Abstract

Background: Identifying the burden of disease and its inequality between geographical regions is an important issue to study health priorities. Estimating burden of diseases using statistical models is inevitable especially in the context of rare data availability. To this purpose, the spatio-temporal model can provide a statistically sound approach for explaining the response variable observed over a region and various times. However, there are some methodological challenges in analysis of these complex data. Our primary objective is to provide some remedies to overcome these challenges.

Method: Data from nationally representative surveys and systematic reviews have been gathered across contiguous areal units over a period of more than 20 years (1990 – 2013). Generally, observations of areal units are spatially and temporally correlated in such a way that observations closer in space and time tend to be more correlated than observations farther away. It is critical to determine the correlation structure in space-time process which has been observed over a set of irregular regions. Moreover, these data sets are subject to high percentage of missing, including misaligned areal units, areas with small sample size, and may have nonlinear trends over space and time. Furthermore, the Gaussian assumption might be overly restrictive to represent the data. In this setting, the traditional statistical techniques are not appropriate and more flexible and comprehensive methodology is required. Particularly, we focus on approaches that allow extending spatio-temporal models proposed previously in the literature.

Since statistical models include both continuous and categorical outcomes, we assume a latent variable framework for describing the underlying structure in mixed outcomes and use a conditionally autoregressive (CAR) prior for the random effects. In addition, we will employ misalignment modeling to combine incompatible areal units between data sources and/or over the years to obtain a unified clear picture of population health status over this period. In order to take parameter uncertainties into account, we pursue a Bayesian sampling-based inference. Hence, a hierarchical Bayes approach is constructed to model the data. The hierarchical structure enables us to “borrow information” from neighboring areal units to improve estimates for areas with missing values and small number of observations. For their general applicability and ease of implementation, the MCMC methods are the most adapted tool to perform Bayesian inference.

Conclusion: This study aims to combine different available data sources and produce precise and reliable evidences for Iranian burden of diseases and risk factors and their disparities among geographical regions over time. Providing appropriate statistical methods and models for analyzing the data is undoubtedly crucial to circumvent the problems and obtain satisfactory estimates of model parameters and reach accurate assessment.

Keywords: Burden of diseases, Iran, misalignment, spatio-temporal models, study profile


Introduction

Evaluating the burden of diseases and related risk factors is essential to identify key health priorities. Some studies have investigated the global and national pattern of disease.
addition to spatial correlation, the data have been collected over time. Consequently, observations closer in time tend to be more correlated than observations farther away. The mentioned advantages of spatio-temporal correlation between areal units make it possible to impute the estimates that are rare but needed in the burden of diseases studies.

The present article aims to discuss the statistical challenges in analysis of disease burden and propose novel models in spatio-temporal framework and appropriate approaches to handle these challenges.

**Method**

**Data Sources**

Data from nationally representative surveys and systematic reviews have been collected over contiguous areal units (provinces, districts and census tracts) through a period of more than 20 years (1990 – 2013). These nationally representative surveys include Non-Communicable Disease Surveillance Surveys (NCDSS), National Health Surveys (NHS), Demographic Health Survey (DHS), Census information and Household Expenditure Surveys, Hospital Data Survey,11 Outpatient Data and other national health surveys. Table 1 presents an overview of available data sources, their areal unit, and measurement time. As shown, the areal units and the time spans are different between surveys and it is not an easy task to combine these data to produce a unified result.

**Statistical challenges and possible solutions**

In the following section, we will discuss practical challenges that we will encounter in the data analysis step and we will briefly explain how statistical methods can be extended to overcome these limitations.

**Modeling Spatio-temporal correlation structure**

Spatial correlation of response variables across studied region violates assumption of independence in ordinary regressions. One way to overcome these drawbacks is the extension of ordinary regression to include spatial random effects. In this way, model can easily capture any over-dispersion or spatial autocorrelation that remains after accounting for available covariates. A well-known framework used in this context is Bayesian hierarchical model with conditional autoregressive (CAR) prior for these random effects. In this setting, it is essential to define an adjacency matrix for areal units that characterize the neighboring structure and importance weight of each neighbor. Typical approaches for defining these weights are based on the distance between two area centroids and sharing a common boundary. Accordingly, model estimate at a given location is associated with nearby estimates, so it reflects some kind of autocorrelation.12 This modeling framework is flexible and can be implemented via Markov chain Monte Carlo (MCMC) simulations. For their general applicability and ease of implementation, the MCMC methods are the most adapted tool to perform Bayesian inference. However, in our study these methods face several limitations. When model involves a large number of highly correlated latent variables, the conventional MCMC algorithm may converge slowly or even fail to converge. To deal with this obstacle and alleviate the convergence problem through MCMC methods, we will use and assess some advanced techniques which have been suggested in recent years (e.g. Inverse Bayes Formula13–16 and Integrated Nested La-place Approximation17 as alternatives to MCMC). Using these techniques, the statistical inference and model fitting can be carried out more efficiently.

