The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns

KH Jacobsen¹,²* and JS Koopman¹

Accepted 1 March 2005

Background Hepatitis A virus (HAV) infection confers long-term immunity, so mathematical analysis of age-specific seroprevalence in populations can reveal changes in the infection rate over time. HAV transmission is related to access to clean drinking water, personal hygiene and public sanitation.

Methods We used an SIR (susceptible-infectious-recovered) compartmental model with age structure to fit a time-dependent logistic function for HAV force of infection for 157 published age-seroprevalence data sets. We then fit linear regression models for socioeconomic variables and infection rate.

Results The proportion of the population with access to clean drinking water, the value of the human development index (HDI), and per capita gross domestic product (GDP) are all inverse predictors of HAV infection rates. Declining infection rates were observed in 65.6% of the surveys.

Discussion This work demonstrates the utility of HAV seroprevalence studies to reveal patterns of change in force of infection and to assess the association between socioeconomic risk factors and transmission rates.

Keywords Hepatitis A virus, seroprevalence studies, socioeconomic development, water

Hepatitis A virus (HAV) infection is a common infection, responsible for about 1.4 million new infections worldwide each year.¹ Infection is generally acquired by the faecal-oral route either through person-to-person contact or ingestion of contaminated food or water.² Low income, low educational level, crowding and lack of access to safe drinking water and sanitation facilities are associated with increased HAV infection.³ As socioeconomic status (SES) and access to safe drinking water are increasing, the HAV infection rate is declining in most parts of the world.³ However, because HAV infection in children is often asymptomatic but most infected adults present with jaundice and other potentially severe symptoms, this decrease in the infection rate has a paradoxical effect. As socioeconomic conditions improve, individuals become infected at a later age when disease is more severe. Thus, hepatitis A morbidity may increase as the incidence rate of infection decreases.

Population age-immunity structure is important in predicting HAV transmission patterns. In areas with very high endemicity, ≥30–40% of children acquire infection before 5 years of age and almost all persons have been infected by early adulthood.¹ In these regions there is very little hepatitis A disease. In areas with low endemicity, few children have anti-HAV antibodies, and many adults remain susceptible. Epidemics are uncommon in highly endemic areas because most adults have acquired immunity but in regions with lower levels of adult immunity HAV infections may occur primarily as outbreaks.⁴

Because infection with HAV generally confers lifelong immunity to all strains of HAV, age-specific seroprevalence rates are indicators of the level of susceptibility to severe disease in a population. Furthermore, the shapes of population age-seroprevalence curves reflect the changes in force of infection that have occurred over the lifetimes of the people in the populations (Figure 1). For example, in a population that had a high infection rate until 40 years ago and then experienced a rapid increase in SES and access to improved water sources, the population seroprevalence curve would show that nearly all adults >40 years have antibodies while few persons <40 years of age have immunity. A population with a more gradual decline in the infection rate over time would have a rounder age-seroprevalence curve. A population with a consistently high force of infection would have a steep rise in seroprevalence counts.
in young children. Thus, it is possible to use data about age-seroprevalence distributions in real populations to determine both when changes in the infection rate occurred and the rate of change over time.

In this paper we use a deterministic model to calculate time-dependent infection rates for 157 sets of age-seroprevalence data from surveys conducted across the globe. We assess regional differences in infection rates and the trend of declining infection rates seen in most regions of the world. We also use linear regression to examine measures of socioeconomic development which might be related to changes in the infection rate over time.

Methods
Mathematical model
We formulated a mathematical model that represents the transmission dynamics for populations with various age and immunity structures. An SIR (susceptible-infectious-recovered) compartmental model with 80 age groups (one-year age compartments for ages 1–80) was used. We assumed that infants entered the population at 1 year of age, when maternal immunity had waned, and all adults were removed from the population before age 81. The distribution of the population by age was estimated using international data from the US Census Bureau, and age-specific death rates were calculated to maintain population size and age-structure.

