Study Protocol

Bayesian Autoregressive Multilevel Modeling of Burden of Diseases, Injuries and Risk Factors in Iran 1990 – 2013

Amir Kasaeian PhD Candidate[1,2], Mohammad Reza Eshraghian PhD[1], Abbas Rahimi Foroushani PhD[1], Sharareh R. Niakan Kalhori PhD[2,3], Kazem Mohammad PhD•[1], Farshad Farzadfar MD MPH DSc•[2,4]

Abstract

Background: Statistical modeling and developing new methods for estimating burden of diseases, injuries and risk factors is a fundamental concern in studying the country health situation for better health management and policy making. Bayesian autoregressive multilevel model is a strong method for this kind of study though in complex situations it has its own challenges. Our study aims to describe the way of modeling space and time data through an autoregressive multilevel model and address challenges in complex situation.

Method: We will obtain data from different published and unpublished secondary data sources including population-based health surveys (e.g. NHS, DHS, STEP) at national and provincial levels and we also assess epidemiological studies via systematic review for each disease, injuries and risk factor over the period of 1990 – 2013. These data generally have a multilevel hierarchy and also time correlation. However, statistical analysis of diseases, injuries and risk factors data is primarily facing the problem of information scarcity. Data are generally too scarce to ensure reliable estimates in many practical problems. Also, there may be nonlinear changes over time, different kind of uncertainties in data and incompatible geographical data. We describe Bayesian autoregressive multilevel modeling approach that provides a natural solution to these problems through its ability to sensibly combine information from several sources of data and available prior information. In this hierarchy model levels of each hierarchy borrow information from each other and also lower levels borrow information from higher levels. We will fit the model using Markov Chain Monte Carlo (MCMC) methods because of its capabilities and benefits in complex cases.

Discussion: Our analyses will include different existing sources of data in Iran for 24 years through a rational and reasonable model to estimate burden of diseases, injuries and risk factors for Iran at national, regional and provincial levels while considering several kinds of uncertainties. Comprehensive and realistic estimates are always an issue of request that will be obtained through a suitable statistical modeling considering all dimensions and then can be used for making better decision in real situations.

Keywords: Autoregressive time series, burden of diseases, Iran, MCMC, multilevel models, NASBOD

Cite this article as: Kasaeian A, Eshraghian MR, Rahimi Foroushani A, Niakan Kalhori SR, Mohammad K, Farzadfar F. Bayesian autoregressive multilevel modeling of burden of diseases, injuries and risk factors in Iran 1990 – 2013. *Arch Iran Med.* 2014; **17(1):** 22 – 27.

Introduction

Precise assessment of global, regional, and country health conditions and trends is crucial for evidence-based decision making for Public Health.[1] The Global Burden of Disease Study (GBD) is the latest and most reliable analysis to reveal the importance of taking different approaches to the challenges facing global health.[2] The GBD study results provide us a data-rich structure for comparing the effects and burden of different diseases, injuries, and risk factors on premature death and disability between populations.[3–13] But these results are not for within population, which means nothing is known about what's going on within a country, explicitly Iran here. Knowing and comparing health situation within regions and provinces helps to understand the differences and similarities better and also better map the emerging epidemics of diseases which in turn helps health policy makers to be able to allocate resources more efficiently and prevent hazardous effects and extra burden of those diseases. The only study of burden of diseases and injury in Iran dates back to 2003 which was conducted only in six provinces which showed a significant disparity between these six provinces and also indicated a transition from communicable to non-communicable and road traffic injuries.[14]

In line with the GBD study, National and Subnational Burden of Diseases study 2013 (NASBOD) is a systematic effort to efficiently utilize financial and scientific capitals of previous studies in Iran.[15] It also takes into account care systems, current available data on health systems and viable, systematic and relevant nationwide studies carried out in the previous years.

