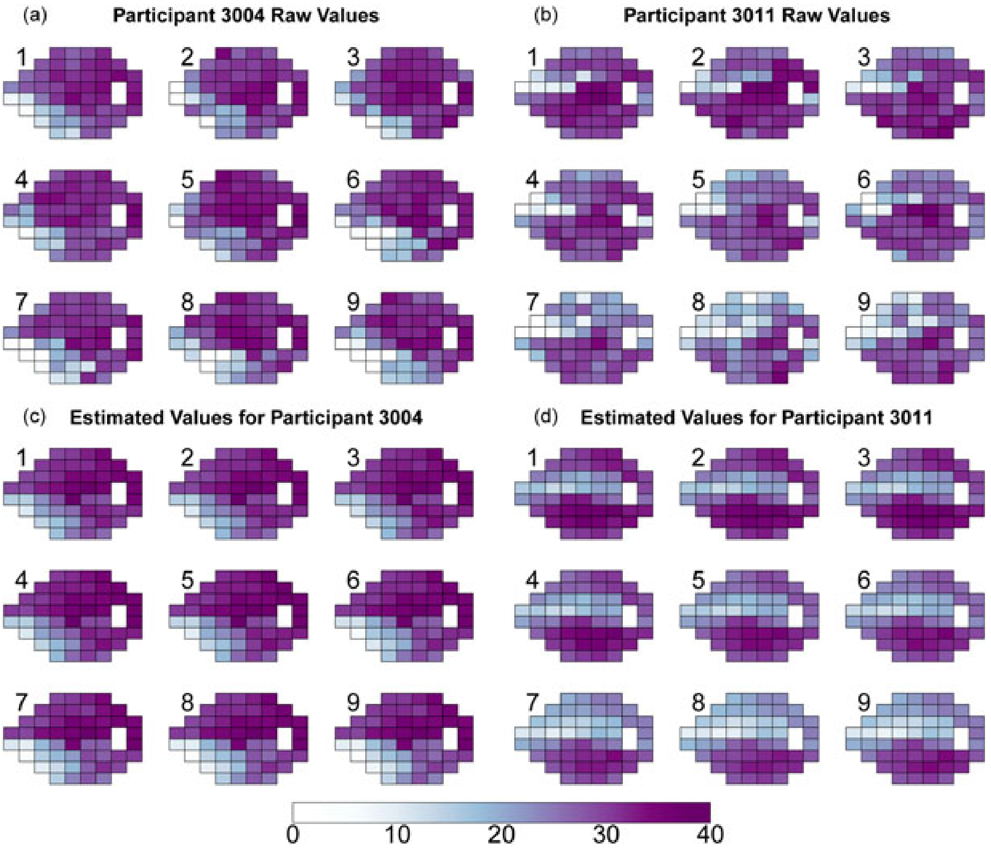

Figure 4. Raw and estimated decibel measurements at each location for two individuals. The raw decibel measurements for the nine time points are presented for participant 3004 in (a) and participant 3011 in (b). The predicted values for the nine time points are presented for participant 3004 in (c) and participant 3011 in (d).

our model, the visual field measurement estimates capture this pattern and smooth the entire surface (Figure 4c). Contrary to the first person, the second individual in Figures 4b and 4d has glaucoma progression in the upper left part of his or her eye (Figure 4b). Again, the DROIIDS model is able to smooth these variable observations and provide a more realistic underlying visual field measurement trend that is occurring (Figure 4d).

At each location, a simple linear regression line can be fit to the T smoothed visual field measurement estimates by using time, $t=1, \ldots, T$, as predictors. This will produce a smoothed linear estimate of deterioration over time representing an average rate of change throughout the visits. Clinicians can use these rates to determine the areas within the central 24° of the visual field that have the fastest glaucoma progression. Two example rate plots are displayed in Figure 5. Figure 5a shows a high progression in the lower left part of her visual field for the individual shown in Figures 4a and 4c, while the majority of the visual field has high progression for the individual in Figures 4b and 4d. For the first individual (Figure 5a), if we compare these high rates of change with the pre-defined nerve fiber bundle regions presented in Section 5.1 (Figure 1), we notice that regions 4 and 5 are the main areas affected in this individual.

These rates have great potential to clinicians and researchers. If there are medications that have a larger protective effect in certain nerve fiber bundle regions compared with others, then these plots would help clinicians determine the best prescription for the individual. In addition, clinicians and researchers can compare the progression rates and areas affected with individual demographics to determine if there are certain risk factors associated with certain patterns of progression.

6. Discussion

Glaucoma is a blinding disease affecting people's vision with no known cure. However, through collaborative research, many treatments have been developed that helps slow the progression of the disease. Using data from the Mechanism of Parametric Variability study, we were able to measure the progressive rates of changes among the 52 participants with glaucoma in the study who returned at every follow-up. Analyses were performed using our introduced DROIIDS model. The model was designed in a Bayesian Hierarchical framework that incorporates a dimension reduction portion and individual baseline covariates. Through these two parts, the data were smoothed and rates of change were calculated for each location. These rates have potential to help clinicians process the glaucoma progression of subjects at return follow up visits.

The dimension reduction matrix was created using a Varimax rotation on the first six EOF basis vectors of the data. The variable regions highlighted in each basis vector are close to the pre-defined nerve fiber bundle regions introduced by Garway-Heath et al. [19]. This resemblance helps support the method chosen to create the dimension reduction matrix. Individual covariates are incorporated to help explain residual variability not captured specifically by the dimension reduction matrix.

