Appendix to Comparability of different methods for estimating influenza infection rates over a single epidemic wave

Vernon J Lee, Mark I Chen, Jonathan Yap, Jocelyn Ong, Wei-Yen Lim, Raymond T P Lin, Ian Barr, Jimmy B S Ong, Tze Minn Mak, Lee Gan Goh, Yee Sin Leo, Paul M Kelly, Alex R Cook

March 10, 2011

Statistical methodology

All four approaches to estimate infection rates require information from multiple sources. The Bayesian statistical paradigm (1) is particularly well suited to combining information from different sources while allowing for proper propagation of uncertainty. For all four methods, we use non-informative, flat prior distributions on all parameters unless external data were available to provide an informative prior distribution. For these informative prior distributions, we started with a non-informative prior, fit a model to the external data, and used the resulting posterior distribution *including all within model uncertainty* as the informative prior for the main analysis. Where an analytical form could not be found for it, the posterior distribution from fitting to external data was approximated by a (multivariate) Gaussian distribution.

When conjugate priors were not available, models were fit using Markov chain Monte Carlo integration, with univariate Gaussian proposal distributions centred on the current value and the Metropolis–Hastings acceptance algorithm (1). Proposal bandwidths were selected by trial and error until satisfactory mixing of MCMC chains was achieved; convergence of MCMC output was assessed by visual inspection of trace plots. When mixing was slow, chains were run for longer and thinned to reduce autocorrelations, but for models in which the MCMC routine mixed ostensibly quickly through the posterior distribution, no thinning was used.

Method 1: paired serological surveys

Method 1 used paired serology, with a baseline blood sample and the highest follow up sample, either mid- or post-epidemic. A "fourfold rise" (from baseline in (1:t,1:2t) to (1:4t,1:8t) for $t \in \{5, 10, 20, \ldots\}$; note that although this does not guarantee a fourfold rise in the uncensored titres standard terminology is to call it a fourfold rise, a convention we abide by here) is deemed to be a seroconversion. The proportion seroconverting must however be scaled up to account for imperfect sensitivity of serology. To do this, we used data on seroconversions associated with RT-PCR confirmed infection from (2) and (3).

Data used:

- number of seroconversions;

- number of possible seroconversions;
- estimated sensitivity of seroconversion.

The number of seroconversions and number at risk were taken from a seroepidemiological study performed in Singapore during the first wave (3). The sensitivity of seroconversion relative to RT-PCR confirmed infection was estimated from historic (2) and contemporary (3) studies. The estimates from these two studies individually were close enough to warrant merging the two and using the historic data alongside the 2009-H1N1 data.

Notation:

- $n =$ the number of individuals in follow-up—known;
- $x =$ the number of individuals that seroconverted—known;
- $\sigma_1 =$ sensitivity of the seroconversion test versus RT-PCR, i.e. the probability of a seroconversion given a virologically confirmed infection;
- $p =$ the proportion infected in the population.

Full model for data and parameters:

$$\begin{aligned} x &\sim \mathrm{Bin}(n, \sigma_1 p) \\ \sigma_1 &\sim \mathrm{Be}(675, 174) \\ p &\sim \mathrm{U}(0, 1). \end{aligned}$$

The prior distribution for σ_1 comes from taking the conjugate prior $\mathrm{U}(0,1) = \mathrm{Be}(1,1)$ to a binomial model for seroconversion given RT-PCR confirmed infection using the data in (2) and (3).

Simplified representation of model:

$$\begin{aligned} \hat{p} &= \frac{x}{n\hat{\sigma}_1} \\ \hat{\sigma}_1 &= 0.79 \end{aligned}$$

Assumptions:

- The findings from (2) and (3) generalise to our study population. The two papers show similar results, supporting this assumption.
- Our cohort is a random sample from the population—the fact that it is not means there may be a systematic and unquantifiable bias in the results.
- Seroconversions are independent—some study participants had co-habitants also in the study, leading to a systematic underestimate of the uncertainty. The magnitude of the bias is likely to be small due to the small fraction of households with multiple members in the study.

