Chapter 13
Bayesian Modelling to Assist Inference on Health Outcomes in Occupational Health Surveillance

Nicholas J. Tierney, Samuel Clifford, Christopher C. Drovandi, and Kerrie L. Mengersen

Abstract *Objectives:* Occupational Health Surveillance (OHS) facilitates early detection of disease and dangerous exposures in the workplace. Current OHS analysis ignore important workplace structures and repeated measurements. There is a need to provide systematic analyses of medical data that incorporate the data structure. Although multilevel statistical models may account for features of OHS data, current applications in occupational health medicine are often not appropriate for OHS. Additionally, typical OHS data has not been analysed in a Bayesian framework, which allows for calculation of probabilities of potential events and outcomes. This paper's objective is to illustrate the use of Bayesian modeling of OHS. Three analytic aims are addressed: (1) Identify patterns and changes in health outcomes; (2) Explore the effects of a particular risk factor, smoking and industrial exposures over time for individuals and worker groups; (3) identify risk of chronic conditions in individuals. *Method*: A Bayesian hierarchical model was developed to provide individual and group level estimates and inferences for health outcomes, FEV1%, BMI, and Diastolic and Systolic blood pressure. *Results*: We identified individuals with the greatest degree of change over time for each outcome, and demonstrated how to flag individuals with substantive negative health outcome change. We also assigned probabilities of individuals moving into "at risk"

N. J. Tierney (✉)
ARC Centre of Excellence for Mathematical and Statistical Frontiers, Parkville, VIC, Australia

School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia

Department of Econometrics and Business Statistics, Monash University, Melbourne, Clayton, VIC, Australia
e-mail: nicholas.tierney@monash.edu

S. Clifford · C. C. Drovandi · K. L. Mengersen
ARC Centre of Excellence for Mathematical and Statistical Frontiers, Parkville, VIC, Australia

School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia

K. L. Mengersen et al. (eds.), *Case Studies in Applied Bayesian Data Science*, Lecture Notes in Mathematics 2259, https://doi.org/10.1007/978-3-030-42553-1_13

health categories 1 year from their last visit. *Conclusion*: Bayesian models can account for features typically encountered in OHS data, such as individual repeated measurements and group structures. We describe one way to fit these data and obtain informative estimates and predictions of employee health.

13.1 Introduction

Occupational Health Surveillance (OHS) is the systematic collection, analysis, and dissemination of employee exposure and health data to facilitate early detection of disease and dangerous exposures in the workplace [1]. Australian employers have a responsibility to identify, assess, and control risks arising from workplace hazards [2, 3]. There is a rigorous methodology for OHS data collection, but a surprising lack of agreement about analysis of these data [4]. Indeed, industry OHS data collection is often targeted for managing risk and implementing engineering controls. Consequently, many of the analyses conducted in industry focus on the likelihood of exposure rather than the impact of these risk factors on health. Moreover, current practices may ignore important data structures such as repeated measurements and workplace structures. This results in inferences not being applicable for individuals over time, or for groups with similar exposures within the workplace.

There is a need to provide systematic analyses of medical data that incorporate workplace structure, relevant to risk factors. An example of such a workplace structure is segmentation of the workplace into similar exposure groups (e.g., as in [5]). Moreover, such analyses need to incorporate typical features of OHS data, in particular where individuals have multiple health measurements, or single repeated measurements over time, or missing data. These analyses should provide both individual and group health predictions, and should improve the understanding of exposure effects on the workplace population as a whole, as well as similar exposure groups and individuals. Such analyses could flag individuals and groups for further health monitoring. The absence of such analyses in industry means that chronic disease and dangerous work environments may go unidentified and that health funding is not optimally or effectively targeted.

These features of OHS data described above can be accounted for with multilevel statistical models. These are in wide use in epidemiology [6, 7], and have been used in occupational health medicine to evaluate decline in lung function for ceramic fibre workers [8], assess impacts of asbestos [9], measure decline from cystic fibrosis [10] and model leptospirosis in abattoir employees [11].

However, applications of multilevel models in occupational health medicine do not quite mimic the analyses conducted in standard industry environments, as they might ignore individuals with only one measurement, population minority groups, or workplace structures [8, 9]. This is likely due to the fact that the goal of these papers is often to demonstrate the use of a new method [9, 10, 12], or discover new health risk factors or exposures [13–17].

In contrast, the goal of OHS analysis is to provide individual predictions for health, understand the effect of exposures on groups, and monitor exposures and health over time, so that individuals and groups at risk of some disease can be flagged for further health monitoring. Thus, ignoring cases with only one medical visit or analysing only subsets of the population for reasons such as sufficient sample size, can increase bias and/or variance of estimates.

Bayesian models provide a pragmatic framework for this research problem, as they provide simple and effective ways of analysing small effects, and provide a rich set of results that can be interpreted with probabilistic statements. Bayesian methodology also allows for direct comparison between groups and individuals, and provides probabilities on potential events and outcomes.

Bayesian techniques have been recently applied in occupational health, with [5] demonstrating the use of hierarchical models to combine monitoring data and professional judgement from occupational hygienists to facilitate decision making. Bayesian hierarchical models have also been applied to quantify chemical exposure variation in human populations [18], and to combine two data sources from animal studies and human industrial studies to create informative priors to estimate human lung function changes [19].

Industry data used in the literature typically consist of longitudinal data collected from employees in a particular industry, or set of industries. These data are used to evaluate the effect of the working conditions, such as long hours with no sleep [20, 21], metal smelting, and other exposures [8–12, 17]. Other OHS studies focus on small populations, using experiments to evaluate effects of increasing some exposure on health [22, 23], or larger cross-sectional studies using registry data, cohort studies, or surveys to evaluate the effect of an environmental exposure on diseases such as rhinitis or cystic fibrosis [10, 16].