Another important feature of this study is to obtain trends of outcomes during this period. Since observations of areal units over time are correlated, special strategy for modeling trends is required. Note that observations are simultaneously correlated across space and time, so statistical methodology for space or time separately is not reasonable. Hence, we will use spatio-temporal models which can simultaneously consider spatially correlated residual and time trends. Note this model can consider both linear and non-linear trends that may happen in real phenomena. Therefore, it can overcome the limitations of the traditional regression models.

**Misalignment**

Linking several data sets from different sources such as national health surveys, censuses or systematic reviews in which data were recorded based on different areal units require special methodology.18 For example, consider we want to model fasting plasma glucose that has been measured by NCDSS and NHS. These measurements are recorded based on different areal units. In fact, while NCDSS areal unit is recorded at districts level, NHS is recorded at province level. A problem arises here is how to combine data from two different levels and especially how to define adjacency structure of areal units that are not compatible for these two data sets. Two remedies are possible: We can omit the district level data and use only province level data for both these data sources. This ad-hoc method does not provide satisfactory estimates of the model parameters. In fact, the two sources of information are valuable in different ways. A more sophisticated solution is to design a modeling framework that can consider different level of data aggregation. This strategy, which has been popular in recent years, is based on utilizing misalignment techniques to combine data from different sources.

In addition to spatial misalignment, the number of cities and provinces has changed during this period. There were a total of 24 provinces in 1990 which has increased to 31 provinces in 2013. It should be noted, the province boundaries has changed after division. So, administrative divisions produce incompatible areal units over the course of study. This problem become more serious at districts level since the number of districts is nearly doubled in this period.

Thus, incompatible areal units between data sources and/or over the years necessitate modeling spatio-temporally misaligned data.19,20

**Dealing with data scarcity**

Data scarcity is a major shortcoming of this study and it should be considered thoroughly. The target responses are not measured for all districts and/or in all years. For example, NCDSS has been measured only in 6 years and 49 districts have been studied in 2011. Indeed, available data points are far fewer than what is necessary to produce reliable subnational inference during this period.

The most common remedy for this problem is to consider unmeasured responses as missing data. Latent class model provides a flexible framework for dealing with missing data. In this framework missing data along with model parameters will be considered unknown and will be estimated via the proposed Bayesian
Table 1. Available data sources and their time span

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Areal unit</th>
<th>Time span</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHS</td>
<td>Province</td>
<td>1991, 2000</td>
</tr>
<tr>
<td>DHS</td>
<td>Province</td>
<td>2000</td>
</tr>
<tr>
<td>Household Expenditure</td>
<td>District, Province</td>
<td>1984 – 2011</td>
</tr>
<tr>
<td>Hospital Data Survey</td>
<td>Hospital</td>
<td>1996 – 2013</td>
</tr>
<tr>
<td>Outpatient Data</td>
<td>Medical Science University</td>
<td>1998 – 2012</td>
</tr>
<tr>
<td>Systematic Review</td>
<td>District, Province</td>
<td>1990 – 2013</td>
</tr>
</tbody>
</table>

Good methods for the analysis of categorical and continuous outcomes facilitate infer-
ences and prediction, this assumption might be overly restrictive to repre-
sent the data. In fact, the violation of those assumptions results in degradation of model performance.

Therefore, new statistical model and methodology are required for this kind of responses. All required programs are written through R and WinBUGS statistical packages, which are two commonly used programming languages especially in the Bayesian context.

Model selection criteria
There are a wide variety of models that can be used to estimate trends of diseases and risk factors. These models may be different in terms of covariates, statistical methodology or even the form of response variable. So model selection criteria are required to choose the best model among all possible ones. Usually we want to generalize the results of the models, and estimate how accurately a predictive model performs in practice. So the best model is selected based on better out of sample performance. Hence the data is divided into two parts, performing the analysis on one part and validating the results on the other part.

Three types of criteria were used to select the best model. They include absolute or relative error in prediction, error in estimating accurate trends and coverage of uncertainty intervals. One model may perform better in one criterion and worse in the other. So it is sensible to combine the results of single models that are best based on these different criteria. Indeed the final ensemble model is a weighted average of these single models. There are many methods including Bayesian Model Averaging which can be used to develop weights for these single models.

Missing data imputation for older age group
Another challenge is how to deal with important groups that had not been measured in none of the surveys. For example, metabolic risk factors had not been measured for age groups above 65 years in NCDSS surveys. Hence, we don’t have any information about these groups and need to extrapolate. The statistical strategy for this issue is somewhat different from modeling missing data in which interpolating rather than extrapolating is required.

To overcome this problem for example in metabolic risk factors, we will use 47 surveys from other countries that have measured metabolic risk factors for all age ranges. Detailed description of this methodology is explained elsewhere. However, in Brief: I. First, we select individuals aged 30 – 60 years in both NCDSS and 47 international surveys and estimate the linear slope of metabolic risk factors versus age separately by provinces and surveys, respectively. Then we divide both set of the slopes into three groups based on the tertile values of international slopes.