Table 1: Posterior mean and 95% credible intervals for parameters of method 1

Parameter	Estimate	Lower bound	Upper bound
σ_1	0.79	0.77	0.82
p	0.17	0.14	0.20

Estimation:

The joint distribution of (σ_1, p) is sampled via Markov chain Monte Carlo.

Method 2: cross-sectional serological surveys

Method 2 uses cross-sectional serological samples of the population. We emulate this by delinking the paired serology data. In both samples, a titre of 1:40 is taken as evidence of exposure to the virus. To account for pre-existing antibodies (resulting from cross reaction to other strains or measurement error), the estimated proportion above this threshold pre-epidemic is "subtracted" from the estimated proportion post-epidemic.

Data used:

- number of people at pre- and post-epidemic with titres above 1:40;
- number of people in these two samples;
- cross sectional antibody data from participants with RT-PCR confirmed infection.

The emulated cross-sectional data are derived from ref (3), as is the sensitivity estimate.

Notation:

- n_1 = the number of individuals giving samples at baseline—known;
- n_2 = the number of individuals at final follow up—known;
- x_1 = the number of individuals with high (above 1:40) initial titres—known;
- x_2 = the number of individuals with high final titres—known;
- q = the proportion of the population with naturally high titres, pre-infection (presumed);
- σ_2 = the proportion of the population with high titres at final follow that were infected and had low initial titres;
- p = the proportion infected in the population.

Table 2: Posterior mean and 95% credible intervals for parameters of method 2

Parameter	Estimate	Lower bound	Upper bound
σ_2	0.79	0.67	0.88
q	0.03	0.02	0.04
p	0.12	0.09	0.17

Full model for data and parameters:

$$\begin{aligned}
x_1 &\sim \mathrm{Bin}(n_1, q) \\
x_2 &\sim \mathrm{Bin}(n_2, q + (1-q)\sigma_2 p) \\
\sigma_2 &\sim \mathrm{Be}(46, 12) \\
q &\sim \mathrm{Be}(3, 55) \\
p &\sim \mathrm{U}(0, 1).
\end{aligned}$$

The prior distributions for q and σ_2 come from taking the conjugate priors $\mathrm{U}(0,1) = \mathrm{Be}(1,1)$ to a binomial model for high titres at baseline and follow up for individuals with RT-PCR confirmed infection during the study using the data from (3).

Simplified representation of model:

$$\begin{aligned}
\hat{p} &= \frac{x_2/n_2 - x_1/n_1}{(1 - x_1/n_1)\hat{\sigma}_2} \\
\hat{\sigma}_2 &= 0.79.
\end{aligned}$$

Assumptions:

- The sensitivity and initial high titres findings generalise to our study population.
- Our cohort is a random sample from the population—which, again, it is not.
- Titre levels are independent (i.e. no household effects).
- We treat these as two independent cross-sectional surveys, but in actuality they are paired (see method 1 for the method accounting for pairing).

Estimation:

The joint distribution of (σ_2, q, p) is sampled via Markov chain Monte Carlo.

Method 3: ILI data from sentinel GPs

Data used:

- daily number of ILI consults at GP sentinel network;
- daily number of GPs reporting;

- interview of participants in serological study and their seroconversion status to estimate consultation rates;
- sensitivity of seroconversion (as method 1).

ILI consult data come from a GP sentinel network previously described (4). The questionnaire data are previously unpublished and were collected from the same group of patients as provided sera in ref (3). Sensitivity data sources are as in method 1.

Notation:

Unobserved:

- $p =$ the proportion of the population infected;
- $p_1 =$ the proportion of infections leading to ILI consults;
- $p_2 =$ the proportion of ILI consults that correspond to actual infection;
- $\sigma_3 =$ the probability of seroconverting given (RT-PCR confirmed) infection;
- $N^s_{IVS} =$ the number of people infected, visiting primary care with an ILI and seroconverting in the serology–questionnaire study;
- $N^s_{IS} =$ the number of people infected and seroconverting in the serology–questionnaire study;
- $N^s_I =$ the number of people infected in the serology–questionnaire study;
- $N^p_V =$ the number of people visiting primary care with an ILI in the population as a whole;
- $\kappa =$ a parameter controlling the variability of GP consultations around the mean.