A flow diagram of the model, which consists of ordinary differential equations, and definitions of model parameters are shown in Figure 2. The equations used are as follows:

$$\frac{dS_i}{dt} = aS_{i-1} - \lambda S_i - (\mu_i + a)S_i$$
$$\frac{dI_i}{dt} = aS_{i-1} + \lambda S_i - (\mu_i + a)I_i$$
$$\frac{dR_i}{dt} = aR_{i-1} + \rho I_i - (\mu_i + a)R_i$$

where $S_i$ is the proportion of susceptible persons, $I_i$ is the proportion of infectious persons, and $R_i$ is the proportion of immune (recovered) persons in the total population. The subscript $i$ indicates age-specific values, $a$ represents aging, $\mu_i$ is the age-specific death rate, $\rho$ is the recovery rate, and $\lambda$ is the infection rate. For the first age group a birth function, rather than an aging function, was used to add susceptible one-year-olds to the population.

This model structure requires a number of key assumptions. (i) All persons in the population are susceptible to infection until they become infected. (ii) Infection produces lifelong immunity, which can be detected by serological examination. (iii) Mortality due to infection is negligible. (iv) Populations are homogeneous with respect to susceptibility and exposure. (v) The samples consist of individuals who have spent their entire lives in the same population. This assumes, for instance, that members of low-endemicity populations have not travelled to endemic areas. (vi) The immunization rate is very low such that we can assume that immunity is the result of infection and not immunization.

We used a three-parameter symmetric logistic function to describe the infection rate $\lambda$ over time:

$$\lambda(t) = \frac{\lambda_\infty}{1 + \theta \alpha^{-t}}$$

where $\lambda_\infty$ is the constant representing the asymptote of the past force of infection, $\theta$ and $\alpha$ are constants, $0 < \theta \leq 1$, and $0 < \alpha < \infty$. We set $t = 0$ as the time 100 years before the collection of data. We assumed that the initial population had very high seroprevalence, with all children acquiring immunity in early childhood. The conclusions of the model were not sensitive to the assumption of very high endemicity at the start of the run because the model was run long enough to replace the entire population.

The value of the infection rate, $\lambda$, in a given time period (set as one month) determined the proportion of susceptible persons...
in the model population who acquired new infections during that time period. The rate and magnitude of change in $\lambda$ during the 100 years the model was run defined the shape of the age-seroprevalence curves. The infection rate over time, $\lambda(t)$, was described by $\lambda_\infty$, $\theta$ and $\alpha$.

For each of the 157 data sets we analysed, we used Berkeley Madonna software to fit the values of $\lambda_\infty$, $\theta$ and $\alpha$ that generated age-seroprevalence curves that most closely matched the actual data. These values were fit by writing a least squares function for each dataset. The equation of the optimizing function was:

$$\sum_{j=1}^{k} m_j (\hat{p}_j - p_j)^2$$

where $\hat{p}$ was the reported fraction of a particular age group with anti-HAV antibodies from the literature, $p$ was the seroprevalence for that age group generated by the computer, and $m$ was the age range in each of the $k$ number of age groups $j$. The optimizing function minimized the average difference between observed seroprevalence rates and the corresponding seroprevalence rates generated by the computer for each age or age group.

Each model was run for 100 years, ending at the time of data collection, and the calculated infection rate at the time of data collection was recorded. Age-group seroprevalence rates for model populations, such as the proportion of adults aged $\geq 30$ with immunity, were calculated from the one-year age compartments. We also calculated the derivative of the equation for the infection rate so that we could find both the slope of the change in $\lambda$ over time and the time when the decrease in $\lambda$ was most rapid:

$$\lambda(t)' = \frac{\lambda_\infty \theta \alpha^2 (\ln \alpha)}{(\alpha^2 + \theta)^2}.$$

**Data collection**

**Seroprevalence data**

An extensive search of the hepatitis literature identified 172 reports of anti-HAV seroprevalence by age group. From the analysis, 15 small datasets, defined as averaging $<5$ persons per age within the age range covered by the study, were removed. The remaining 157 datasets (Table 1) included 7 studies from sub-Saharan Africa, 6-12 31 from the Americas, 13-36 40 from Asia and the Pacific, 37-62 62 from Europe, 63-108 and 17 from the Middle East and North Africa, 109-124.