This study is an endeavor to assess and evaluate the burden of diseases at national and provincial levels in Iran by means of the most recent valid and reliable qualitative and quantitative research methods and experiences taken from Global Burden of Diseases 2010 (GBD). Moreover, knowledge, expertise, and skills of health experts and distinguished scholars in the field have been optimally and efficiently used to come up with absolutely precise computa-

Authors' affiliations: [1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran, [2]Non-Communicable Diseases Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran, [3]Social Determinants of Health Research Centre, School of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran, [4]Endocrinology and Metabolism Research Center, Endocrinology and Metabolism Research Institute, Tehran University of Medical Sciences, Tehran, Iran.

•Corresponding author and reprints: Kazem Mohammad PhD, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. Address: Poursina Avenue, P.O. Box 6446, Tehran 14155, Iran. Tel: 021 88951396; Email: mohamadk@tums.ac.ir.

Farshad Farzadfar MD DSc, Non-communicable Diseases Research Center, Endocrinology and Metabolism Research Institute, Tehran University of Medical Sciences, Tehran, Iran. Address: 4th floor, No. 4, Ostad Nejatollahi St, Enqelab Ave, Tehran, Iran. Postal Code: 1599666615, Tel/Fax: 98-21-88913543, E-mail: f-farzadfar@tums.ac.ir.

Accepted for publication: 3 December 2013

tions of burden of diseases.

The most significant benefit of the NASBOD study is the possibility of comparison and contrast of health conditions in different provinces and regions, an advantage or prerogative which is conducive to the fostering health and hygiene in these regions, which provides Health Policy Makers with the necessary documents for better health policy making and resource allocation.

Also, a comparison of the situations at provincial levels will ultimately boost the health condition throughout the country in a fair and balanced manner.

So, the first key step in the NASBOD study is to utilize the statistical modeling and improved methods for estimating time trends of diseases, injuries, and risk factors. We will use the advanced methods and when necessary expand the current methods to develop Bayesian time series multilevel models for 31 provinces from 1990 to 2013. We will present the data, methods and the key finding of these models. We will estimate national, regional, and provincial 1990 – 2013 trends and their uncertainties in population's mean (whatever measure is) of diseases and injuries or risk factors for all provinces in four regions included in NASBOD study to allow meaningful time and provincial comparison. Iran is divided into four regions (eastern south, north and eastern north, west, and center) on the basis of two criteria: epidemiological homogeneity, and geographical contiguity. The study covers urban and rural areas of the country.

The main purpose of this article is to explain a Bayesian autoregressive multilevel model and all its components together with challenges in complex data which will appear in the study of burden of diseases, risk factors and injuries.

Methods

Study design

Statistical analysis of diseases, injuries and risk factors data is primarily facing the problem of information scarcity. Data are generally too scarce to ensure reliable estimate in many practical problems. In the present study there are 24 provinces at the beginning year in 1990, however during a period of 24 years, there are 31 provinces at the ending of this study. This means there should be at least 576 data points that this is very unlikely in a study of diseases, risk factors and injuries. This problem is more serious for the first half of data warehousing interval. It is also obvious that we encounter geographical incompatibility which is the other issue of concern.

We describe a Bayesian autoregressive multilevel modeling approach that provides a natural solution to these problems through its ability to sensibly combine information from several sources of data and available prior information. Such modeling strategies that capture geographical and time patterns in the data will reduce estimation error.

We will develop this model to estimate prevalence of diseases and injuries or mean of risk factors by age group, sex and province over the time period of 1990 – 2013. We do analyze each gender independently and make estimates for each age group-province-year unit.

In this multilevel model provinces are nested in subregions, subregions are nested in regions, and regions nested in country level. Accordingly, lower level(s) borrow information from higher levels and also levels of each hierarchy borrow information from each other. In fact, there is an concurrence for borrowing information depending on the level of availability and scarcity of data so that the richer is the data the less borrowing within and across levels will be needed and vice versa.

The other point is that trends might not be linear over time; this non-linearity will be modeled in the form of a linear trend plus a smooth non-linear trend, both hierarchically.

Also, because of heterogeneity between community-based studies they might have larger variation than nationally representative studies. The model is able to capture this variation through including a time–varying offset for non-provincial data. These variation components were estimate empirically.