Notwithstanding these studies, there are no examples of Bayesian hierarchical modelling and analysis of typical OHS data with applications in an industry context. This paper analyses OHS data from selected industrial sites around Australia to identify risk factors for health outcomes. The multilevel model adopted is a Bayesian hierarchical model providing individual and group level estimates and inferences.

The aims of the analysis are threefold, and are focused on the following health outcomes: lung function (as a percentage of predicted Forced Expiratory Volume in 1 s (FEV_1, $FEV_1\%$), Body Mass Index (BMI), and systolic and diastolic blood pressure. These health outcomes were selected as they are clinically substantive in the case study population and are well known in OHS. The first analytic aim is to identify patterns and changes in health outcomes. The second aim is to explore the effects of a particular risk factor: smoking and industrial exposures over time for individuals and worker groups. The third aim is to identify risk of chronic conditions (such as obesity, hypertension, and obstructive/restrictive lung disease) in individuals.

13.2 Method

13.2.1 Case Study Data

The case study considered in this paper is typical of many large companies that are involved in a range of activities, such as construction, mining, manufacturing and agriculture. For reasons of confidentiality, the particular industry and associated sites are not named here. The data are comprised of over 3000 employee medical records from nearly 2000 individuals located at a number of sites. Each observation is a medical visit, and while most employees have one or two visits, some have over ten visits over a 10 year period. Employees are typically grouped by their workplace exposure; for example, Administration employees are less likely to be exposed to environmental factors such as dust or noise, compared to maintenance exposure groups. In this way, Administration provides a useful control group to compare to the other exposure groups. Employees may change positions within the company over their career and thus may also change their exposure group. The pattern of measurements over time for individuals are illustrated in Fig. 13.1, which displays individual measurements of lung function (FEV1%) over visits for selected exposure groups.

The frequency of medical visits changes for each exposure group, as certain exposure groups require more frequent medical examinations to ensure that they

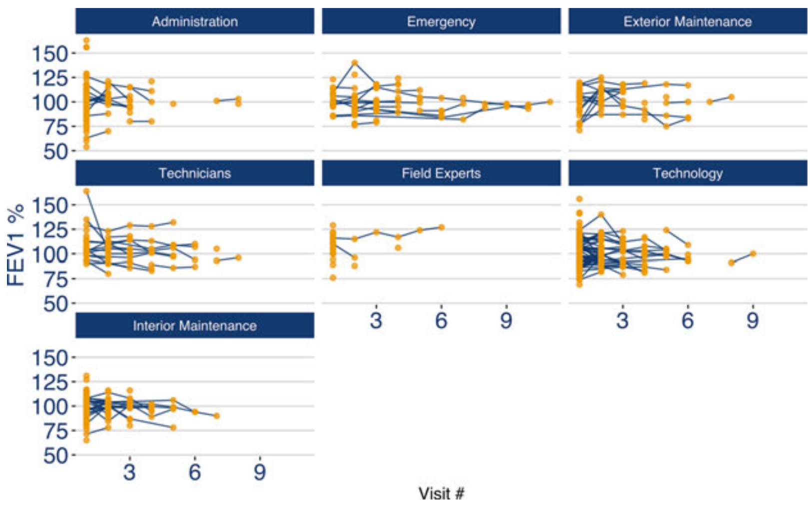

Fig. 13.1 Individual employee lung function (FEV1%) over their medical visit number (1, up to 10), for each exposure group. A sample of 50% of employees is used to reduce overplotting. Individuals are linked by a line between observations. A point on the 2nd or later visit which is not joined to previous points by a line indicates individuals who have changed exposure group. Broken lines and individual floating points without lines indicate where individuals have changed exposure group. Note that the number of days between visits varies by individual and exposure group

are fit for work. The times between visits for each worker were not equally spaced, with median number of days since first visit being 1028 (IQR = 193–3363), or 2.8 years (IQR = 0.5–9.2 years). Gender (male or female) and smoking status (ever smoker or never smoker) were also recorded. Dust data were not recorded for some dates and were interpolated using a loess model [24] fitted for each exposure group, so that the values corresponded to medical examination dates. Interpolated values should be treated with care, and explored with visual and numerical summaries.

13.2.2 Ethics

The Queensland University of Technology Human Research Ethics Committee assessed that this research met the conditions for exemption from HREC review and approval in accordance with section 5.1.22 of the Australian National Statement on Ethical Conduct in Human Research.

13.2.3 Patient and Public Involvement

The development of the research questions and outcomes were informed by discussion with health practitioners who helped collect the data. The patients were not involved in the results, design, or recruitment. The paper will be shared with the medical practitioners for their use in future designs. We thank the health practitioners and patients involved in the data collection.

13.2.4 Modelling

We construct four multilevel Bayesian hierarchical models. Each model predicts one of the four outcomes: lung function (FEV1%), Body Mass Index (BMI), systolic blood pressure, and diastolic blood pressure.

Let Y_{ij} be the ith individual's jth health observation, at a time day_{ij} after their first visit. We assume that Y_{ij} follows a normal distribution with mean μ_{ij} and variance σ_y^2. Let β_{0i} and β_{di} be respectively the individual intercept and individual health trend coefficient associated with the jth day for the ith person; these individual parameters are centered around an overall intercept β_{0c} and an overall slope β_{dc}, the effect of the number of days since arriving at the workplace. Thus β_{di} is the linear trend over time for the health characteristics of interest for the ith individual, over and above the overall population effect. Let β_g be the effect of being female (compared to being male); let β_s be the effect of being a smoker (compared to a never smoker), and let $\sum_{k=1}^{n_{\text{exposure}}-1} \beta_k I(\text{exposure}_{ij} = k)$ be the effect of a workplace exposure, where $I(.)$ indicates whether an individual i at a visit j is in exposure group k, with the baseline exposure group set to Administration. Thus the model for a particular health outcome is represented as:

$$Y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_y^2)$$

with

$$\mu_{ij} = \beta_{0i} + \beta_{di}\,\text{day}_{ij} + \beta_g\,\text{gender}_{ij} + \beta_s\,\text{smoke}_{ij} + \beta_p\,\text{dust}_{ij} + \sum_{k=1}^{n_{\text{exposure}}-1} \beta_k I(\text{exposure}_{\text{ij}} = k)$$

$$\beta_{0i} \sim N(\beta_{0c}, \sigma_0^2)$$

$$\beta_{di} \sim N(\beta_{dc}, \sigma_d^2)$$

for $i = 1 \ldots n_I$, $j = 1 \ldots n_{0i}$, $k = 1 \ldots, n_E$, where n_I is the total number of individuals, n_{0i} is the number of observations for each individual, and n_E is the number of exposure groups.

