Forecasting Exposure to Occupational Back Pain: 2013 – 2040

Kyle R. Heuton

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2015

Committee:

Christopher J.L. Murray

Haidong Wang

Kyle Foreman

Program Authorized to Offer Degree:

Global Health

University of Washington

Abstract

Forecasting Exposure to Occupational Back Pain: 2013 – 2040

Kyle R. Heuton

Chair of the Supervisory Committee:

Christopher J.L. Murray, MD, DPhil

Department of Global Health

A new model is proposed and evaluated to produce forecasts to 2040 of exposure to occupational back pain, a risk factor in the Global Burden of Disease 2015 Study's comparative risk assessment. Summary exposure values are created from the population distribution of exposures for 188 countries, 13 age groups, and both sexes. The model is a Bayesian autoregressive model with a Leroux continuous autoregressive prior to smooth over both age groups and location. Global exposure to ergonomic factors that cause occupational back pain decrease for every country and both sexes. All-ages male exposure decreases from a mean of 0.19 in 2013 to a mean of 0.17 in 2040, while all-ages female exposure decreases from a mean of 0.17 to 0.15 in 2040. These forecasts will allow for the creation of detailed burden of disease forecasts that include risk attributions. Additionally, this forecasting framework allows for investigation of hypothetical policies and their influence on burden through reduction in exposure to occupational risk factors.

Introduction

Continuing in the tradition of the Global Burden of Disease 1990 Study (GBD 1990) [1], the 2015 iteration of the Global Burden of Disease Study (GBD 2015) will again include projections of mortality and disability into the future. This paper focuses on the methods and results responsible for a small section of those projections, the forecasts of exposure to occupational risks resulting in Low Back Pain (LBP). LBP is a common health condition [2] with large economic implications, causing loss of income for individuals and negatively impacting a society's available workforce. [3] Of the 291 conditions reported in the Global Burden of Disease 2010 Study (GBD 2010) revealed that LBP is the largest contributor to disability in the world. [4]

Previous Projections of Global Measures of Health

Forecasts of global disability and mortality were last produced as part of the Global Burden of Disease Study 2004. [5] These forecasts used relatively simple methods for projecting mortality and disability, similar to those used in the original GBD 1990. Mortality was modeled in log space as a linear model according to Equation 1.

$$\ln M_{a,s,c} = C_{a,s,c} + \beta_1 \ln Y + \beta_2 \ln HC + \beta_3 (\ln Y)^2 + \beta_4 T + \beta_2 \ln SI$$

Equation 1: The model used in GBD 2004 for mortality projections.

The dependent variable is the natural log of mortality for a given age a, sex s, and cause-cluster c. Here C is a constant term that also varies by age, sex, and cause, Y represents a country's GDP per capita, HC represents human capital, T the year, and SI is Smoking Impact, used for tobacco and smoking related health outcomes. In addition to only projecting mortality and disability due to 10 high-level "cause-clusters," these projections were reported only at the geographic region level, and were not country specific. Three different "scenarios" were considered in these projections: a baseline, or "business-as-usual" scenario, along with and

optimistic and a pessimistic scenario. For most causes, these scenarios were created by using different covariate values for the baseline, optimistic, and pessimistic scenarios. For example, the baseline scenario used the World Bank's GDP per capita projections, while the pessimistic scenario assumed growth rates at 50% of the baseline. The optimistic scenario assumed growth would be 40% higher than the baseline, except for certain countries with recent rapid growth, such as China and India. For certain causes, such as HIV/AIDS, additional assumptions were made in the creation of optimistic and pessimistic scenarios. The baseline HIV/AIDS scenario assumed 80% ART coverage globally by 2012, while the optimistic scenario also assumed increased prevention activities in addition to 80% ART coverage, and the pessimistic scenario assumed only 60% coverage would be achieved.

Forecasting in the GBD 2015

The work described in this paper comes from a forecasting framework that models at a much more detailed cause level and will report results at the country level. Additionally, models more complex than linear regressions are considered, allowing individual causes to have models that more appropriately capture their unique sources of variance. In addition to modeling mortality and disability at a more detailed level, the GBD 2015 projections will forecast drivers of health and exogenous variables, such as national income, future population and human capital, future levels of educational attainment, and the exposure variables from the GBD 2015 comparative risk assessment. By forecasting all of these variables simultaneously, the GBD 2015 forecasts can go beyond 3 scenarios to any number of detailed scenarios. Health outcomes and how they develop over time can then be projected for specific levels of income, education, or risk exposure.