Another problem that might occur is the non-linear association between prevalence and/or mean measurements and age since the association might change in different ages especially in older age groups. In such a condition, we will use cubic spline age model or flexible age model to deal with this kind of non-linearity.

We will determine the values of different kinds of uncertainties such as sampling uncertainty in the original data, uncertainty associated with inconsistency between years in national data, uncertainty relevant to data sources that are not provincial, uncertainty associated with statistical methods for crosswalking between prevalence (categorical measure) and mean (continuous measure).

Finally, a Bayesian model with Markov Chain Monte Carlo (MCMC) algorithm will be fitted. Then, samples taken from posterior distribution of model parameters, which represent uncertainties, will be used to achieve posterior distribution of prevalence or mean of each diseases or risk factors. An important benefit of Bayesian models and specifically MCMC fits is that uncertainty reproduces indeed from the data itself in an integrated and direct way. Uncertainty intervals are also computed for prevalence and mean.

Data sources

We will obtain data from different published and unpublished secondary data sources including population-based health surveys (e.g. NHS, DHS, STEP) at national and provincial levels and also epidemiological studies via systematic review for each disease, injury and risk factor. Some data are obtainable from censuses, household expenditure surveys, demographic surveillance, and disease and death registries. Data from systematic review are evaluated via a quality assessment process used in GBD to review the included studies and to exclude the poor studies. This process has three parts including general information of the study, quality of sampling, and quality of measurement. Data from population-based or community-based surveys, household expenditure surveys and censuses also will be included in the study after data cleaning for plausible ranges of variables and outliers detection. All available data will be extracted in age- and sex-stratified form for each year and province including information for mean or prevalence (depending on the analysis), sample sizes, standard deviations, standard error and/or confidence intervals, and estimates will embed the survey weights in age group-sex-province-year groups.

Since the mean of measure and its uncertainty are inputs of the model, in the analyses of risk factors, for the studies that reported mean of risk factors with only confidence intervals around the mean, we estimated standard error as one-fourth of confidence interval width. For studies reported mean, sample size and standard deviations (SD), we estimated the standard error (SE) as SD/(n^0.5).

Covariates

A covariate is a variable that has a positive or negative relationship with a disease, risk factor or injury in the NASBOD study.

We will use covariates to inform the estimation process in our models. For conditions with lots of data, covariates play a minimal role in the estimation process however for conditions with little data, the role of covariates is very important. In fact, time-varying province-level covariates can help informing the units in a setting where there are scarce data or conflicting data from multiple sources.

Some of the frequently used covariates associated with the risk factors or diseases under study are (i) urbanization, measured as proportion of province's population that lived in urban areas, (ii) province availability of multiple food types for their citizens' consumption, (iii) wealth index, estimated from assets, which were asked in yearly household expenditure surveys, (iv) years of schooling, which is educational attainment in years obtainable from household expenditure surveys, (v) population density, proportion of the province with population density over 1000 people per square kilometer, (vi) mean BMI, mean body mass index (kg/m^2) for males and females older than 20 years, , (vii) age-specific fertility rate (births per 1,000 women) obtainable from census, (viii) completeness of vital registration (% of deaths captured) obtainable from census and vital registration data, (ix) vehicles, 2+4 wheels (per capita) accessible from country Road Statistics. However, some variables like neonatal mortality rate (per 1000), diabetes prevalence (% of population), smoking prevalence (% of population), systolic blood pressure (mmHg), and indoor and outdoor air pollutions which are indeed estimates from NASBOD study will be used as covariates for estimating of other diseases, risk factors and injuries.

Crosswalking

To analyze the data collected from NASBOD, the first step is to use homogenous data. Non-homogenous data will lead to wrong estimates. Sometimes depending on the primary outcome we need to attain continuous measurement from prevalence or vice versa, for example mean FPG from diabetes prevalence since the relevant study has reported just prevalence. Other example is when one measurement can be obtained from other measurement for example FPG from HA1C. Unification of different metrics of one measurement is also the other issue which is necessary. One example of unification is to translate period prevalence of alcohol intake to point prevalence of alcohol intake.