In the absence of other information, all of the regression coefficients were allocated independent normal priors with a mean of 0 and a variance of $D_1 = 10^3$.

$$\beta_{0c}, \beta_{dc}, \beta_{di}, \beta_g, \beta_s, \beta_p, \beta_k \sim N(0, D_1)$$

Priors on $\sigma_0, \sigma_y, \sigma_d$ were set to a uniform distribution with bounds of zero and $D2$, where $D_2 = 100$ for BMI and $\text{FEV}_1\%$, and $D_2 = 50$ for Systolic and Diastolic blood pressure. D_2 is intended to better reflect the variation in BMI and $\text{FEV}_1\%$ compared to blood pressure. Note also that we do not recommend automatically choosing set values for the uniform, but to instead choose sensible bounds based on the problem at hand.

$$\sigma_y, \sigma_0, \sigma_d, \sim \text{Uniform}(0, D_2)$$

Note also that the priors used for the β terms are proper priors, which produce a proper posterior. In some cases improper priors such as an infinite uniform prior might be used, but these are sometimes not valid choices (See [25] and [26] for more details). It is worthwhile to consider the choice of prior for the variance terms. Although we have used inverse gamma and uniform priors, other weakly informative priors could be considered, such as a half-t-prior (represented as a half-Cauchy) [27]. It is important to not automatically choose uniform or half-t-priors, but to explore options during model building.

Data processing and manipulation were implemented using the R statistical programming language [28] and various R packages [29–34]. To ensure reproducibility, the paper was written using rmarkdown and knitr [35, 36]. Potential outliers in the data were checked administratively and confirmed for biological plausibility in the context of the workforce under consideration. Given this, we elected to include them in the analyses. Moreover, the modelling goal is to identify those who are risk, so removing outliers seems counter to that goal. The model was run for 20,000 iterations (10,000 burnin) using JAGS [37, 38]. Thinning was applied to the analysis, removing every 20th value to assist in reducing autocorrelation and for computational storage. We note that thinning is not absolutely necessary in an analysis, and should be assessed case by case [39]. We note that other software such as STAN, WinBUGS or OpenBUGS, Nimble, and greta could also have been used [40–43]. The diagnostics for MCMC convergence were predominantly graphical and statistical [44, 45]. Graphical evaluation included expert examination of posterior density plots, traceplots and autocorrelation plots of parameters. Statistical evaluation included calculation of the Geweke diagnostic and effective sample size.

Missing values were imputed from their respective posterior conditional distributions as part of the Bayesian analysis. Posterior estimates of each parameter, including mean, 95 and 80% credible intervals, and probability of being negative were calculated after burnin. An effect was nominated as substantive if the corresponding credible interval did not contain zero. The probability of individual health outcomes reaching the threshold value of being a chronic condition was also calculated. Individuals were identified as being "at risk" if the corresponding estimates of the parameter for change over time, β_{di}, contained 0 in the 95% credible intervals, and β_{di} was far away from zero. For the purposes of exposition, individuals with 3 or more visits were selected as examples to explore further.

Patterns and trends in health outcomes were examined by exploring individuals' change over time and identifying substantive effects, addressing analytic aim 1. The effects of smoking and industrial exposures over time for individuals and exposure groups were examined by evaluating substantive effects of smoking and dust for each outcome, and finding those exposure groups substantively different from the Administration population, addressing aim 2. To identify future risk of chronic conditions, 1 year forecasts for each individual and corresponding 95% credible intervals were calculated from the respective posterior predictive distribution, and the probability of having a chronic condition in 1 year was obtained, addressing aim 3. Model fit was evaluated by examining the proportion of observed values lying within the 95% and 80% posterior predictive intervals [46, 47].

13.3 Results

13.3.1 Demographics

In the case study dataset, the population was predominantly male (86%), with the mean overall age being 35.8 years. Males were older on average, but not significantly so, compared with females. For all exposure groups there were more

Table 13.1 Percent of the population in selected exposures

Exposure	% of population
Technology	18–20
Administration	8–10
Interior maintenance	8–10
Technicians	6–7
Emergency	5–6
Exterior maintenance	4–5
Field experts	2–3

males than females, except in the Administration exposure group. The proportion of individuals in selected exposure groups is shown below in Table 13.1.

13.3.2 Model Fit

Figure 13.2 shows the posterior predictions for each outcome plotted on the y axis against the observed values on the x axis. The points represent the observed values and the corresponding posterior means with vertical lines representing the respective 95% posterior predictive intervals. A line of perfect prediction runs from the bottom left to the top right corner. The points and lines are shown in red to indicate when the observed value lies outside of the 95% posterior predictive interval.

Model fit was assessed by visual inspection of Fig. 13.2, and by assessing the percentage of observed values that lie within nominated posterior predictive intervals (Table 13.2). The models for BMI and FEV1% had very high proportions of observed values in the 95% intervals and 80% intervals, indicating reasonable model fit.