There is no consensus on the best approach to be used for forecasting mortality and other demographic variables. [6] The broad categories of models typically used include: using a model mortality scenario, target setting and interpolating towards a target mortality, extrapolating

mortality as a function of time without the use of covariates, and the approach used in the GBD 2015, forecasting using covariates as exogenous drivers of mortality. [7] While non-covariate approaches are sometimes preferred for their simplicity and ability to capture underlying trends and ignore noise, several limitations have been noted. These models often fail to adapt to diverse settings, performing well for one cause but not capturing the unique situation of another, or predicting an individual age group well but forecasting an inconsistent age pattern of mortality. [8] While using covariates in forecasting a challenging undertaking that requires forecasting a complete time series for every covariate used, its primary advantage is that it enables the investigation of specific scenarios. It also captures the effects of exogenous shocks, which are often of critical global health importance but missed by models lacking covariates.

Methods

Data Sources

Data on the exposure to ergonomic factors in the workplace that result in LBP were taken from the GBD 2015 study. As the origin of these data are not yet described elsewhere, the methodology behind creating this exposure time series for every country, sex, and age group in GBD 2015 is described here briefly.

Raw data were obtained from the International Labor Organization's (ILO) ILOSTAT Database. [9] The ILO's database contains information on participation in specific industries for select countries, years, sexes, and age groups, as well as the fraction of each country, year, sex, and age group participating in the workforce. However, this dataset is incomplete. To create a complete time series for both sexes from 1990-2015, including all countries and age groups of the GBD 2015 study, a 3 stage modeling procedure was used. First a robust linear regression was performed, predicting the amount of participation in each industry with explanatory variables such as educational attainment, income per capita, population density, and latitude and

longitude for agricultural industries. The resulting residuals from this procedure were then used in a spatio-temporal regression to predict neighboring residuals. These predicted residuals were then used as a mean function in a Gaussian Process Regression (GPR) used to draw on uncertainty from the input data as well as priors on its smoothness to produce the final estimates. This procedure closely mirrors the one used for an individual model in the GBD 2015's Cause of Death (COD) modeling strategy. [10] Beyond mortality, this approach has proved useful in forecasting covariates such as tobacco consumption and smoking prevalence. [11] This procedure results in 1000 draws of exposure for every location, sex, age group, and year. Exposure values have a lower bound of 1.0, where 1.0 is the theoretical minimum level of exposure. These 1000 draws create an exposure curve representing the distribution of exposure for members of each location, sex, age group, and year population.

After this procedure is finished, exposure data is available for 188 countries, 13 5-year age groups, both sexes, and all years 1990-2013. Because this is an occupational risk factor, data do not exist for the age groups below age 15. The 13 age groups available for this dataset are the 5-year age groups beginning at 15-19 year olds, and ending with the terminal age group, 80+.

Summary Exposure Variable Creation

In order to forecast future exposure, the exposure distributions calculated with the method described above were simplified to a single value, a Summary Exposure Value (SEV). SEV's were created by integrating over the exposure curve, and multiplying by that risk factor's relative risk. The relative risks used were taken from the GBD 2010's Comparative Risk Assessment (CRA). [12] These relative risks were taken from the best existing meta-analyses from studies with a direct clinical and causal pathway exists between exposure and health outcomes. This integration procedure turns the distribution into a single number for ever location, age, sex, and year. This value is then scaled, so that a population with the theoretical maximum exposure

would have an SEV of 1.0, and all other SEVs are scaled proportionally. The equation for calculating SEV is presented below in Equation 2.

$$SEV_{l,a,s,t} = \frac{\int p\left(Y_{l,a,s,t}\right) RR(Y_{l,a,s,t})\, dY_{l,a,s,t}}{RR^{max} - 1}$$

Equation 2: The formula used for calculating Summary Exposure Values

Here $SEV_{l,a,s,t}$ represents the SEV for a given location l, age group a, sex s, and year t. It is calculated by integrating over the proportion of the population exposed $p\left(Y_{l,a,s,t}\right)$ multiplied by the relative risk at that exposure level $RR(Y_{l,a,s,t})$ over all levels of exposure, and then scaling by the maximum observed relative risk for any location, year, age, and sex, minus 1.

This results in 1000 draws of an SEV for ever location, age, sex, and year. Using the SEV allows for modeling a single value over time, rather than modeling an entire distribution of different exposure levels.