In this case we perform a regression of alternative definition on gold standard definition (using surveys which report multiple metrics) and use the beta generated as the adjustment factor for a given definition to estimate the desired dimension. The only point is that the necessary data should be relatively high and the overlapping information from the same source is needed to generate relationships. This technic is the so-called crosswalk or metadata mapping method.

Totally, crosswalk is a method of data conversion that enables searching data across heterogeneous resources and is a useful tool for making similar data comparable.

Statistical analysis

Multilevel models that are also called hierarchical, mostly because of the parameters of the within-level regressions at the lowest, controlled by the hyper-parameters of the upper-level model, are the basis of our analysis. The multilevel modeling allows estimating heterogeneity within as well as across levels or units.[16]

The study model has six components. The first component that is the multilevel hierarchy has province-specific intercept and linear time slope. The second component of the model is the nonlinear time effect. Covariate effect is the third one. Age is the other important component which will be smoothed via a cubic spline. Since there are different kinds of data sources one component is considered for study random effect and finally variance component which is multiplicative with study random effect.

The multilevel hierarchy component of the time trends

An important trait of multilevel models is that each parameter refereeing to a specific group or unit borrows information from comparable parameters of other groups or units with similar characteristic. In other words, a shrinkage effects towards the population mean is present while using multilevel models. The volume of the shrinkage depends on the variance between the random parameters. This can be quiet beneficial, especially when a small number of individuals is observed in some groups. In such cases, there is large reduction of the uncertainty since information from other groups or units with smaller variability is incorporated in the posterior estimates.[17] This is our main rationale for using Bayesian multilevel models.

In our project, studies are nested in provinces, provinces are nested in sub-regions, sub-regions, are in turn nested in regions and all nested in the country. This is the structure of the data. The benefit of this structure as mentioned is that we can have partial-pooling estimates from the model. Partial-pooling is a compromise between two extremes; non-pooling and complete pooling. Complete pooling is when we combine all observations of a given level and non-pooling is the opposite. In this scenario, multilevel estimate of a given province is approximated by a weighted average of the observation in the province (the un-pooled estimate, $\overline{y}_j$) and the mean over all provinces (the completely pooled estimate, Y_{all}). So, depending on the availability and sparseness of data in each level, the model benefits from borrowing information by means of non-pooling, partial pooling and complete pooling.[18] This situation is repeated in each hierarchy.

Nonlinear time effect component

Nonlinear changes in time at each province will be captured using a term which is the sum of province, sub-region, region and the country and each of these four components is assigned a Gaussian autoregressive prior to allow the model to distinguish the extend of nonlinearity exist at each level.[19, 20]

In particular, the vectors of each component have a normal prior with zero mean and precision parameters. The model-estimated precision parameters will determine the degree of smoothing at each level. We will expect the provincial precision parameter to be the lowest and the country precision parameter to be the highest as the provincial trends of a disease has more extra-linear variability than the country trend. Identifiability of parameters is an issue of concern here which will be achievable by constraining some conditions. The benefit of this constraint which is compelling orthogonality between the linear and nonlinear part of the time trend is that each can be explained independently. For provinces with no data, we will take the Moore-Penrose pseudoinverse for computation because of some technical matters.[21]

Covariate effects component

The covariates which we will use in our model are categorized in two group; province-level covariates and study-level covariates. Province-level covariates include covariates like (i) wealth

index, (ii) urbanization, (iii) multiple food types based on a principal component analysis (PCA) on Household Expenditure Data, (iv) years of schooling, (v) body mass index, (vi) completeness of vital registration and etc.

The effects of some of these province-level covariates on the risk factors or diseases will be allowed to change linearly over time. Theses covariates will be smoothed using moving average methods. We may use triangular weighting to reduce the influence of yearly variation of covariates.

