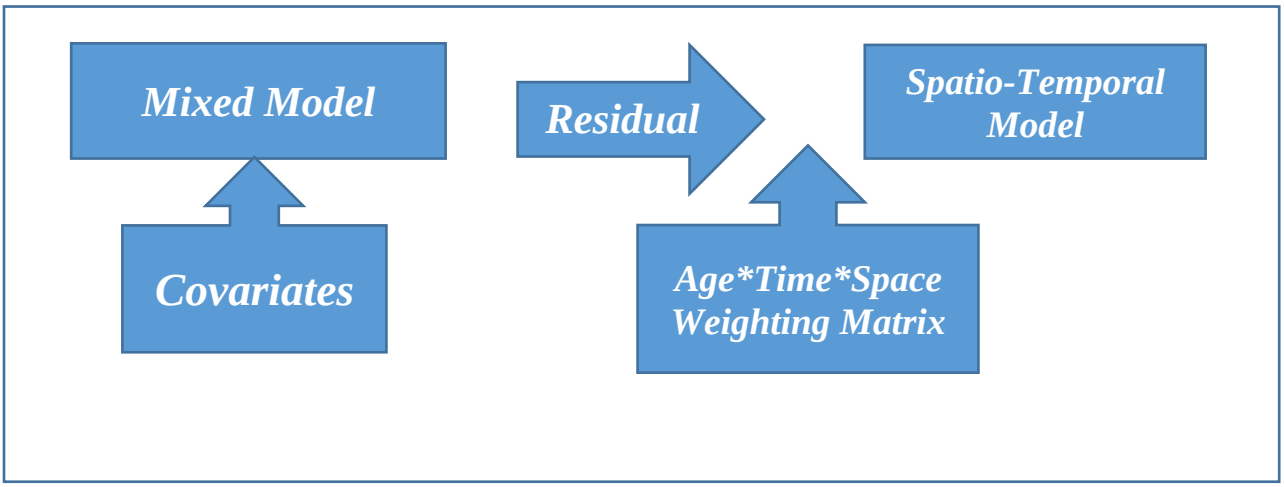
Supplementary file 3. Statistical Analysis Detail

Spatio-temporal models are used when the study data is collected geographically and also over time. Both spatial and temporal dimensions must be considered simultaneously when analyzing these data. In other words, considering only one temporal or spatial dimension of the correlation does not take into account the other dimension, thereby reducing the study power. The model used in this study should be capable of taking into account spatial and temporal correlations simultaneously. It is also important to consider the differences between age and sex in the estimates of this model. This model consists of two components. First, a random effects model was fitted to the data. The fitted model was defined as follows:

$$\mathrm{Log}it(prevalence\ rate) = \alpha + \beta_1 sex + \beta_2 age + \beta_3 year + \beta_4 YOS + \beta_5 WI + \beta_6 urbanizratio + b_i$$

The variables used in this model include mean mean years of education (YOS) and household assets index (WI), and urbanization as a covariate and population logarithm as a compensation variable to improve prediction values. Also the random effect b_i is related to each province. Second, the predicted response values and residuals were extracted from this model. The residuals of this step were used in the spatio-temporal model to investigate the effects of variables such as age, sex, time, and location that the random effects model could not justify.

The procedure used for the spatio-temporal model is that any data in the dataset can affect other data. The magnitude of this effect varies depending on the spatio-temporal as well as the proximity of the two data sets. The effects are calculated and weighted for each of the temporal, spatial, and age correlations by their proximity matrices. Finally, by combining these three matrices, the overall weighting matrix is obtained. By applying this matrix to the data, we obtain the expected values of the residuals for the predicted values. Finally, the sum of the predicted residuals of this step and the predicted values of the regression model of the random effects of the final estimated values are obtained.



A simple exponential function was used to form the age variable weight matrix as follows:

$$W_{a_{i,j}} = \frac{1}{e^{\omega * (agegroup_i - agegroup_j)}}$$

As it is evident, with increasing distance between the two age groups, the assigned weight decreases. Also if two observations are in the same age group have the highest weight possible. The parameter ω plays a decisive role in model estimations. For cases with thin data, decreasing the ω parameter can increase the level of smoothing on the age variable. On the other hand, when the difference between age groups is high, the smoothing rate of the model on the age variable can be reduced by increasing the ω value.
To construct the time-weight matrix, a matrix scheme similar to that used in the local Loess regression was used. This scheme is defined as follows:

$$W_{t_{i,j}} = (1 - (\frac{|year_i - year_j|}{argmax(|year_i - year_j|) + 1})^{\lambda})^3$$

As we can see here, by increasing the interval of the two observations, the weight of the observations decreases and vice versa. The main difference between the above scheme and that used in the Loess model is the existence of a parameter λ that can control the smoothness of the results on the time variable. This value is determined by the number and distribution of data, as stated for the age variable.