For our sample data we assumed that reported seroprevalence rates provide a correct measure of population immunity and that the sample population was representative of the source population. In particular, we assumed that the distribution of study participants by age within reported age groups reflected the actual distribution by age of the source population and that the sample population was not skewed towards the young or old within each age group.

**Water data**

The United Nations’ estimates of the percentage of each nation’s population with access to improved water in 1980, 1990 and 2000, or as many years as were reported, were collected. 125  A line was fit through these points and used to estimate the level of water coverage for each study at the time of data collection.

### Table 1 Location of included studies, by region

<table>
<thead>
<tr>
<th>Region</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa ($n = 7$)</td>
<td>Cameroon, Dem. Rep. Congo, Ethiopia, Senegal, Sierra Leone, South Africa (2)</td>
</tr>
<tr>
<td>Americas ($n = 31$)</td>
<td>Argentina (2), Belize, Bolivia (3), Brazil (10), Canada (3), Chile (2), Costa Rica, Dominican Republic, Jamaica, Mexico, Peru, Uruguay (2), USA, Venezuela (2), Australia (2), China (9), Hong Kong (2), India (6), Indonesia (2), Japan (8), Korea, Malaysia, Nepal, Singapore, Sri Lanka, Thailand (5), Vietnam</td>
</tr>
<tr>
<td>Asia ($n = 40$)</td>
<td>Australia (2), China (9), Hong Kong (2), India (6), Indonesia (2), Japan (8), Korea, Malaysia, Nepal, Singapore, Sri Lanka, Thailand (5), Vietnam, Austria (2), Belgium (2), Czech Republic, Denmark, France, Germany (4), Greece (5), Greenland, Iceland (2), Italy (13), Netherlands (2), Poland (4), Portugal (2), San Marino</td>
</tr>
<tr>
<td>Europe ($n = 62$)</td>
<td>Austria (2), Belgium (2), Czech Republic, Denmark, France, Germany (4), Greece (5), Greenland, Iceland (2), Italy (13), Netherlands (2), Poland (4), Portugal (2), San Marino</td>
</tr>
</tbody>
</table>
When study populations were clearly identified as urban or rural, an urban- and rural-specific water coverage estimate was also recorded. When data on water coverage for high-income nations (Europe, United States, Canada, Australia and Japan) were missing, we assumed that those nations had 100% water coverage. Data for six of the surveys conducted in low-income areas were not available.

**Socioeconomic data**

The Human Development Index (HDI) calculated by the United Nations Development Programme (UNDP) incorporates data on life expectancy, education and standard of living [gross domestic product (GDP)] into an index number. The HDI for each nation in 1975, 1980, 1985, 1990, 1995 and 2001, or as many years as were reported, were collected and a linear fit used to estimate HDI for each study year.126 Data for four surveys were unavailable. We also collected the United Nations estimates of the per capita GDP in current international dollar purchasing power parity (PPP) for 1986, 1990 and 2000, or as many years as were reported.127 A line was fit through these points and used to estimate GDP per capita for each study at the time of data collection. Data for 17 surveys were unavailable.

Linear regression was used to fit models for predicting the infection rate at the time of data collection. Predictor variables included the proportion of the population with access to clean water, HDI, per capita GDP and survey year. Simple and multiple regressions, with and without interaction terms, were fit.

**Results**

Force of infection plots (Figure 3) and their corresponding age-seroprevalence curves (Figure 4) were obtained for 157 datasets using the techniques described in the Methods section.

A crude value for the average annual infection rate in each region was calculated by taking the mean infection rate of all the studies from that region (Table 2). Average annual infection rates were highest in Africa (0.60, 0.6 or 600 infections per 1000 susceptible persons per year), followed by the Americas (0.34), the Middle East (0.22), and Asia (0.21). Rates in Europe (0.01, or 10 infections per 1000 susceptible persons per year) were the lowest of any region.