However, the study-level covariates include study coverage and study-level urbanization. The study-level coverage covariate which explains types of data used has four categories: (i) provincial data with sampling weight (ii) provincial data without weight (iii) district data, and (iv) individual community data.

The next study-level covariate explains whether the study has been conducted in rural, urban or rural and urban area population.

These two covariates will help us account for data sources biasness. Since non-provincial studies mostly are performed in areas of special regard or thought because of a health problem, their results will not be representative of the whole province. They might also have larger variation than provincial representative studies. As mentioned before the model considers a time-varying offset for district and community data, and additional variance components for district and community data and for provincial data without sampling weights. These variance components were estimated empirically and let provincial data with sampling weights to have a stronger effect on estimates than other sources.

The covariates and their interactions will be chosen based on substantive thoughtfulness and their predictive power through influence on the model's Deviance Information Criterion.[22] We are not seeking causal effects of these independent variables.

Age association component

Almost all risk factors and diseases have a nonlinear association with age, for example for some diseases age association might flatten or even decrease in older ages. So, we will use a cubic spline model to smooth this association.[23] We will use a baseline age and then subtract all age values from that baseline.

Since the age association between provinces might change further from differences of provincial means, province-specific random spline coefficients will be accommodated in the model with normal distribution of zero mean and σ^2 variances that each σ^2 has a flat improper prior.[24]

We treated age as a continuous variable in this model. This is the reason we extracted age groups from studies as narrow bands (5 years) to use their mid-point as continuous measurements.

Study-specific random effects component

The study-specific random effect enables age groups from the same study to have fluctuations after accounting for other variables in the model. We appoint a normal prior with variance depending on the coverage of the study i for each study-specific random effect, e_i :

$$var\,(e_i)=\begin{cases}\nu_w & if\ study\ i\ is\ weighted\ provincial\\ \nu_u & if\ study\ i\ is\ unweighted\ provincial\\ \nu_d & if\ study\ i\ is\ district\\ \nu_c & if\ study\ i\ is\ community\end{cases}$$

Weighted provincial studies may not reflect the province's mean or prevalence of the measurement under study even after accounting for sampling variability. So, the term ν_w enables us to explain this variability. ν_w can also explain study design and quality matters. We assume random effects from community studies have greater variance than random effects from district studies and so forth i.e.: $\nu_w < \nu_u < \nu_d < \nu_c$. This constrain indicates that studies with limited coverage are not only have greater or lesser than the province mean or prevalence, but also have more variability.

Residual age-study variation

Age patterns inside communities within a given province may differ and may not be consistent with its province age pattern. This kind of within-study variation will not be captured by e terms as they are the same across all observations in a given study. Thus, an additional variance component for each study, τ_i^2, will be accommodated in the model:

$$\tau_i^2=\begin{cases}\tau_w^2 & if\ study\ i\ is\ weighted\ provincial\\ \tau_u^2 & if\ study\ i\ is\ unweighted\ provincial\\ \tau_d^2 & if\ study\ i\ is\ district\\ \tau_c^2 & if\ study\ i\ is\ community\end{cases}$$

Again there is less variation in weighted provincial studies than unweighted provincial studies and so on i.e.: $\tau_w < \tau_u < \tau_d < \tau_c$. This consideration for model comprises the smooth age in residual terms not only for each province but also for each study to have its own cubic spline in age.

Computation

We will fit the model using Markov Chain Monte Carlo (MCMC) method. All statistical computation programs will be written and done in R language. As we know well, to achieve better estimations from the model we should jointly sample random effects with their hyperparameters since there is a heavy dependency between parameters.[25] We will not marginalize over mean parameters in the model since this may cause off-diagonal structure into the likelihood covariance and need manipulating large variance-covariance matrices to calculate this marginal likelihood.

A main step in running MCMC is to ensure the MCMC sampler will converge to the posterior distribution and that estimating is fast enough to return sufficient number of independent posterior draws.[24] For each model, we will start with 20 chains in parallel at randomly-selected starting values. Then, after 5000 iterations of burn-in to harmonize the Metropolis proposal variances, we will run each chain 50000 more iterations. Next, we will combine the first 20 chains and thinned them by a factor of 200 to achieve chain of length 5000 with which to generate results. A benefit of MCMC is that uncertainty generates naturally from the data via estimation in an integrated and simple manner.