Figure 13.3 shows four selected "at risk" individuals and their posterior mean and credible intervals for the health characteristics systolic blood pressure, FEV1%, diastolic blood pressure, and BMI. The proportion of individuals "at risk" for each health outcome, and the mean and standard deviation for each health outcome for those at risk and not at risk, are shown in Table 13.3. Individuals were identified as "at risk" in this case according to whether their parameter estimates for change over time β_d were the furthermost away from zero (and did not contain zero in the credible interval) for the health characteristics systolic blood pressure, FEV1%, diastolic blood pressure, and BMI. As described in the method section, these individuals had 3 or more visits.

Table 13.4 shows the posterior mean, 95% credible interval, and probability of being negative for each of the risk factors considered, namely smoking, dust, and days, since commencement.

The number of days since first visit had a substantive effect on all outcomes, and was associated with a decrease in $FEV_1\%$, and a decrease in BMI, diastolic and systolic blood pressure. Dust did not have a substantive impact on any outcomes, but was associated with an 11% chance of decreased BMI, a 22% chance of

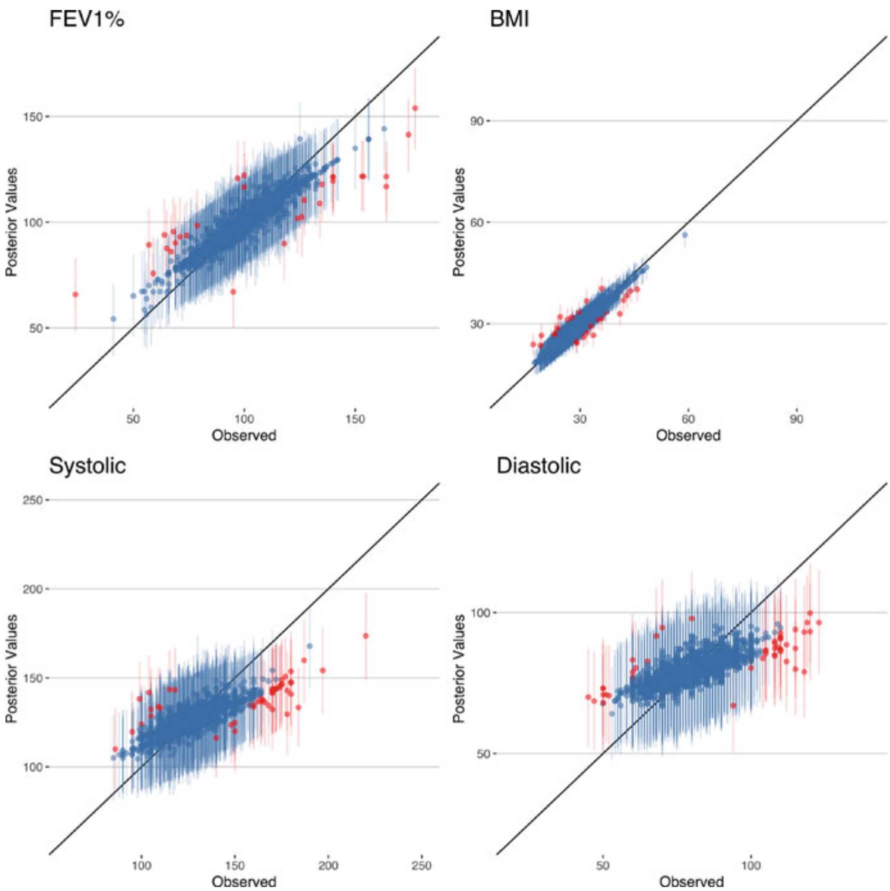

Fig. 13.2 Observed values for each health outcome plotted against the posterior mean values (point) with their respective 95% posterior predictive interval (lines). A line of perfect prediction is shown. The points and lines are shown in red to indicate when the 95% posterior predictive interval lies outside the line of perfect prediction

Table 13.2 Percent of observed values inside the 95 and 80% posterior prediction intervals and RSS for each model outcome

Outcome	Inside 95% PI	Inside 80% PI
BMI	99.09	97.52
Diastolic BP	98.38	90.18
FEV1 (%)	98.96	95.91
Systolic BP	98.32	92.15

decreased FEV_1%, a 31% chance of decreased systolic blood pressure, and a 73% chance of decreased diastolic blood pressure. Being a smoker was associated with substantively decreased FEV_1%, a 100% chance of decreased FEV_1 %, a 95% chance of decreased BMI, a 13% chance of increased diastolic blood pressure, and a 24% chance of increased systolic blood pressure.

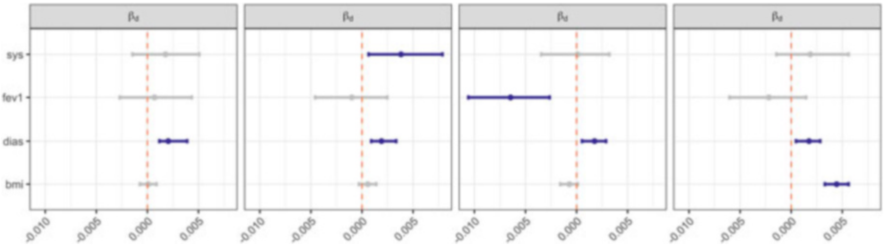

Fig. 13.3 Individuals whose parameter estimates for change over time, β_d, were the furthest away from zero (and did not contain zero in the credible interval) for the health characteristics systolic blood pressure, FEV1%, diastolic blood pressure, and BMI

Table 13.3 The proportion of individuals with 3 or more visits who were 'at risk' (95% credible interval for change over time did not include zero) and the health outcomes for the 'at risk' and 'not at risk' groups

Summary	BMI	FEV1%	Systolic	Diastolic
Proportion at risk	0.1	0.02	0.01	0.98
Mean (SD) not at risk	27.68 (3.84)	100.19 (12.36)	126.46 (13.32)	73.89 (10.98)
Mean (SD) at risk	30.55 (5.22)	102.00 (23.76)	140.77 (19.01)	80.94 (10.04)