The ability of water and two measures of socioeconomic development to predict infection rates was assessed using simple linear regression (Table 3). The estimated percentage of the population with access to an improved water source, as defined by the United Nations, is a significant predictor of the force of infection ($r^2 = 0.37$) (Figure 5). A correlation for region-specific mean water coverage rates and mean forces of infection showed a very strong correlation ($r^2 = 0.98$) (Figure 6).
The HDI and per capita GDP are also significant independent predictors of infection rate (Table 3). When variables for both water coverage and HDI are included in the model, both water coverage (P < 0.0001) and HDI (P = 0.038) are significant predictors (r^2 = 0.39). When variables for both water coverage and GDP are included in the total model, water remains a significant variable (P < 0.0001) and the correlation increases to r^2 = 0.42, although GDP is not significant (P = 0.354).

Study year is an independent predictor of infection rate for both Europe (r^2 = 0.12, P = 0.005) and Asia (r^2 = 0.11, P = 0.035), where water coverage rates in high-income nations have not changed significantly over the survey years. In most

Table 2 Regional statistics for force of infection

<table>
<thead>
<tr>
<th>Total</th>
<th>Africa</th>
<th>Americas</th>
<th>Asia</th>
<th>Europe</th>
<th>Middle East</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of surveys</td>
<td>157</td>
<td>7</td>
<td>31</td>
<td>40</td>
<td>62</td>
</tr>
<tr>
<td>Average infection rate (at time of data collection) (mean ± SD)</td>
<td>0.18 ± 0.41</td>
<td>0.60 ± 0.24</td>
<td>0.34 ± 0.74</td>
<td>0.21 ± 0.37</td>
<td>0.01 ± 0.02</td>
</tr>
<tr>
<td>Surveys from 1990 and later (n = 83)</td>
<td>0.19 ± 0.36</td>
<td>0.49 ± 0.25</td>
<td>0.25 ± 0.44</td>
<td>0.31 ± 0.45</td>
<td>0.01 ± 0.01</td>
</tr>
<tr>
<td>Surveys from before 1990 (n = 74)</td>
<td>0.16 ± 0.47</td>
<td>0.62 ± 0.22</td>
<td>0.68 ± 1.35</td>
<td>0.11 ± 0.22</td>
<td>0.02 ± 0.03</td>
</tr>
</tbody>
</table>

Studies with decreasing force of infection (%): 66 0 42 57 97 41

Mean % of susceptible adults (ages 30–80): 13 0 7 8 24 3

Table 3 Linear correlation for force of infection

<table>
<thead>
<tr>
<th>Total</th>
<th>Africa/ Middle East</th>
<th>Americas</th>
<th>Asia</th>
<th>Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>r^2 = 0.37</td>
<td>r^2 = 0.24</td>
<td>r^2 = 0.60</td>
<td>r^2 = 0.08</td>
</tr>
<tr>
<td>P &lt; 0.0001</td>
<td>P = 0.032</td>
<td>P &lt; 0.0001</td>
<td>P = 0.075</td>
<td>P = NA</td>
</tr>
<tr>
<td>n = 151</td>
<td>n = 19</td>
<td>n = 30</td>
<td>n = 40</td>
<td>n = 62</td>
</tr>
<tr>
<td>HDI</td>
<td>r^2 = 0.26</td>
<td>r^2 = 0.48</td>
<td>r^2 = 0.30</td>
<td>r^2 = 0.33</td>
</tr>
<tr>
<td>P &lt; 0.0001</td>
<td>P = 0.0003</td>
<td>P = 0.001</td>
<td>P = 0.0001</td>
<td>P = 0.0002</td>
</tr>
<tr>
<td>n = 153</td>
<td>n = 23</td>
<td>n = 31</td>
<td>n = 39</td>
<td>n = 60</td>
</tr>
<tr>
<td>GDP</td>
<td>r^2 = 0.13</td>
<td>r^2 = 0.52</td>
<td>r^2 = 0.08</td>
<td>r^2 = 0.12</td>
</tr>
<tr>
<td>P &lt; 0.0001</td>
<td>P = 0.0003</td>
<td>P = 0.133</td>
<td>P = 0.052</td>
<td>P = 0.038</td>
</tr>
<tr>
<td>n = 140</td>
<td>n = 21</td>
<td>n = 31</td>
<td>n = 33</td>
<td>n = 55</td>
</tr>
<tr>
<td>Year</td>
<td>r^2 = 0.01</td>
<td>r^2 = 0.08</td>
<td>r^2 = 0.01</td>
<td>r^2 = 0.11</td>
</tr>
<tr>
<td>P = 0.137</td>
<td>P = 0.192</td>
<td>P = 0.656</td>
<td>P = 0.035</td>
<td>P = 0.005</td>
</tr>
<tr>
<td>n = 157</td>
<td>n = 24</td>
<td>n = 31</td>
<td>n = 40</td>
<td>n = 62</td>
</tr>
</tbody>
</table>