II. For each tertile, we will estimate overall and province specific intercept via a fixed effects model that contains the effects of 5-year age groups and provinces on NCDSS data set for men and women separately.

III. For each tertile, we fit the same fixed effect model in 47 international surveys. Note that this time data includes older age groups. After removing survey specific effect, beta coefficient of older age groups can be obtained.

IV. For older age groups, the province specific mean and variance estimates of each metabolic risk factor can be obtained through adding the overall intercept, province specific intercept of NCDSS and older age group coefficient from relevant tertile of international surveys.

Conclusion
In the present study, we assess a wide variety of diseases, injuries and risk factors along with powerful statistical methods to characterize the health status of Iranian people and related changes...
during a period of over 20 years. To our knowledge, there is only one Iranian national burden of disease study that was conducted in 2003 by the Ministry of Health and Medical Education. This study was conducted at the national level and included six selected provinces. Hence, the results just render an overall picture of the society and more detailed studies are required to provide comprehensive evidences for the whole country.

Some studies have investigated the national and subnational pattern of disease worldwide. In a study conducted in Mexico, burden of disease was investigated at the national and subnational levels, and subnational disparities between states were identified to set health priorities.

Recently conducted burden of disease studies, such as Global Burden of Disease 2010, pay more attention to quantifying the variations among areal units and they explicitly incorporates these variations in the modeling framework. In the present study, we have developed a Bayesian hierarchical model with spatial and temporal correlation. This modeling strategy is comparable to Global Burden of Disease 2010 and provides a flexible framework that enables us to “borrow information” from neighboring areal units and nearby time periods to obtain more accurate and efficient estimates. Using this additional information for imputing missing values, this model outperforms other alternative imputation methods such as Amelia II package. Also additional prior information can be incorporated in the model and it is also possible to estimate via small number of observations.

Since in the Burden of disease study we primarily use existing data sets rather than gathering new information, the problem of different areal units is prevalent. Also administrative divisions result in increased number of provinces and districts, creating many incompatible areal units especially at district level. This problem is less pronounced in Global Burden of Disease since the areal units such as countries, almost remains constant. We have used spatio-temporal misalignment model to combine the results from disparate data sources and to produce a unified clear picture at province or district level as well as national level.

Two commonly used models that can be employed in this setting are multilevel and spatial models. Both models are well functioning to quantify this heterogeneity, however extra information about distances between observations will be used in the latter. In this study we mainly discussed spatial models, but both spatial and multilevel models will be used in the analysis of burden of disease to compare their predictive ability and reduce any model dependency in the final results. In addition Ensemble model which is the weighted average of the best spatial and multilevel models will be used as a final model. It has been shown that ensemble models outperform any single model and are less sensitive to model specification bias and produce smaller prediction error. These models incorporate both uncertainty for any single model and uncertainty due to different model specifications.

As mentioned, data scarcity is one of the major shortcomings of this study. Available data points are far fewer than what is necessary to produce reliable inference. Therefore, an attempt is made to use all available information and a sophisticated modeling strategy to compensate for the lack of sufficient and accurate databases. In the presence of more reliable data for each province-year, less complicated models are required. Nevertheless, as expected modeling just can solve part of this problem and estimates with wide uncertainty intervals is obtained for some areas with small number of observations. In this study we have tried to provide an accurate picture of health status in Iran based on available data sources and the results can inform policy makers about the current and future health status of society and reveal possible gaps across different geographic regions.

Author’s Contribution

General designing of paper: Farshad Farzadfar, Majid Jafari Khaledi, Kazem Mohammad, Hojjat Zeraati, Mahmood Mahmoudi, Mahboubeh Parsaein

Designing of models: Majid Jafari Khaledi, Farshad Farzadfar, Kazem Mohammad, Hojjat Zeraati, Mahmood Mahmoudi, Mahboubeh Parsaein

Writing primary draft: Mahboubeh Parsaeian, Farshad Farzadfar, Kazem Mohammad, Majid Jafari Khaledi, Gelareh Rahimizahalayeh

Manuscript revision: Farshad Farzadfar, Majid Jafari Khaledi, Kazem Mohammad, Mahboubeh Parsaeian, Hojjat Zeraati, Mahmood Mahmoudi, Iman Navdidi

Rule of funding Sources

Our research was funded by Iran’s Ministry of Health and Medical Education of Islamic Republic of Iran and Setad-e-Efzarane Farmane Imam. The spatio-temporal analysis of these date sets is the PhD thesis of Mahboubeh Parsaeian in Biostatistic.

Acknowledgments

The authors thank our colleagues in Non Communicable Disease Research Center and appreciate the collaboration of organizations and researchers that provided necessary data for this project. We would also like to express thanks to DeMasoud Moradi for his precise editing of the text and Dr. Sirin Djalalinia and Ms Rosa Haghigh Shenas for thier efforts on managing coordinative and administrative processes.

References