Table 13.4 Estimated posterior mean, 95% credible intervals and probability of the effect of Day, Dust, and Smoking parameters being less than zero

Terms	BMI	FEV1%	Systolic	Diastolic
Days	6.09×10^{-4}	-8.44×10^{-4}	1.35×10^{-3}	1.78×10^{-3}
	$(4.72 \times 10^{-4}, 7.52 \times 10^{-4})$	$(-1.41 \times 10^{-3}, -2.51 \times 10^{-4})$	$(7.30 \times 10^{-4}, 1.97 \times 10^{-3})$	$(1.34 \times 10^{-3}, 2.17 \times 10^{-3})$
	0	9.98×10^{-1}	0	0
Dust	0.11	0.33	0.24	−0.21
	(−0.06, 0.3)	(−0.52, 1.15)	(−0.72, 1.16)	(−0.92, 0.51)
	0.11	0.22	0.31	0.73
Smoking	−0.33	−1.87	0.47	0.52
	(−0.72, 0.07)	(−3.24, −0.55)	(−0.8, 1.79)	(−0.38, 1.44)
	0.94	1	0.24	0.13

All exposure group effects were compared to the baseline Administration, and the effects for each outcome over all exposure groups are shown in Fig. 13.4. BMI was substantively lower in Maintenance, Emergency, and Field Experts. Diastolic blood pressure was substantively lower in Technology and Field Expert exposure groups. Technologists had substantively lower systolic blood pressure, and Technicians had substantively higher FEV_1%.

Figure 13.5 shows the same four selected individuals previously identified as being "at risk", from Fig. 13.3, and the observed and predicted values for health outcomes. Observed outcomes are shown as blue points and model mean posterior

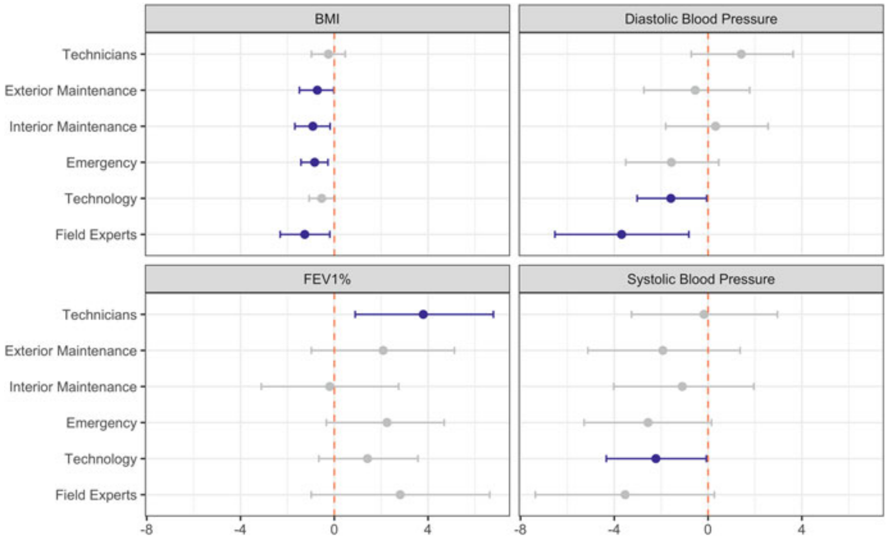

Fig. 13.4 Posterior mean and 95% credible interval for exposure group parameters, for each model. The baseline exposure group is Administration

values are shown as a blue line. The dark blue ribbon around the blue line represents an 80% credible interval, and the light blue ribbon the 95% credible interval. A 1 year forecast is shown as a red line extending from the blue line, and similarly the 80% and 95% credible intervals are displayed. A dotted line is shown for each outcome, which represents a clinically relevant threshold of chronic disease for each outcome. Individual probabilities of chronic condition in the 1 year forecast are labelled directly on Fig. 13.5.

13.4 Discussion

This paper set out to develop a Bayesian approach to analysing OHS data and to illustrate three analytic aims focussed on the health outcomes lung function (FEV_1%), Body Mass Index (BMI), and systolic and diastolic blood pressure. Aim 1 investigated patterns and trends in the health outcomes. Aim 2 explored the effects of smoking and industrial exposures over time for individuals and worker groups. Aim 3 identified future individual risk of chronic conditions.

Aim 1 was addressed by examining individual change over time and identifying individuals with the greatest degree of change over time for each outcome. We demonstrated how one could then assess the overall change over time for these individuals, which could be used to identify trends in other health outcomes. We also identified the proportion of individuals who fell into an “at risk” category.