Figure 5 Water Coverage vs Force of Infection (r^2 = 0.37). Each point represents one of the 157 studies included in this analysis. Infection rates were calculated from age-seroprevalence data using our mathematical model. Water coverage data is from the UN [124]

Figure 6 Correlation for Regional Means (r^2 = 0.985). Mean water coverage is the estimated proportion of each region’s population with access to safe drinking water. The estimates for regional infection rates were calculated by taking the mean of the infection rates calculated for studies in each region.
nations with complete water coverage and relatively low forces of infection, the infection rate has continued to decrease over time. To assess possible co-effects of SES, water coverage and study year, interaction terms were added to the model. None of the interaction terms are significant predictors of the infection rate.

Declining infection rates were observed in 65.6% of the surveys (Table 2). None of the studies from Africa showed a decrease in infection rate whereas nearly all of the studies from Europe (97%) did. About half the studies from the Middle East (41%), the Americas (42%) and Asia (57%) indicated decreases in force of infection. An analysis of the factors that predict a decreasing infection rate, regardless of the value of the infection rate at the time of data collection, showed that higher values for HDI, per capita GDP and water coverage are strongly associated with a declining infection rate. Of the surveys >80% showing decreasing infection rates were from nations reporting complete water coverage. The mean water coverage rate was 80% for surveys indicating no decline compared with 97% for surveys showing a decline. Mean HDI values were 0.65 and 0.83, respectively. Although surveys that showed no decrease in the infection rate were found for the entire range of values for water coverage and HDI, a decline in infection rate was seen only in surveys from areas with an HDI value >0.58 or a water coverage rate >55%.

The proportion of adults aged ≥30 without immunity to HAV, an indication of the risk of severe hepatitis A disease, ranged from 0% in Africa to nearly 24% in Europe. The proportion of susceptible adults in surveys showing a decreasing infection rate was 18% compared with only 4% in surveys showing no decrease.

Discussion

We used a symmetric logistic function to fit a time-dependent infection rate for 157 age-seroprevalence datasets. We found evidence that HAV infection rates are decreasing in most regions of the world and that water development is crucial in reducing transmission rates. This work demonstrates the utility of HAV age-seroprevalence plots.132,133 No previously published works for water-related transmission of HAV would improve both our understanding of the factors that lead to decreasing transmission and our ability to model the various modes of transmission of HAV. Serological data will be tremendously useful in developing more detailed studies of transmission.

More detailed information on the changes in SES over time would also improve our model. We focused on the relationship of water, SES, average GDP and force of infection at the time of data collection so as to minimize the error that could be introduced by using water and socioeconomic data collected at different times using different methods and standards. We, therefore, did not fully explore the information on past infection rates provided by the model. In particular, our analysis of the experience of older persons in the population in this paper was limited since the immunity of older adults is probably more reflective of their exposure experience in the distant past than in the recent past.

Given that HAV has multiple modes of transmission, safe drinking water does not eliminate risk of infection. We observed that even in areas with 100% water coverage, the infection rate is decreasing over time. One limitation of the force of infection model used is that it is unable to fit infection rates for urban/rural status may have caused errors in the water coverage estimates. The estimate of regional infection rates did not account for variations in sample size or quality among included studies.