	Male	Female
Occupational Back Pain SEV Mean (95% CI)	0.19 (0.09 – 0.55)	0.17 (0.029 – 0.63)

Table 1: Global mean and 95% confidence intervals for all-ages exposure to occupational back pain in 2013

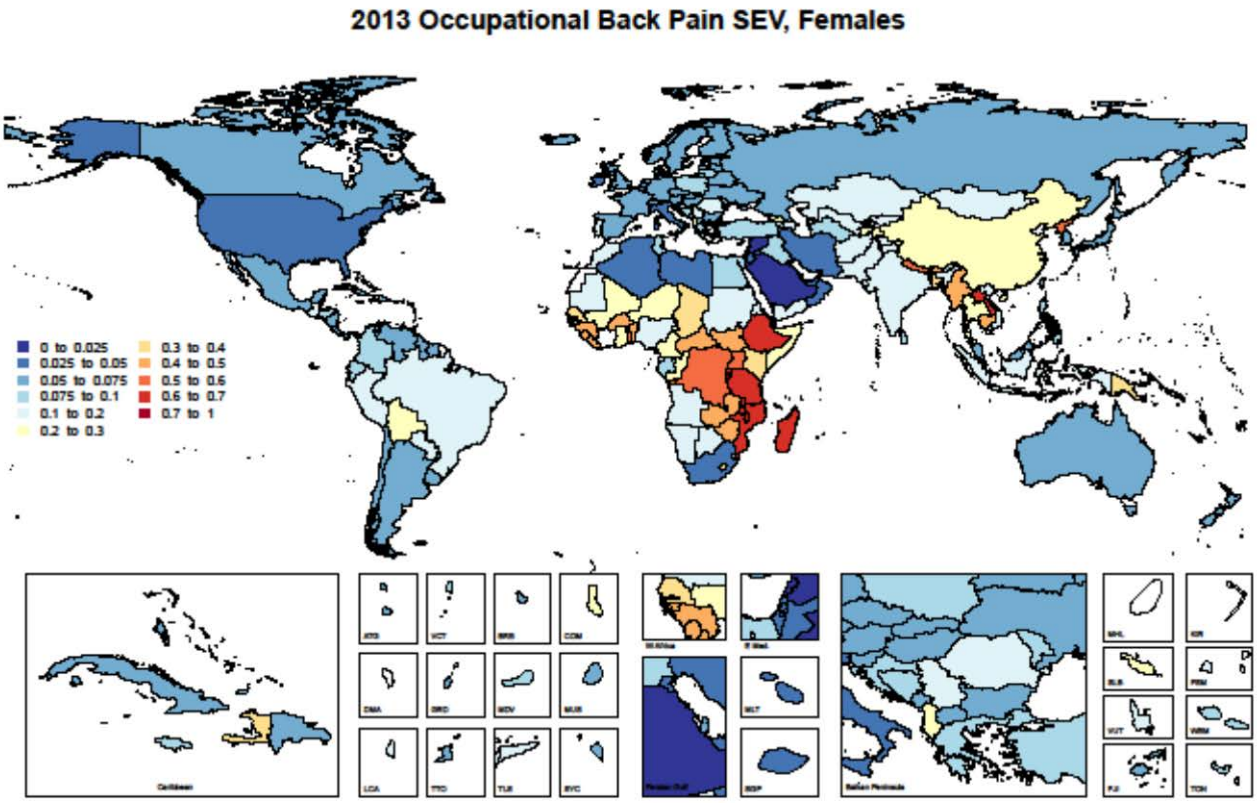

Figure 1: Map of all-ages exposure to occupational back pain SEV values for females in 2013