Observed:

- $N^s_S =$ the number of people seroconverting in the serology–questionnaire study;
- $N^s_{VS} =$ the number of people seroconverting and visiting primary care with an ILI in the serology–questionnaire study;
- $N^s_V =$ the number of people visiting primary care with an ILI in the serology–questionnaire study;
- $\pi = 1/2486$ the proportion of primary care consultations attributable to a single GP in our study—assumed known;
- $N^s =$ the number of people in the serology–questionnaire study;
- $N^p =$ the number of people in the population (we restrict attention to resident adults throughout);
- $D_t =$ the daily number of ILIs reported on day t;
- $F_t =$ the number of GPs faxing or emailing a report on day t;
- $\mathcal{T} =$ the length of time of the two studies.

Full model for data and parameters:

$$
\begin{array}{rcl}
N^s_{IS} & \sim & \mathrm{Bin}(N^s_I, \sigma_3) \\
\sigma_3 & \sim & \mathrm{Be}(675, 174) \\
(N^s_S - N^s_{VS}) & \sim & \mathrm{Bin}(N^s_{IS} - N^s_{IVS}, \sigma_3) \\
N^s_{IVS} & \sim & \mathrm{Bin}(N^s_{VS}, \sigma_3) \\
N^s_{IVS} & \sim & \mathrm{Bin}(N^s_{IS}, p_1) \\
N^s_{IVS} & \sim & \mathrm{Bin}(N^s_V, p_2) \\
N^s_I & \sim & \mathrm{U}(0, N^s) \\
D_t & \sim & \mathrm{NegBin}(\mu_t = F_t \pi N^p_V / \mathcal{T}, k = \kappa) \\
\kappa & \sim & \mathrm{U}(0, 10) \\
p_1 & \sim & \mathrm{U}(0, 1) \\
p_2 & \sim & \mathrm{U}(0, 1) \\
p & \sim & \mathrm{U}(0, 1) \\
N^p_V & = & p p_1 N^p / p_2
\end{array}
$$

Simplified representation of model:

$$
\begin{array}{rcl}
\hat{p} & = & \dfrac{\hat{N^p_V}\hat{p_2}}{N^p \hat{p_1}} \\
\hat{p_1} & = & \dfrac{N^s_{VS}}{N^s_S} \\
\hat{p_2} & = & \dfrac{N^s_{VS}}{N^s_V \hat{\sigma}_3} \\
\hat{N^p_V} & = & \dfrac{\sum_t \frac{D_t}{F_t}}{\pi}.
\end{array}
$$

Note that alternative simplified point estimates of the number of ILI consults in the population exist and it is not clear which gives the least bad representation of the estimate accounting for all within model uncertainty.

Assumptions:

- The sensitivity findings generalise to our study population.
- Our cohort is a random sample from the population.
- Seroconversions are independent (i.e. no household effects).
- The GPs in the surveillance network are representative.
- Observed ILIs in the GP study are negative binomial distributed, i.e. inflated by a factor κ relative to a Poisson to account for the inhomogeneous pattern of consults caused by the epidemic wave and day of week effects.

Table 3: Posterior mean and 95% credible intervals for parameters of method 3

Parameter	Estimate	Lower bound	Upper bound
σ_3	0.79	0.77	0.82
p_1	0.20	0.13	0.28
p_2	0.67	0.47	0.86
N^s_{IS}	124	112	137
N^s_{IVS}	24	20	30
κ	2.95	1.88	4.53
N^p_V	116 000	100 000	133 000
p	0.15	0.10	0.25

Estimation:

A two-stage approach is taken. In the first round, Markov chain Monte Carlo is used to sample the joint distribution of $(p_1, p_2, \sigma_3, N^s_{IVS}, N^s_{IS})$. A bivariate Normal distribution is then used to approximate the joint distribution of (p_1, p_2) which is then used as an informative prior for round two. In the second round, Markov chain Monte Carlo is used to sample the joint distribution of the population level estimands $(p_1, p_2, \kappa, N^p_V)$.