Model checking

Modeling composite datasets has a risk of overfitting and underfitting and achieving tradeoff between these two needs a great attention. The perfect model is elastic enough to capture important complications while still keeping its external validity and interpretability.

We will examine our model using posterior predictive checks to verify that we have not neglected any key interaction out and also will use cross-validation to ensure we have not over-fit our data. Posterior predictive checks are well-designed and smart tool for intuitively inspecting how well a model fits our data.[26] We will compare observed datasets with a given replicated datasets, e.g. 500,

from model's posterior predictive distribution for other risk factors. Whenever the difference between this prediction and the observed data becomes smaller, this means our model is consistent with data.

For cross-validating the model, we will divide the provinces into five non-overlapping parts of equal sizes so that all five groups become similar regarding rich and sparse density. For each group of provinces we will do a 10-fold cross-validation so that we drop out 10 percent of data and fit the model on the remaining 90 percent. Then, we calculate the prediction error of the fitted model when predicting that 10 percent of data. We will do this for every 10 percent and combine the 10 percent estimates of prediction errors.[27] At each iteration of the MCMC we will draw a prediction from the main model and will build 95 % prediction intervals from predictions across all iterations.

Discussion

The NASBOD study is the first comprehensive assessment of burden of diseases, risk factors and injuries across provincial and regional levels over recent years in Iran.[15] The only study of burden of diseases and injuries in Iran dates back to 2003 which was only conducted in six provinces. The results showed a significant disparity between these provinces and also transition from communicable diseases to non-communicable diseases and road traffic injuries.[14] But, the NASBOD study is the first subnational burden of diseases study in Iran and even in the Middle-East and one of the few subnational studies all over the world.[28–31] We will obtain long-term trends of prevalence of diseases, risk factors and injuries under NASBOD study for each age group, sex, province, sub-region, region and the whole country. Then we will estimate health inequalities respectively. All the time trends will be reported together with their uncertainty intervals. We will report estimates for all province-years, subregion-years, region-years that many of them suffers from poor data.

As mentioned before provincial and subnational studies of burden of diseases inside countries provided health policymakers with a solid perspective of health situation all over the country and therefore helped them in better health management and future planning to control the progressive epidemics of all dominant diseases. The other advantage of the present study compared with the only previous one in Iran and other subnational studies in the world is that its methodological and analytical approach is very close to GBD study 2010 guidelines together with their main investigators involvement. What mentioned above are just the epidemiological achievements of such a study which will be a helpful landmark for policy makers in health systems.

The NASBOD project achievements are not only very important from epidemiological perspective but also from statistical point of view because of handling the complexities existing in the nature of this study. These complexities will be modeled with new advanced statistical models especially Bayesian autoregressive multilevel models as explained in this paper.

Though the detailed main discussion on the results will be provided after running the model and releasing the results, we can talk about strengths and benefits of the model under the study now.

Multilevel models are of the rare approaches for modeling aggregated data like what we encountered in NASBOD study.

One of the main advantages of multilevel models is assessing different levels effects. Considering higher level units as a random sample (parts of a distribution) allows quantification of higher level variation in the total population and therefore leads to unbiased standard error estimates and independent residuals of the model.[16] Another advantage is that missing data which are frequently occurred in large surveys are handled very simply via these models.[16]

Though handling missing data is one of the advantages of multilevel models and we will just use these models together with informative priors to impute missing information, our model suffers from data scarcity especially in older age groups and the earlier time of the study. Our model also suffers from low quality and non-representativeness of data at the earlier time of the study mainly before 2000. Thus, relatively large amount of data will make our inferences more robust.