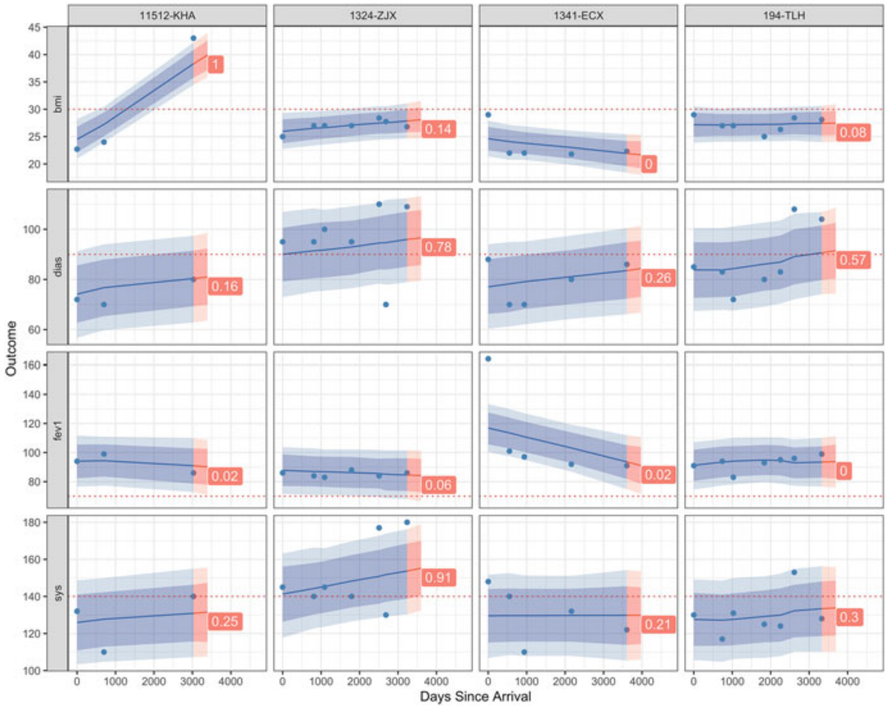

Fig. 13.5 Individuals from Fig. 13.3 and their observed (points), predicted (blue region and line), and forecasted (red region and lines) values for the health outcomes shown with 80% (darker region) and 95% credible intervals (lighter region). Labels show the probability of the individual having a clinically defined chronic disease at the forecasted timepoint

Identifying those individuals with substantive negative change in BMI, lung function, systolic, and diastolic blood pressure means that medical professionals could flag these individuals as at risk (compared to the overall worker population), and provide more frequent medical attention to better monitor their health.

Aim 2 was addressed by examining the probability that the effects of smoking and dust were different from zero for each outcome, finding those exposure groups that were substantively different from the reference group (Administration), and identifying individuals at risk based on substantive change over time in health outcomes. Smoking was associated with negative health outcomes for lung function and systolic and diastolic blood pressure. This information can be used to further support health policies, such as implementation of tobacco bans in the workplace.

The nominated industrial exposure, dust was not substantively associated with health outcomes in the workplaces in this case study. The relevant parameter estimates had quite wide credible intervals, possibly due to interpolation of the data, which was used to align dust measurement points with health measurements. This demonstrates that frequent measurements of industrial exposures of concern can provide more certainty in the measurement of effects. Interestingly, results identified that employees in Administration should be more closely monitored and perhaps should be the focus of health interventions and healthy worker programs in workplaces. Providing descriptive statistics of the outcomes for at risk and not at risk populations (Table 13.3) provides medical professionals with a measure of how meaningfully different these populations are, and facilitates more targeted health and wellness programs.

Aim 3 was addressed by calculating posterior predictions and corresponding intervals and 1 year forecasts for all individuals. This allows medical professionals to assign a probability that an individual might move into an "at risk" category 1 year from their last visit. This means that individuals may be flagged as "at risk" and further action can be taken, perhaps in the form of more frequent medical visits to more closely monitor their health measures. his demonstrates how forecasting could identify "at risk" individuals, by placing a threshold on the probability of health outcomes being medically classified as chronic or acute conditions. Some individuals might cross the threshold over the observed times, whilst others might be predicted to enter the threshold with a given probability over the next year.

The Bayesian hierarchical model accounts for important features of data such as multiple measurements for individuals, and the exposure group structure in the workplace. The definition of a Bayesian credible interval as a range of probable values for a parameter makes it easier to communicate model inferences. The model also provides probabilities of interest directly, conditional on the data. This is a useful complement to credible intervals. Forecasting of future observations in a Bayesian framework also allows for probabilistic statements based directly on the posterior predictive distribution. The Bayesian framework naturally includes additional uncertainty due to imputation of missing values. These features compare favourably to their frequentist modelling counterparts.

Extensions to the Bayesian models developed here are also straightforward. For example, an obvious next step in analysis might be to add interactions into the model, such as smoking and dust, or BMI and blood pressure. One relatively straightforward way to explore the impact of interactions is to evaluate the Bayes factors for each variable, approximated using the Savage-Dickey density ratio, which only requires samples from the posterior [48, 49]. This can add time to the model building process, in terms of deciding upon the most useful model, but is worth the effort if the practitioner is genuinely interested in one or two interaction terms. As with any working population, there may be some healthy cohort effect [50, 51], where, being employable, employees are healthier than the general population. The methods provided in this study identify employees and groups that are different from the population. Combining this information with reference chronic conditions provides a more comprehensive approach which might otherwise have missed healthy employees.

It is also possible that there may perhaps be less measurement error for long term employees; here a model that predicts the number of visits for each individual may be useful, where the number of visits n_{0i} for each individual is the outcome. Additionally, there may be some correlation between the slope and the intercept, which could be accounted for by modelling them as coming from some bivariate normal distribution [46]. It is acknowledged that the model fit is not ideal, particularly with respect to underestimation and overestimation of very high and low values, respectively. While this regression to the mean is to be expected given the random effects terms, the fit could be improved for the other outcomes, perhaps by including interactions as previously discussed.

As far as we are aware, this is the first time Bayesian methods have been applied to this kind of OHS data. It is our hope that this paper can serve as one way to fit and interpret these data, and serve as encouragement for researchers in the field of OHS, to include Bayesian approaches in their analytic toolkit. The ultimate ambition is to provide more informative evidence-based OHS assessments for a healthier workforce and more profitable workplaces.

13.5 Summary

13.5.1 Strengths and Limitations of the Study

- Strength: This is the first application of Bayesian methods to typical data found in occupational health surveillance.
- Strength: The methods used account for important features of data such as multiple measurements for individuals, and the group structure of exposure groups in the workplace.
- Strength: The model allows for groups and individuals to be flagged as "at risk", enabling proactive action on individual health.
- Strength: The definition of a Bayesian credible interval as a range of probable values for a parameter makes it easier to communicate model inferences.
- Strength: The model provides probabilities of interest directly, conditional on the data, which is a useful complement to credible intervals that makes effects and uncertainty simpler and easier to communicate to health practitioners.
- Limitation: No account was taken of the healthy worker effect, so whilst the focus of the paper is on employees rather than the general population, the analysis may be biased if healthier employees remain longer in the industry.
- Limitation: The model used vague priors, and so future work could explore the use of more informative priors, based on, for example, previous data collected in similar fields.