Method 4: Laboritory surveillance and ILI data from sentinel GPs

Here, we replace the estimated probability of H1N1 given ILI consult from the questionnaire associated with the serology study by an estimate from lab surveillance.

Data used:

- daily number of ILI consults at GP sentinel network;
- daily number of GPs reporting;
- weekly numbers of swabs of patients presenting with ILI testing positive and negative to H1N1;
- sensitivity of RT-PCR from (2);
- sensitivity of seroconversion (as method 1);
- interview of participants in serological study and their seroconversion status to estimate consultation rates.

Data sources as per method 3, with in addition laboratory data from Singapore's National Laboratory (published in part in ref (5)).

Notation:

Unobserved:

- L_w^I = the number of people sampled for lab testing that were infected in week w;
- p_1 = the proportion of infections leading to ILI consults;
- p_2 = the proportion of ILI consults that correspond to actual infection aggregated over the study period of the sero-cohort questionnaire, subsequently replaced by lab-derived estimates;
- p_{2w} = the proportion of ILI consults that correspond to actual infection in week w;
- σ^{PCR} = the probability of RT-PCR confirmed infection given *any* form of confirmed infection;
- σ^{sero} = the probability of seroconverting given (RT-PCR confirmed) infection;
- N_{IVS}^s = the number of people infected, visiting primary care with an ILI and seroconverting in the serology–questionnaire study;
- N_{IS}^s = the number of people infected and seroconverting in the serology–questionnaire study;
- N_{Vt}^p = the number of people visiting primary care with an ILI in the population on day t;
- N_{IVt}^p = the number of infected people visiting primary care with an ILI in the population on day t;
- N_{IV}^p = the number of infected people visiting primary care with an ILI in the population over the course of the study;
- N_I^p = the number of people infected in the population.

Observed:

- L_w^+ = the number of people sampled for lab testing that test positive in week w;
- L_w^- = the number of people sampled for lab testing that test negative in week w;
- N_S^s = the number of people seroconverting in the serology–questionnaire study;
- N_{VS}^s = the number of people seroconverting and visiting primary care with an ILI in the serology–questionnaire study;
- $\pi = 1/2486$ the proportion of primary care consultations attributable to a single GP in our study—assumed known;
- N^p = the number of people in the population (we restrict attention to resident adults throughout);
- D_t = the daily number of ILIs reported on day t;
- F_t = the number of GPs faxing or emailing a report on day t;
- $w(t)$ = the week that contains day t.

Full model for data and parameters:

$$
\begin{aligned}
L_w^+ &\sim \mathrm{Bin}(L_w^I, \sigma^{\mathrm{PCR}}) \\
L_w^I &\sim \mathrm{Bin}(L_w^+ + L_w^-, p_{2w}) \\
\sigma^{\mathrm{PCR}} &\sim \mathrm{Be}(731, 62) \\
N_{IS}^s &\sim \mathrm{Bin}(N_I^s, \sigma^{\mathrm{sero}}) \\
(N_S^s - N_{VS}^s) &\sim \mathrm{Bin}(N_{IS}^s - N_{IVS}^s, \sigma^{\mathrm{sero}}) \\
N_{IVS}^s &\sim \mathrm{Bin}(N_{VS}^s, \sigma^{\mathrm{sero}}) \\
N_{IVS}^s &\sim \mathrm{Bin}(N_{IS}^s, p_1) \\
N_{IVS}^s &\sim \mathrm{Bin}(N_V^s, p_2) \\
p_1 &\sim \mathrm{U}(0, 1) \\
p_2 &\sim \mathrm{U}(0, 1) \\
\sigma^{\mathrm{sero}} &\sim \mathrm{Be}(675, 174) \\
p_{2w} &\sim \mathrm{U}(0, 1) \\
N_{Vt}^p &\sim \mathrm{Ga}(1, 1/100\,000) \\
D_t &\sim \mathrm{Po}(F_t \pi N_{Vt}^p) \\
N_{IVt}^p &\sim \mathrm{Bin}(N_{Vt}^p, p_{2w(t)}) \\
N_{IV}^p &= \sum_t N_{IVt}^p \\
N_I^p &= \frac{N_{IV}^p}{p_1} \\
p &= \frac{N_I^p}{N^p}
\end{aligned}
$$