It is clear from the literature that many approaches have been developed for missing data imputation but almost all of them use simple methodology like bootstrapping just like Amelia[32] and do not have a strong assumptions or justifications to be used in complex situations resembling NASBOD study. The only disadvantage is that modeling process and interpretation of the results may be complex which both can be passed off through advanced computers, efficient statistical programming and carefulness. Data gaps may be the main limitation of our study just like what occurred in modeling GBD study 2010.[10–13] The other limitation is the geographical incompatibility that occurs at the provincial levels which is not a serious problem in multilevel modeling since we have only slight changes during the study period and it can be handled with tricky techniques. But it may be a serious problem at district levels and more advanced models should be developed at this phase of study in near future.

The other sensible models which can be engaged in NASBOD study is the Spatio-temporal models.[33] which will be developed and fitted in parallel and then will be mixed with the present Bayesian autoregressive multilevel models to develop ensemble models which will produce independent model and more reliable and accurate estimations. Ensemble models are weighted combinations of the posterior distributions of individual models and provide lower error for point estimates and more accurate uncertainty intervals.[34–36] Moreover, ensemble models catch uncertainty due to both the parameters in any single model and the uncertainty of predictions owing to differences in specification across models.

Generally speaking, the main advantages of the mentioned model is estimating long-term trends using a Bayesian autoregressive multilevel model to predict mean and prevalence of risk factors and diseases by including non-linear age associations and time trends, incorporating study coverage as well as variance component in the model. Coverage-specific offsets and variances enable our model to use all the data and track provincial representative studies more than other data sources. We expect coverage-specific variance components to be greater and have larger uncertainty for data sources with less representativeness and ultimately uncertainly intervals achieved from the Bayesian model that represent the true availability of information.

Though we are to develop a sophisticated model based on real needs and existing complexities in real situations to estimate missing information this does not obviate the need for gathering more and more qualified data.

All mentioned about modeling and its challenges in complex conditions itself creates careers for young researchers to learn and train more and more and this capacity building ultimately will lead to knowledge production in the country.

As a bottom line, achieving estimations of time trends after mod-

eling all diseases, risk factors and injuries under NASBOD study can help anybody who works in health systems, specially Health Policy Makers and also politicians to trace, understand and monitor epidemiological transition of non-communicable diseases in all over the country and then launch prevention plan to reduce the burden of non-communicable diseases, risk factors and injuries and consequently achieve the new health goal of the World Health Assembly in 2012,[2] which is reducing avoidable mortality from non-communicable disease (NCDs) by 25 % by 2025 (the 25 by 25 goal).

Authors' Contributions

General design prepared by Farshad Farzadfar and Amir Kasaeian. Designing of models prepared by Farshad Farzadfar, Kazem Mohammad, Amir Kasaeian, Mohammad Reza Eshraghian and Abbas Rahimi Foroushani. The primary draft was prepared by Amir Kasaeian and revised by all co-authors. All authors have given approval to the final version of the manuscript.

Acknowledgments

The study is granted by Ministry of Health and Medical Education of Islamic Republic of Iran and Setad-e-Ejraie Farmane Imam. The authors would like to express thanks to Dr.Masoud Moradi for his precise editing of the text and Ms Rosa Hagh Shenas for her efforts on managing coordinative and administrative processes.