Acknowledgments The authors would like to thank Dr. Xing Lee and Dr. Nicole White for their advice and collaboration with this project.

Data Sharing For reasons of employee confidentiality and industry sensitivity, the dataset used in the case study is not available for sharing. However, the seminal features of the dataset are well described in the paper. Code used for the statistical data analysis for this work can be found at: https://github.com/njtierney/njtbatohs-chapter.

Contributors NJT conducted literature survey, statistical analysis, created visualisations, and wrote the first draft. SC provided critical feedback, assisted in developing the statistical model and

in presenting results. CCD provided assistance in developing the statistical model and in critical feedback of the paper. KLM provided initial description of the model, assisted in developing the code, and provided critical feedback of the paper and presentation of results. All authors approved the version of the paper for publishing, and agreed to respond to questions that may arise regarding integrity of the work.

Funding This research was jointly funded by an Australian Postgraduate Award (APA), the Australian Technology Network Industry Doctoral Training Centre (IDTC), the Australian Research Council, and the ARC Centre of Excellence for Mathematical and Statistical Frontiers. CCD was supported by an Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE160100741).

Competing Interests None declared.

Ethics Approval The QUT University Human Research Ethics Committee.

References

1. P.A. Schulte, D. Trout, R.D. Zumwalde, E. Kuempel, C.L. Geraci, V. Castranova, et al., Options for occupational health surveillance of workers potentially exposed to engineered nanoparticles: state of the science. J. Occup. Environ. Med./Am. Coll. Occup. Environ. Med. **50**(5), 517–26 (2008)
2. S.W. Australia, *Australian Work Health and Safety Strategy 2012–2022: Healthy, Safe and Productive Working Lives* (Australian Government - Safe Work Australia, Canberra, 2012)
3. I. Firth, *Simplified Occupational Hygiene Risk Management Strategies*, Tullamarine. (Australian Institute of Occupational Hygienists, Westmeadows, 2006)
4. L. Lewis, D. Fishwick, Health surveillance for occupational respiratory disease. Occup. Med. **63**(5), 322–34 (2013)
5. S. Banerjee, G. Ramachandran, M. Vadali, J. Sahmel, Bayesian hierarchical framework for occupational hygiene decision making. Ann. Occup. Hyg. **58**(9), 1079–93 (2014)
6. J. Lynch, G.D. Smith, A life course approach to chronic disease epidemiology. Annu. Rev. Public Health **26**, 1–35 (2005)
7. C. Duncan, K. Jones, G. Moon, Context, composition and heterogeneity: using multilevel models in health research. Soc. Sci. Med. **46**(1), 97–117 (1998)
8. R.T. McKay, G.K. LeMasters, T.J. Hilbert, L.S. Levin, C.H. Rice, E.K. Borton, et al., A long term study of pulmonary function among US refractory ceramic fibre workers. Occup. Environ. Med. **68**(2), 89–95 (2011)
9. E. Algranti, E.M.C. Mendonça, E. Hnizdo, E.M. De Capitani, J.B.P. Freitas, V. Raile, et al., Longitudinal decline in lung function in former asbestos exposed workers. Occup. Environ. Med. **70**(1), 15–21 (2013)
10. R.D. Szczesniak, G.L. McPhail, L.L. Duan, M. Macaluso, R.S. Amin, J.P. Clancy, A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. Ann. Epidemiol. **23**(12), 771–7 (2013)
11. E.A.J. Cook, W.A. de Glanville, L.F. Thomas, S. Kariuki, B.M. de Clare Bronsvoort, E.M. Fèvre, Risk factors for leptospirosis seropositivity in slaughterhouse workers in western kenya. Occup. Environ. Med. **74**, 357–365 (2016)
12. F. de Vocht, H. Kromhout, G. Ferro, P. Boffetta, I. Burstyn, Bayesian modelling of lung cancer risk and bitumen fume exposure adjusted for unmeasured confounding by smoking. Occup. Environ. Med. **66**(8), 502–8 (2009)
13. P.S. Bakke, V. Baste, R. Hanoa, A. Gulsvik, Prevalence of obstructive lung disease in a general population: relation to occupational title and exposure to some airborne agents. Thorax **46**(12), 863–70 (1991)