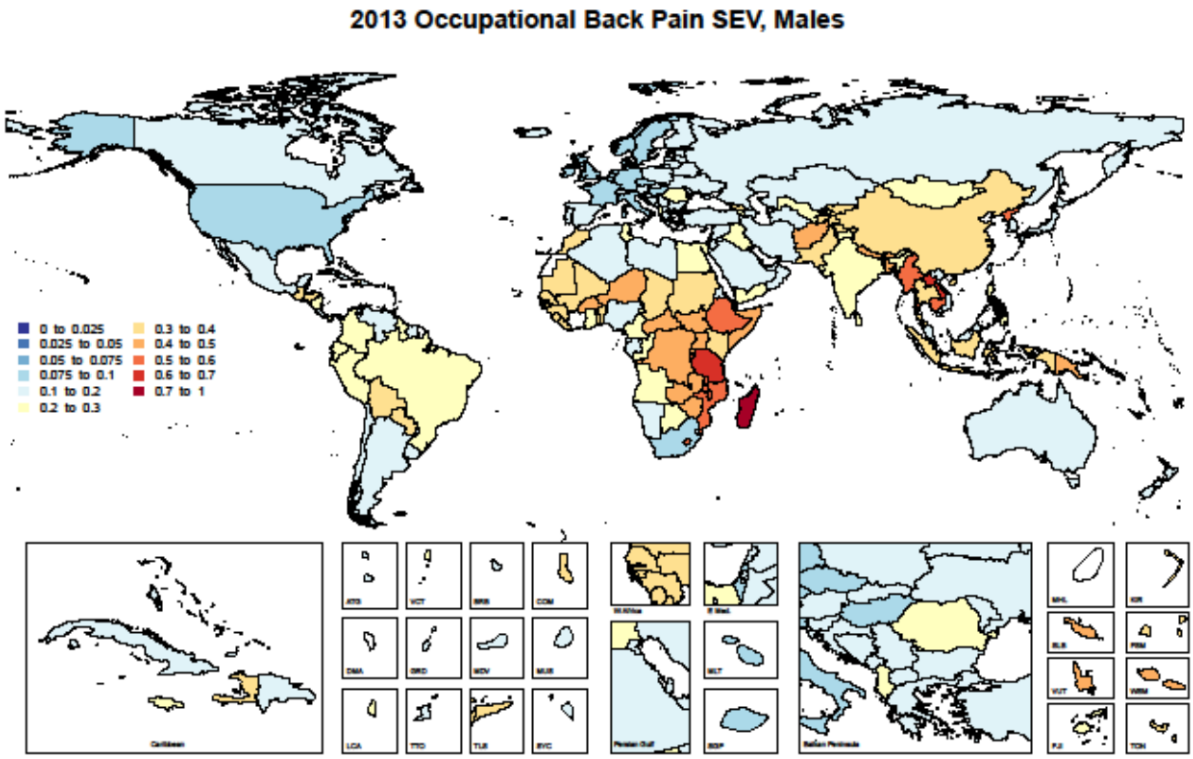

Figure 2: Map of all-ages exposure to occupational back pain SEV values for males in 2013

Descriptive summary statistics for the SEV values in the last year with historical data, 2013, are presented by sex in Table 1. All-ages values were calculated by taking the population weighted average of the individual age groups' SEV's. Although individual countries have SEV's above 0.7, the mean is only 0.19 for males and 0.17 for females, with 95 percentiles of 0.55 and 0.63 respectively. SEV levels by country are shown for males in Figure 1 and females in Figure 2. Here the highest SEV values are observed for Sub-Saharan Africa and South East Asia. This is largely due to these region's high population participation in agricultural industry, and the high relative risk associated with agricultural work.

Anchor Model

To forecast this SEV into the future, we developed the Anchor Model, a Bayesian autoregressive model with a LeRoux Conditdional AutoRegressive prior on age and location (LCAR). [13] The

dependent variable used in the model was the logit of SEV is shown below in Equation 3. Data from each sex were modeled separately.

$$y_{l,a,t} = logit\left(SEV_{l,a,t}\right)$$

Equation 3: The Anchor Model's dependent variable

$$y_{l,a,t} \sim N\left(\mu_{l,a,t}, \sigma\right)$$

Equation 4: The data likelihood function used in the Anchor Model

This data-likelihood function used for this Bayesian model assumes that this dependent variable is normally distributed about some mean μ with variance σ, as shown in Equation 4.

The mean was modeled as being normally distributed about a piecewise linear autoregressive model centered on an "anchor" year. This anchor year was typically selected to be the last year in which historical data are present, which was 2013 for this study. By centering the model on this value, it ensures that the forecasted results do not have a discontinuity with the historical data when predictions are made.

$$\begin{cases} \mu_{l,a,t} \sim N\left(\kappa\mu_{l,a,t+1} - \ \left(\beta\Delta_t^{t+1}X + \theta_{la}\right), \tau\right) \ if \ t < anchor \\ \mu_{l,a,t} \sim N\left(\kappa\mu_{l,a,t-1} + \ \left(\beta\Delta_t^{t+1}X + \theta_{la}\right), \tau\right) \ if \ t > anchor \end{cases}$$

Equation 5: The formulation of the Anchor Model.