References

1. Chan M. From new estimates to better data. *The Lancet*. 2012; **380(9859):** 2054.
2. Horton R. Non-communicable diseases: 2015 to 2025. *The Lancet*. 2013; **381(9866):** 509 – 510.
3. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, et al. 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*. 2013; **380(9859):** 2224 – 2260.
4. Murray CJ, Vos T, Lozano R, Naghavi M, Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2197 – 2223.
5. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2163 – 2196.
6. Salomon JA, Wang H, Freeman MK, Vos T, Flaxman AD, Lopez AD, et al. Healthy life expectancy for 187 countries, 1990–2010: a systematic analysis for the Global Burden Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2144 – 2162.
7. Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2129 – 2143.
8. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2095 – 2128.
9. Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, et al. Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2013; **380(9859):** 2071 – 2094.
10. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2· 7 million participants. *The Lancet*. 2011; **378(9785):** 31 – 40.
11. Danaei G, Finucane MM, Lin JK, Singh GM, Paciorek CJ, Cowan MJ, et al. National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5· 4 million participants. *The Lancet*. 2011; **377(9765):** 568 – 577.
12. Finucane MM, Stevens GA, Cowan MJ, Danaei G, Lin JK, Paciorek CJ, et al. National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9· 1 million participants. *The Lancet*. 2011; **377(9765):** 557 – 567.
13. Farzadfar F, Finucane MM, Danaei G, Pelizzari PM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in serum total cholesterol since 1980: systematic analysis of health examination surveys and epidemiological studies with 321 country-years and 3· 0 million participants. *The Lancet*. 2011; **377(9765):** 578 – 586.
14. Naghavi M, Abolhassani F, Pourmalek F, Lakeh MM, Jafari N, Vaseghi S, et al. The burden of disease and injury in Iran 2003. Population health metrics. 2009; **7(1):** 9.
15. Farzadfar F, Delavari A, Malekzadeh R, Mesdaghinia A, Jamshidi HR, Sayyari A, et al. NASBOD 2013: Design, definitions, and metrics. *Arch Iran Med*. 2014; **17(1):** 7 – 15.
16. Goldstein H. Multilevel statistical models: Wiley. com; 2011.
17. Ntzoufras I. Bayesian modeling using WinBUGS: Wiley. com; 2011.
18. Gelman A. Data analysis using regression and multilevel/hierarchical models: Cambridge University Press; 2007.
19. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; **88(421):** 9 – 25.
20. Rue H, Held L. Gaussian Markov random fields: theory and applications: CRC Press; 2005.
21. Harville DA. Matrix algebra from a statistician's perspective: Springer; 2008.
22. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; **64(4):** 583 – 639.
23. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine*. 1989; **8(5):** 551 – 561.
24. Marin J-M, Robert CP. Bayesian core: a practical approach to computational Bayesian statistics: Springer; 2007.
25. Chib S, Carlin BP. On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*. 1999; **9(1):** 17 – 26.
26. Gelman A, Meng X-L, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*. 1996; **6(4):** 733 – 760.
27. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*. 2005; **27(2):** 83 – 85.
28. Bradshaw D, Nannan N, Groenewald P, Joubert J, Laubscher R, Nijilana B, et al. Provincial mortality in South Africa, 2000-priority-setting for now and benchmark for the future. *South African Medical Journal*. 2008; **95(7):** 496.
29. Stevens G, Dias RH, Thomas KJ, Rivera JA, Carvalho N, Barquera S, et al. Characterizing the epidemiological transition in Mexico: national and subnational burden of diseases, injuries, and risk factors. *PLoS Medicine*. 2008; **5(6):** e125.
30. Begg SJ, Vos T, Barker B, Stanley L, Lopez AD. Burden of disease and injury in Australia in the new millennium: measuring health loss from diseases, injuries and risk factors. *Medical journal of Australia*. 2008; **188(1):** 36.
31. Asaria P, Fortunato L, Fecht D, Tzoulaki I, Abellan JJ, Hambly P, et al. Trends and inequalities in cardiovascular disease mortality across 7932 English electoral wards, 1982–2006: Bayesian spatial analysis. *International journal of Epidemiology*. 2012; **41(6):** 1737 – 1749.
32. Honaker J, King G, Blackwell M. Amelia II: A program for missing data. R Package version 1.5–5. 2011.
33. Parsaeian M, Farzadfar F, Zeraati H, Mahmoudi M, Rahimighazikalayeh G, Navidi I, et al. Application of spatio-temporal model to estimate burden of diseases, injuries and risk factors in Iran 1990 – 2013. *Arch Iran Med*. 2014; **17(1):** 28 – 32.
34. Vrugt JA, Robinson BA. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research*. 2007; **43(1):** W01411.
35. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*. 2007; **102(477):** 359 – 378.
36. Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*. 2005; **133(5):** 1155 – 1174.