14. G.W. Gibbs, B. Armstrong, M. Sevigny, Mortality and cancer experience of quebec aluminum reduction plant workers. Part 2: mortality of three cohorts hired on or before January 1, 1951. J. Occup. Environ. Med./Am. Coll. Occup. Environ. Med. **49**(10), 1105–23 (2007)
15. J. Hellgren, L. Lillienberg, J. Jarlstedt, G. Karlsson, K. Torén, Population-based study of non-infectious rhinitis in relation to occupational exposure, age, sex, and smoking. Am. J. Ind. Med. **42**(1), 23–8 (2002)
16. K. Radon, U. Gerhardinger, A. Schulze, J.-P. Zock, D. Norback, K. Toren, et al., Occupation and adult onset of rhinitis in the general population. Occup. Environ. Med. **65**(1), 38–43 (2008)
17. A. Olsson, H. Kromhout, M. Agostini, J. Hansen, C.F. Lassen, C. Johansen, et al., A case-control study of lung cancer nested in a cohort of european asphalt workers. Environ. Health Perspect. **118**(10), 1418–24 (2010)
18. K. Shao, B.C. Allen, M.W. Wheeler, Bayesian hierarchical structure for quantifying population variability to inform probabilistic health risk assessments. Risk Anal. **37**, 1865–1878 (2016); An official publication of the Society for Risk Analysis
19. S.M. Bartell, G.B. Hamra, K. Steenland, Bayesian analysis of silica exposure and lung cancer using human and animal studies. Epidemiology **28**(2), 281–7 (2017)
20. W.K. Sieber, C.F. Robinson, J. Birdsey, G.X. Chen, E.M. Hitchcock, J.E. Lincoln, et al., Obesity and other risk factors: the national survey of U.S. Long-haul truck driver health and injury. Am. J. Ind. Med. **57**(6), 615–26 (2014)
21. D. Fekedulegn, C.M. Burchfiel, L.E. Charles, T.A. Hartley, M.E. Andrew, J.M. Violanti, Shift work and sleep quality among urban police officers: the BCOPS study. J. Occup. Environ. Med./Am. Coll. Occup. Environ. Med. **58**(3), e66–71 (2016)
22. D.J. Hendrick, M.J. Connolly, S.C. Stenton, A.G. Bird, I.S. Winterton, E.H. Walters, Occupational asthma due to sodium iso-nonanoyl oxybenzene sulphonate, a newly developed detergent ingredient. Thorax **43**(6), 501–502 (1988)
23. S.C. Stenton, J.H. Dennis, E.H. Walters, D.J. Hendrick, Asthmagenic properties of a newly developed detergent ingredient: sodium iso-nonanoyl oxybenzene sulphonate. Br. J. Ind. Med. **47**(6), 405–10 (1990)
24. S. Steinle, S. Reis, C.E. Sabel, S. Semple, M.M. Twigg, C.F. Braban, et al., Personal exposure monitoring of PM2.5 in indoor and outdoor microenvironments. Sci. Total Environ. **508**, 383–94 (2015)
25. J.P. Hobert, G. Casella, The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J. Am. Stat. Assoc. **91**(436), 1461–73 (1996)
26. F.J. Rubio, M.F.J. Steel, Flexible linear mixed models with improper priors for longitudinal and survival data. Electron. J. Stat. **12**(1), 572–98 (2018)
27. A. Gelman, Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian Anal. **1**(3), 515–34 (2006).
28. R Core Team, R: A language and environment for statistical computing [Internet] (R Foundation for Statistical Computing, Vienna, 2016). Available from: https://www.R-project.org/
29. H. Wickham, R. Francois, Dplyr: A grammar of data manipulation [Internet] (2017). Available from: https://github.com/hadley/dplyr
30. H. Wickham, Tidyr: Easily tidy data with 'spread()' and 'gather()' functions (2017)
31. H. Wickham, Readxl: Read excel files [Internet] (2016). Available from: https://CRAN.R-project.org/package=readxl
32. H. Wickham, Purrr: Functional programming tools [Internet] (2016). Available from: https://CRAN.R-project.org/package=purrr
33. G. Grolemund, H. Wickham, Dates and times made easy with lubridate. J. Stat. Softw. [Internet] **40**(3), 1–25 (2011). Available from: http://www.jstatsoft.org/v40/i03/
34. M. Dowle, A. Srinivasan, T. Short, R Saporta SL with contributions from, E. Antonyan, Data.table: Extension of data.frame [Internet] (2016). Available from: http://r-datatable.com
35. J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, et al., R markdown: Dynamic documents for r [Internet] (2017). Available from: http://rmarkdown.rstudio.com

36. Y. Xie, Dynamic documents with R and knitr [Internet], 2nd edn. (Chapman Hall/CRC, Boca Raton/Florida, 2015) Available from: http://yihui.name/knitr/
37. M. Plummer, et al., JAGS: A program for analysis of bayesian graphical models using Gibbs sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna (2003), p. 125
38. M. Plummer, Rjags: Bayesian graphical models using MCMC [Internet] (2016). Available from: https://CRAN.R-project.org/package=rjags
39. W.A. Link, M.J. Eaton, On thinning of chains in MCMC: thinning of MCMC chains. Methods Ecol. Evol./British Ecol. Soc. **3**(1), 112–115 (2012)
40. Stan Development Team, RStan: The R interface to Stan [Internet] (2018). Available from: http://mc-stan.org/
41. D.J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. **10**(4), 325–37 (2000)
42. P. de Valpine, D. Turek, C.J. Paciorek, C. Anderson-Bergman, D.T. Lang, et al., Programming with models: writing statistical algorithms for general model structures with nimble. J. Comput. Graph. Stat. **26**(2), 403–13 (2017)
43. N. Golding, Greta: Simple and scalable statistical modelling in R [Internet] (2018). Available from: https://github.com/greta-dev/greta
44. D. Lunn, C. Jackson, N. Best, A. Thomas, D. Spiegelhalter, *The Bugs Book: A Practical Introduction to Bayesian Analysis* (CRC press, Boca Raton, 2012)
45. X. Fernández-i-Marín, “Ggmcmc”: analysis of mcmc samples and Bayesian inference. J. Stat. Softw. **70**(9), 1–20 (2016)
46. A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, vol. 2 (Chapman & Hall/CRC, Boca Raton, 2014)
47. J. Besag, P. Green, D. Higdon, K. Mengersen, Bayesian computation and stochastic systems. Stat. Sci. **10**(1), 3–41 (1995); A Review Journal of the Institute of Mathematical Statistics
48. E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal, R. Grasman, Bayesian hypothesis testing for psychologists: a tutorial on the Savage–Dickey method. Cogn. Psychol. **60**(3), 158–89 (2010)
49. J.M. Dickey, The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Math. Stat. **42**, 204–223 (1971)
50. C.Y. Li, F.C. Sung, A review of the healthy worker effect in occupational epidemiology. Occup. Med. **49**(4), 225–9 (1999)
51. A.J. McMichael, R. Spirtas, L.L. Kupper, An epidemiologic study of mortality within a cohort of rubber workers, 1964–1972. J. Occup. Med. **16**(7), 458–64 (1974); Official Publication of the Industrial Medical Association