The model is written above in Equation 5. Here the mean μ is normally distributed about a linear function that has an auto-regressive term on the previous year's mean with coefficient κ, the contribution of covariates X with coefficient β, and random intercept θ. All variables are indexed with subscripts l for location, a for age, and t for year. Using a Bayesian framework to estimate this model allowed for the use of priors to inform the sign and magnitude of different parameters in the model.

$$\log\sigma \sim U(-\infty, \infty)$$

$$\beta \sim U(-\infty, \infty)$$

$$\tau \sim G(1, 0.01)$$

$$\theta_{la} \sim LCAR(\rho_L, \rho_A, \delta)$$

$$logit(\rho_L) \sim U(-\infty, \infty)$$

$$logit(\rho_A) \sim U(-\infty, \infty)$$

$$\delta \sim G(1, 1)$$

Equation 6: Priors used in the estimation of the Anchor Model

The priors used in the estimation of the Anchor Model are listed in Equation 6. We assumed a uniform distribution on the log of the variance to enforce the variance to be positive. Similarly, a uniform distribution was used as the prior for the covariate coefficients β, placing no preference for any sign or magnitude. The variance of the mean is modeled with a gamma distribution prior with shape 1.0 and scale 0.01. The LCAR prior on the location and age random effect allows the model to simultaneously smooth over both age and location, with the delta prior controlling which dimension, age or location, is smoothed more heavily. The variables ρ_L and ρ_A correspond to relative strength of correlation in each dimension. If there exists strong correlation across age groups, ρ_A will be close to 1, and close to 0 in the absence of smooth age group trends. Similarly, ρ_L represents the strength of spatial correlation. Accordingly, the logit of these variables are modeled with a uniform prior over the real numbers, enforcing only that they be between 0 and 1. Finally, a Gaussian distribution with shape 1.0 and scale 1.0 is used as a prior for δ, reflecting the prior belief that smoothing across age should not be inherently stronger than smoothing across space, nor is there a belief that smoothing across space must necessarily be stronger than smoothing across age.

Covariate Selection

One of the major strengths of the forecasting work being done as part of the GBD 2015 is the increased number of exogenous variables and other drivers of health used for forecasting health

outcomes. This is reflected in this model by expanding on the covariates used in [5] to include both the natural log of income per capita and educational attainment. Additionally when forecasting risk factors, any mediating risk in the causal pathway was also included as a covariate, and forecasted simultaneously with each risk factor. However, occupational exposure to ergonomic factors that cause LBP is the only risk factor in the GBD 2015's comparative risk assessment that impacts LBP, so no other risk factors were used as covariates in this model.

Income was forecasted with an autoregressive model that used workforce participation as its only exogenous covariate. Workforce participation was defined as proportion of the population aged 20-64. Age specific population forecasts were taken from the UN's World Population Prospects. [14] Similarly, educational attainment was forecast as an age-specific autoregressive model using both forecasted population and the income. The models for income and education are both described elsewhere in as of yet unpublished sources.

Model Validation

The accepted method of model validation is using out-of-sample predictive validity; when forecasting, the out-of-sample dataset is created by picking a date in the past and holding out all data after that point. [15] To validate this model, the model was fit using only data from 1990-2005, and its out-of-sample predictive validity was measured using data from 2006-2013. Two out of sample metrics were used: Root Mean Squared Error (RMSE), and coverage. Out-of-sample coverage is defined as the proportion of out-of-sample data that lie within with predicted 95% uncertainty intervals. A coverage greater than 95% implies uncertainty intervals are too large, and a model with less than 95% coverage is inaccurate.

Computation

All models were estimated using Template Model Builder (TMB) [16] using a newly developed Python interface for TMB, PyMB. [17]

Results

Exposure to ergonomic factors that cause LBP was forecast for 188 countries, 13 age groups, and both sexes. Summary statistics for 2040 are presented below in Table 2.

	Male	Female
Occupational Back Pain SEV Mean (95% CI)	0.17 (0.03 – 0.51)	0.15 (0.02 – 0.63)

Table 2: Global mean and 95% confidence intervals for all-ages exposure to occupational back pain in 2040.