Missing data are inevitable in trials, and numerous methods have been proposed to handle the missingness over the years. Less than half (43%) of the participants with glaucoma had a complete dataset in the current analysis. Incorporating a mechanism to handle the missingness within the model will allow us to utilize all information collected and is part of our future research.

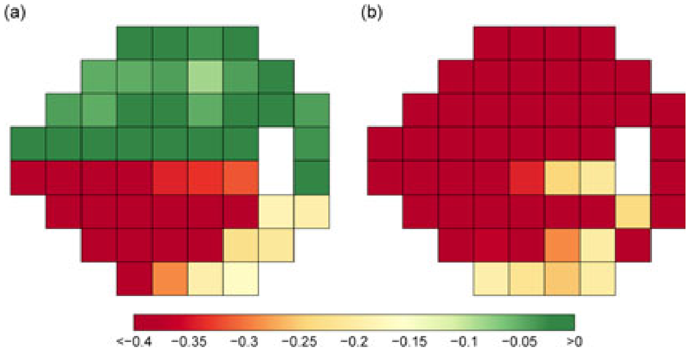

Figure 5. Rates of change plots for participant 3004 (a) and participant 3011 (b).

Copyright © 2016 John Wiley & Sons, Ltd.

Propagator and transition matrices were created and were then fixed in this current analysis. If there were both a sufficient sample size and a large number of repeated measurements on subjects, then these matrices could be estimated in an iterative step with the other parameters.

The model developed within this manuscript is designed to better understand, both individually and epidemiologically, the direction and rate of disease progression. Through the incorporation of multiple individuals' datasets, the model was able to account for excess variability in the data. Clinicians can use the model results to help determine the best prescription for a susceptible individual.

Acknowledgements

This research would like to thank the Mechanism of Perimetric Variability grant (RX000140-03) supported through a VA Merit Review Grant awarded to the Department of Ophthalmology at the University of Iowa from Research to Prevent Blindness, New York, NY. In addition, the authors would like to thank the two reviewers for their suggestions and comments that helped improved the clarity and content of this paper.

References

1. Cressie NA, Wikle CK. Statistics for Spatio-temporal Data. Wiley: Hoboken, N.J., 2011.
2. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1905–1934)* 1927; **115**:700–721.
3. Dukic V, Lopes H, Polson N. Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association* 2012; **107**:1410–1426.
4. Porter AT, Oleson JJ. A path-specific SEIR model for use with general latent and infectious time distributions. *Biometrics* 2013; **69**:101–108.
5. Brown GD, Oleson JJ, Porter AT. An empirically adjusted approach to reproductive number estimation for stochastic compartmental models: a case study of two Ebola outbreaks. *Biometrics* 2016; **72**(2):335–343.
6. Hooten MB, Anderson J, Waller LA. Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology* 2010; **1**:177–185.
7. Lekone PE, Finkenstadt BF. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics* 2006; **62**:1170–1177.
8. Mode CJ, Sleeman CK. Stochastic processes in epidemiology: HIV/AIDS, other infectious diseases, and computers. World Scientific: Singapore, 2000.
9. Hooten MB, Garlick M, Powell J. Computationally efficient statistical differential equation modeling using homogenization. *Journal of Agricultural, Biological, and Environmental Statistics* 2013; **18**:405–428.
10. Whittle P. Topographic correlation, power-law covariance functions, and diffusion. *Biometrika* 1962; **49**:305–314.
11. Hooten MB, Wikle CK. A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics* 2008; **15**:59–70.
12. Wikle CK, Holan SH. Polynomial nonlinear spatio-temporal integro-difference equation models.(Report). *Journal of Time Series Analysis* 2011; **32**:339.
13. Wikle C, Hooten M. A general science-based framework for dynamical spatio-temporal models. *An Official Journal of the Spanish Society of Statistics and Operations Research* 2010; **19**:417–451.
14. Hughes J, Haran M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 2013; **75**:139–159.
15. Porter AT, Holan SH, Wikle CK. Bayesian semiparametric hierarchical empirical likelihood spatial models. *Journal of Statistical Planning and Inference* 2015; **165**:78–90.
16. Oleson JJ, Wikle CK. Predicting infectious disease outbreak risk via migratory waterfowl vectors. *Journal of Applied Statistics* 2013; **40**:656–673.
17. Soebiyanto RP, Adimi F, Kiang RK. Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. *PLoS ONE* 2010; **5**:1–10.
18. Wall M, Woodward KR, Doyle CK, Artes PH. Repeatability of automated perimetry: a comparison between standard automated perimetry with stimulus size III and V, matrix, and motion perimetry. *Investigative Ophthalmology & Visual Science* 2009; **50**:974–979.
19. Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA. Mapping the visual field to the optic disc in normal tension glaucoma eyes1. *Ophthalmology* 2000; **107**:1809–1815.
20. Association AO. Glaucoma. In Glaucoma. American Optometric Association: City, 2015.
21. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2015.
22. Gelman A, Rubin DB. Inferences from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
23. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis/Richard A. Johnson, Dean W. Wichern (5th edn). Prentice Hall: Upper Saddle River, N.J., 2002.

Copyright © 2016 John Wiley & Sons, Ltd.