Studies at the local level have found evidence of an association between type of water source and water storage, the quantity and quality of water available, the reliability of water treatment and sanitation facilities and the number of taps in the home in addition to income, crowding, family size and education.3 Water improvements are closely related to other aspects of socioeconomic development, including decreased crowding, increased education and increased income, so it is difficult to separate out the effects of water access from economic growth. Regardless, the higher correlation between water and infection rate indicates that access to clean water may be very important in reducing HAV transmission. We believe that more accurate information on water coverage levels for the surveys included in our sample, such as for water or provincial level rather than the national level, would probably yield a higher correlation between higher water coverage and lower infection rates. For example, several of the studies included in our analysis indicated that their study site was rural or urban. Using the United Nations’ urban- or rural-specific estimates of water coverage rather than the total estimate of water coverage improved the fit of the models, although we did not include this information in our Results section because few studies could be classified by population density.

Determining the mechanism by which water improvements cause a decrease in transmission is challenging. Declining transmission could result from improved hygiene practices (which reduce direct transmission), better sanitation (which reduces waterborne transmission) or a combination of hygiene and sanitation. The separate effects of increased water quantity and improved water quality are also difficult to distinguish. Increased quantity allows more water for washing and cleaning and a decreased need for water storage, which is associated with increased HAV infection risk.3 Increased quality directly decreases risk of disease transmission because of a reduction in the ingestion of the virus in water. Further studies of the mechanisms for water-related transmission of HAV would improve both our understanding of the factors that lead to decreasing transmission and our ability to model the various modes of transmission of HAV. Serological data will be tremendously useful in developing more detailed studies of transmission.
epidemics. Adding a sine function to our force of infection equation would allow for these epidemic fluctuations in the infection rate to be modelled. However, because epidemics are seen only in low infection rate areas and because our correlations looked only at the infection rate at the time of data collection (which is the same for either force of infection model), this limitation should not change the conclusions of this study. Also, the lack of realistic age structuring of contacts in the model is a limitation that we do not feel invalidates any of our conclusions.

We also chose to treat all samples as independent rather than as part of a longitudinal study. Although we found many examples of data collected over time from the same location (as was the case for some studies from Japan, Italy and other areas), the infection rates for some of the data could not be fit because the seroprevalence rate of the presumed age cohort decreased over time, violating our assumption of lifelong immunity. For example, a series of studies in secondary schools in Bangkok, Thailand, showed decreasing seroprevalence in age cohorts over time, violating our assumption of lifelong immunity. For the seroprevalence rate of the presumed age cohort decreased over time, a force of infection curve was fit to each individual dataset. The information provided by each is valid if independence is assumed.

This analysis of HAV seroprevalence has revealed important patterns about regional risk factors for transmission. Water and socioeconomic development are associated with decreasing infection rates and increasing adult susceptibility levels. In Africa, where transmission is highly endemic, the provision of clean water for both drinking and hygiene is a priority for disease control. Developed countries in Europe, North America and other parts of the world may be rapidly increasing their potential for epidemics as the proportion of individuals susceptible to HAV infection increases. Targeted vaccination for high-risk groups should be considered in these populations. In the transition economies of Asia, Latin America and the Middle East, infection rates are beginning to decrease. The gradient of the decline, whether rapid or gradual, may determine if HAV transmission remains endemic or shifts to epidemic patterns. In this paper we have demonstrated that HAV serology provides an excellent tool for assessing infection rates in populations. Studies that sample different risk groups by water quality, economic status and/or sanitary conditions and gather lifetime histories of exposure should be able to further elucidate risk factors for transmission.

**KEY MESSAGES**

- Seroprevalence by age data can be used to calculate the hepatitis A virus (HAV) infection rate over time.
- Mathematical models indicate that HAV infection rates are declining in most regions of the world.
- Hepatitis A virus infection rate is correlated with access to clean drinking water and socioeconomic indicators.

**References**


608 INTERNATIONAL JOURNAL OF EPIDEMIOLOGY