The mean male SEV of 0.17 and the mean female SEV of 0.15 in 2040 are both lower than their respective mean values in 2013. In fact, every country modeled experiences a decline in occupational back pain SEV. This is due to the anchor model's autoregressive term being driven by historical declines in occupational back pain SEV between 1990 and 2013 and carrying the trend forward. As agricultural workers have the highest relative risk for contracting LBP from work, these declines make sense given the shrinking percentage of the population involved in agriculture. [18]

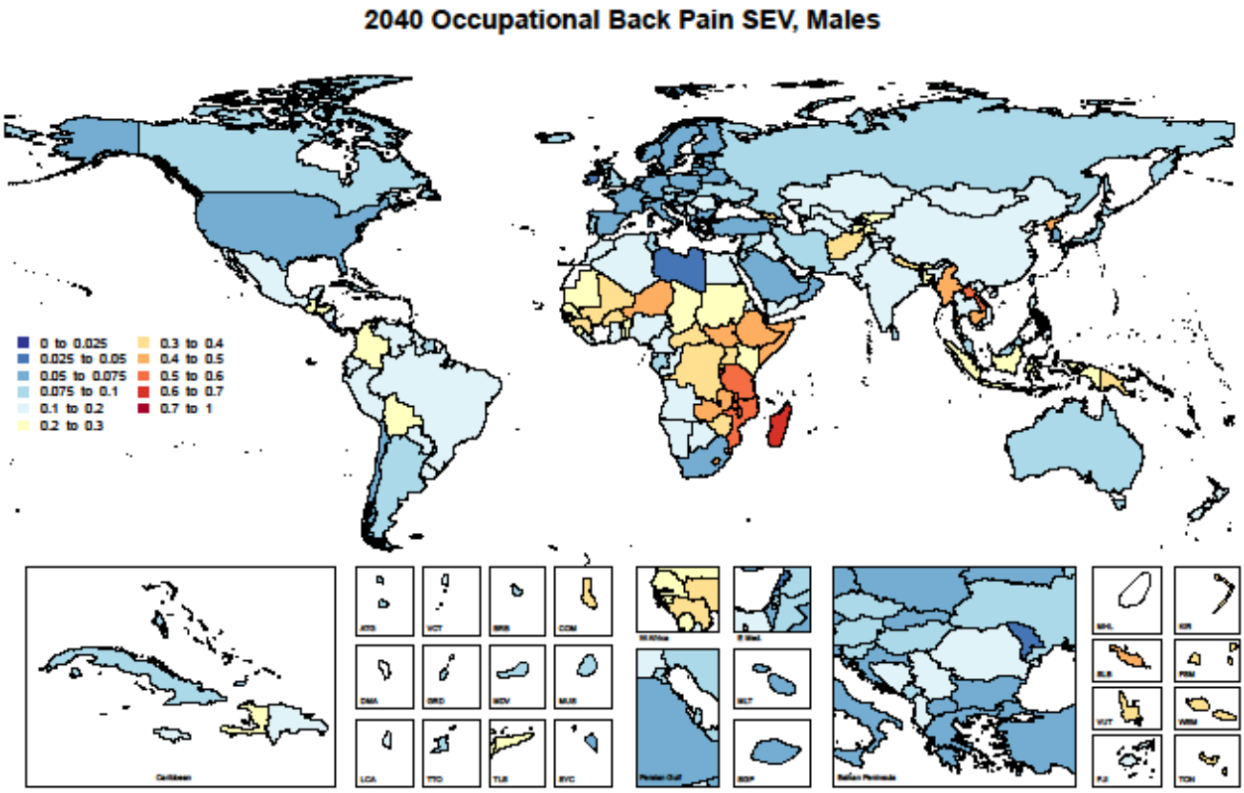

Figure 3: Map of all-ages exposure to occupational back pain SEV values for males in 2040