the matrix with zero and one elements was used to model spatial correlation in the model. A matrix with 31 columns and rows was formed. Each row and column of this matrix represented one of the provinces of the country. If the two provinces had a common boundary, for the same row and column as their common border, the number was one, and if they had no neighborhood, the value was zero.
Finally, all three weight matrices of age, time and place were multiplied to obtain the final weight matrix. Finally, after applying this weight matrix to the model residuals, the residuals adjusted to the predicted amount of the model were combined and the final results were obtained.
Different methods were used to obtain the uncertainty rates at different stages. The first stage variance was closed form. But in the second step, due to the complexity of the calculations and the fact that it was virtually impossible to calculate the variance of estimates by mathematical methods, the simulation was used. The method used was to estimate the values of the random effects model with their variance. Then the random values of the normal distribution were estimated with the mean values. This process was performed 1000 times. Thus, based on the percentiles of the values obtained, the point and distance estimates of the required values were obtained.

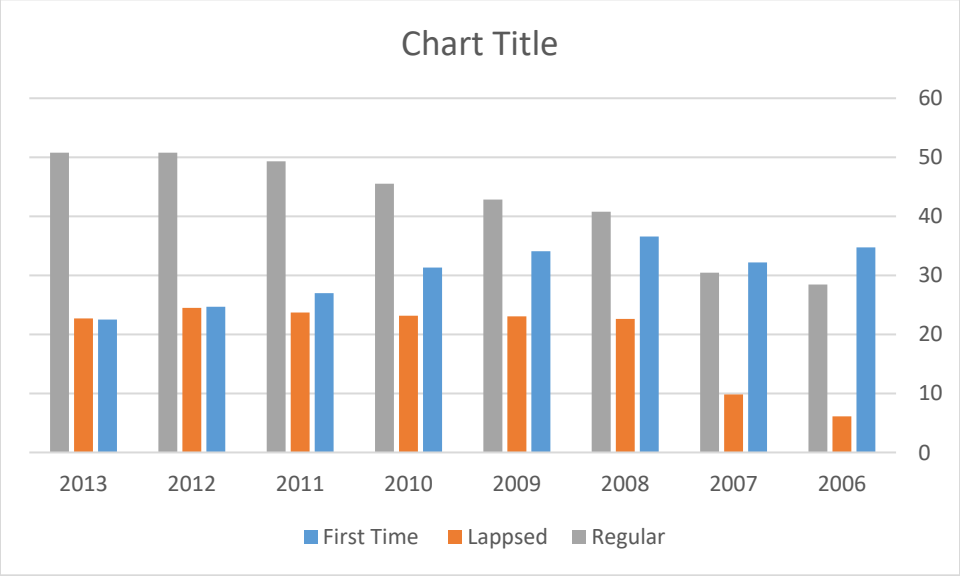
Cross Walk :

A crosswalk is a specification for mapping one metadata standard to another. Crosswalks provide the ability to make the contents of elements defined in one metadata standard available to communities using related metadata standards. We have used linear regression model using logit transformation and covariates (urbanization ,education, and welth index) to enhanced mapping . The RMSE has a "unite" value and the RMSE is not

directly comparable for different "unite" values. However, RMSE values can be predicted for detecting model performance in a calibration period using a validation period as well as comparing individual model performance with other models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}}$$

To do cross-walks, both sets of data must contain at least some combination of information. Since the blood transfusion data had nothing to do with systematic reviews data, we had to first predict all possible combinations using the spatio-temporal analysis model in the blood transfusion data, and then use this information and cross-sharing points. The blood transfusion data contained three different types of donors: first-time, with pervious history, and regular. A first-time donor is a donor who came to donate blood for first time . So their risk of infection is high despite initial screening of donors by a physician, and fortunately they are less than half of blood donors. regular blood donors are persons who donate blood at least 3 to 4 times a year. with history blood donors are people who donate blood at least once a year, and these individuals have a lower risk of infection than the first time blood donores. In this study, a systematic search was done on the numbers and ratio of different type of donation in each year from the literature review and used as weight cores to aggregate the Blood Transfusion data to initiate cross walk . this gave us a better estimate and our results was similar to pupulation base studies.



Finally, using the regression model, the correlation between the systematic review data of hepatitis B and The hepatitis B prevalence data in blood donors was calculated by age, sex , and province using cross walk method. the prevalence of hepatitis B was estimated in areas without data. The challenge with blood transfusion data is that we cannot consider the donor population to be the same as the normal population, and we used the

cross walk method to generalize it to the normal population. Four regression models were performed with log, logit, squared root, norm prevalence, and the best fit model was the logit model used.
The covariate data were obtained from the Iranian Center for Statistics, Demography and Health Survey (DHS) which included demographic characteristics such as age group, gender, school year, urbanization, wealth index.
To determine the degree of uncertainty, the spatio- temporal model error values were used. RMSE was also used to estimate the model validity.
For age standardize , the national population data of Iran in 2016 was used.