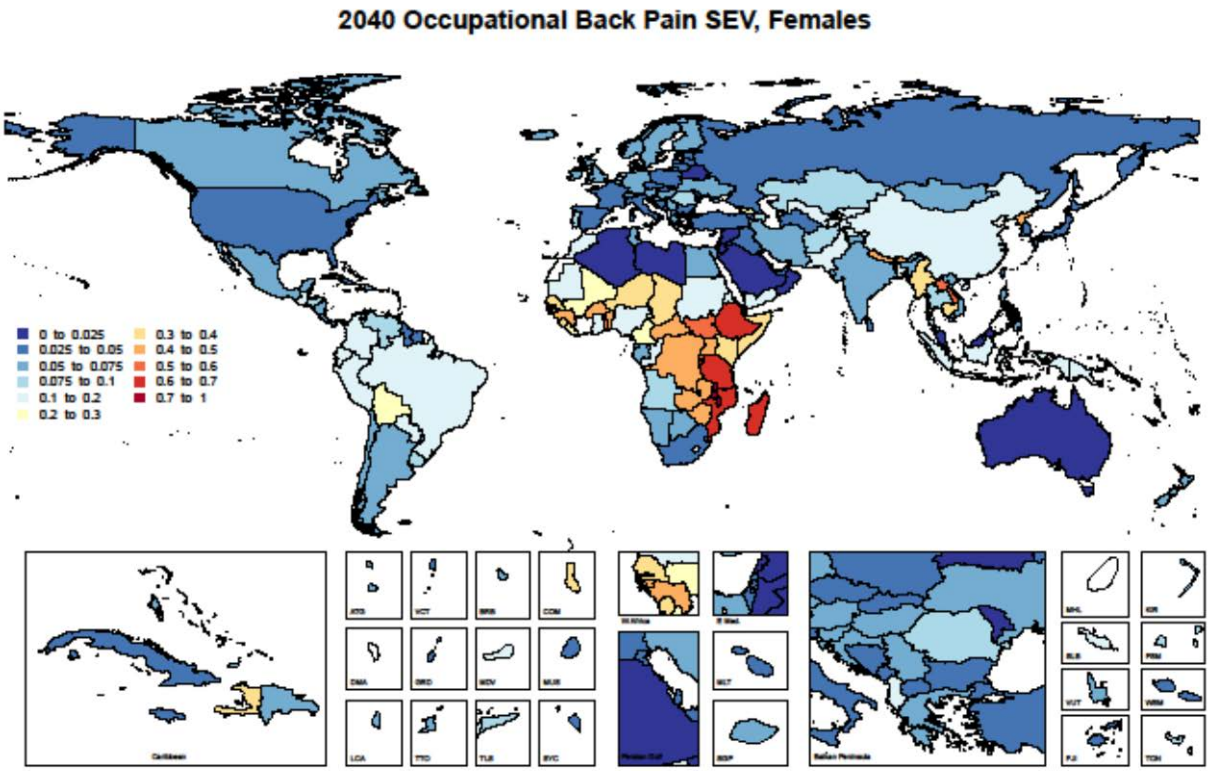

Figure 4: Map of all-ages exposure to occupational back pain SEV values for females in 2040

Figure 3 and Figure 4 show the all-ages SEV levels in each country in 2040 for males and females, respectively. From these maps we can see that the SEV level noticeably decreases for females in Southeast Asia by 2040. For example, female all-ages SEV in China decreases from 0.28 in 2013 to 0.15 in 2040. Eastern Sub-Saharan Africa shows a much slower rate of decline. Female all-ages SEV in Tanzania only decreases from 0.69 in 2013 to 0.64 in 2040. In 2040 Eastern Sub-Saharan Africa remains the region with the highest exposure to occupational ergonomic factors that result in LBP.

Model Validity Results

The Anchor Model was evaluated by fitting the model to the 1990 – 2005 data, and calculating RMSE and coverage on the data from 2006 to 2013.

	Male	Female
RMSE	0.016	0.019
Coverage	76%	76%

Table 3: RMSE and coverage on data from 2006 – 2013 used as out-of-sample predictive validity.

Table 3 shows the predictive validity metrics obtained for this model. RMSE was calculated as deviation from the actual SEV, rather than the logit SEV. This eases interpretability, as SEV is bounded by 0 and 1, and logit SEV can be any real number. The RMSE over all countries and ages are remarkably low, at only 0.016 for males and 0.019 for females. This seems contradictory with the low coverage of just 76% for both male and female models. However, this is explained by the model performing in an all-or-nothing fashion, either the model is accurate and correctly models all data points for a country-age with low RMSE and 100% coverage, or it incorrectly predicts the trend, and forecasts with low RMSE but completely misses the data for almost 0% coverage.

Conclusions and Future Work

In this paper we present a method for forecasting exposure to a risk factor into the future using a robust Bayesian autoregressive model and forecasts of various drivers of health. Although forecasted results from only one risk factor are shown, this model has the potential to produce forecasts of every risk factor in the GBD 2015 comparative risk assessment. Forecasts of exposure to these risk factors are a vital input to forecasting risk attributable mortality and calculating risk attributable burden in the future.

Beyond forecasting burden, this new approach to forecasting the individual risk factors used in the comparative risk assessment allows for policy makers and academics to explore interesting counterfactuals using burden of disease data. An effort could be made to study the effect of different labor laws on a country's exposure to these occupational risk factors. Then, a hypothetical future could be forecasted where some or all countries adopt these policies, and their exposure to risk factors change over time. This would allow careful study of potential

burden averted by acting on a new policy or burden attributable to failure to act. Additionally, the interconnected nature of these forecasts would allow for the study of much more distal effects of any potentially burden-reducing policy. If a policy were to avert some disability or mortality for young people, it will naturally affect the population and burden of disease for older populations in the future. This modeling framework will allow for the exploration of these counterfactuals, and enable policy makers to use the best available data and methods to make better informed decisions.

Acknowledgements

I would like to thank the Institute for Health Metrics and Evaluation, whose Post-Bachelor Fellowship made my work and this thesis a possibility. I would like to acknowledge Kyle Foreman for developing the model discussed in this paper. Special thanks to my colleague Patrick Reidy for developing the code used to produce these results, assistance debugging my own code, and help working through some of the mathematics involved. I thank Astha KC for her instrumental advice and documentation used to understand the GBD approach to modeling occupational risks. Additionally, I would like to acknowledge specifically the IT team at IHME: Serkan Yalcin, Andrew Ernst, Bill Britt, Khan Mai, and others whose tireless efforts resulted in the computational resources necessary for the production of this work. Finally, I would like to thank my wife, Amy, whose love and support made completion of this degree program possible.

References

[1] C. J. L. Murray and A. D. Lopez, The Global Burden of Disease and Injury Volume 1: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020, Cambridge: Harvard University Press, 1996.

[2] D. Hoy, P. Brooks, F. Blyth and R. Buchbinder, "The Epidemiology of low back pain.," *Best Pract Res Clin Rheumatol.,* vol. 24, no. 6, pp. 769-81., 2010.

[3] L. Lidgren, "The bone and joint decade 2000-2010.," *Bull World Health Organ,* vol. 81, no. 9, p. 629, 2003.

[4] D. Hoy, L. March, P. Brooks, F. Blyth, A. Woolf, C. Bain, G. Williams, E. Smith, T. Vos, J. Barendregt, C. Murray, R. Burstein and R. Buchbinder, "The global burden of low back pain: estimates from the Global Burden of Disease 2010 study," *Annals of the Rheumatic Diseases,* 2014.

[5] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLOS Medicine,* vol. 3, no. 11, p. e422, 2006.

[6] H. Wang and S. H. Preston, "Forecasting United States mortality using cohort smoking histories," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 106, no. 2, pp. 393-398, 2009.

[7] G. d. L. J and T. I, "Methods of mortality projections and forecasts," in *National Population Forecasting in Industrialized Countries*, Amsterdam, Swets and Zeitlinger, 1992, pp. 61-74.

[8] F. Girosi and G. King, "Methods without Covariates," in *Demographic Forecasting*, Princeton, Princeton University Press, 2--8, pp. 21-42.

[9] International Labor Organization, "ILOSTAT Database," 31 July 2015. [Online]. Available: www.ilo.org/ilostat. [Accessed 10 August 2015].

[10] K. J. Foreman, R. Lozano, A. D. Lopez and C. J. Murray, "Modeling causes of death: an integrated approach using CODEm," *Population Health Metrics,* vol. 10, no. 1, 2012.

[11] N. M, F. MK, F. TD and e. al., "Smoking Prevalence and Cigarette Consumption in 187 Countries, 1980-2012," *Journal of the American Medical Association,* vol. 311, no. 2, pp. 183-192, 2014.

[12] G. 2. Collaborators, "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet,* vol. 380, no. 9859, p. 2224–2260, 2012.

[13] Y. C. MacNab, "On Gaussian Markov random fields and Bayesian Disease Mapping," *Statistical Methods in Medical Research,* vol. 20, pp. 49-68, 2011.

[14] United Nations Population Division, "World Population Prospects, the 2015 Revision," 29 July 2015. [Online]. Available: http://esa.un.org/unpd/wpp/. [Accessed 12 August 2015].

[15] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting,* vol. 16, no. 4, pp. 437-450, 2000.

[16] K. Kristensen, U. H. Thygesen, K. H. Andersen and J. E. Beyer, "Estimating spatio-temporal dynamics of sizestructured populations," *Canadian Journal of Fisheries and Aquatic Sciences,* vol. 71, pp. 326-336, 2013.

[17] K. Foreman, "PyMB," 14 July 2015. [Online]. Available: https://github.com/kforeman/PyMB. [Accessed 12 August 2015].

[18] The World Bank, "3.2 World Development Indicators: Agricultural Inputs," 2015. [Online]. Available: http://wdi.worldbank.org/table/3.2. [Accessed 10 August 2015].