Simplified representation of model:

$$
\begin{aligned}
\hat{p} &= \frac{\hat{N_I^p}}{N^p} \\
\hat{N_I^p} &= \frac{\hat{N_{IV}^p}}{\hat{p_1}} \\
\hat{p_1} &= \frac{N_{VS}^s}{N_S^s} \\
\hat{p_2} &= \frac{N_{VS}^s}{N_V^s \hat{\sigma}^{\mathrm{sero}}} \\
\hat{N_{IV}^p} &= \sum_t \hat{N_{IVt}^p} \\
\hat{N_{IVt}^p} &= \frac{\hat{N_{Vt}^p}}{p_{2\hat{w}(t)}} \\
p_{2\hat{w}} &= \frac{1}{\hat{\sigma}^{\mathrm{PCR}}} \frac{L_w^+}{L_w^+ + L_w^-} \\
\hat{\sigma}^{\mathrm{PCR}} &= 0.92 \\
\hat{\sigma}^{\mathrm{sero}} &= 0.79 \\
\hat{N_{Vt}^p} &= \frac{D_t}{F_t \pi}
\end{aligned}
$$

Assumptions:

- The sensitivity findings generalise to the lab samples.
- Our cohort is a random sample from the population.
- The GPs in the surveillance network are representative.
- Observed ILIs in the GP study an any particular day given the number of ILIs in the community on that day are Poisson.
- Lab samples were randomly drawn from the population of ILI cases.
- Independence of lab samples, seroconversions, infection status of patients consulting for ILI between and within days.
- The proportion of ILIs with H1N1 infection is piece-wise constant and changes only on Sundays with the beginning of a new e-week.

Estimation:

A three stage procedure is used. In the first round, the joint distribution of $(p_{2w}, \sigma^{\mathrm{PCR}})$ is sampled using MCMC and the lab data. In the second round, the joint distribution of $(p_1, p_2, \sigma^{\mathrm{sero}}, N_I^s, N_{IV}^s)$ is sampled (as in method 3) using Markov chain Monte Carlo. In the final round, the posterior distribution for all other estimands is sampled by Monte Carlo simulation, exploiting conjugacy to obtain beta posteriors for p_1 and gamma posteriors for N_{Vt}^p. The posterior distributions for all other terms can be sampled directly.

Table 4: Posterior mean and 95% credible intervals for parameters of method 4. Estimands that vary with time are not shown.

Parameter	Estimate	Lower bound	Upper bound
σ^{PCR}	0.92	0.90	0.94
σ^{sero}	0.79	0.77	0.82
p_1	0.20	0.13	0.28
p_2	0.67	0.47	0.86
N^s_{IV}	24	20	31
N^s_I	124	113	138
N^p_{IV}	61 000	57 000	64 000
N^p_I	317 000	216 000	499 000
p	0.12	0.08	0.18

Age stratified analyses

We undertook analyses of the infection rate in two ways:

- **age stratified**, in which infection rates in five age groups (20–24, 25–34, 35–44, 45–54 and 55 or over) were analysed independently. For methods using GP ILI consults, we assumed all ages had the same probability of consulting with an ILI when infected and that the proportion of ILIs due to H1N1 infection were the same for all ages. This assumption was needed since there was insufficient information on these proportions when the serological questionnaire was divided by age groups, as the number of ILI cases among seroconvertors within any particular age group was too small to obtain usable estimates. All other parameters were estimated using data from the subset of the data corresponding to that age group only.
- **non-age stratified**, in which we derived estimates of the adult infection rate under the assumption that infection rates are constant for all age groups. The motivation for this assumption is that for some of the age groups, there is considerable uncertainty on the model parameters, which can only be reduced by pooling information. An obvious, though complicated and computer-intensive, alternative to this homogeneity assumption that would also provide narrower estimates would be to develop an hierarchical model (see e.g. (1)) or to introduce a functional form for the effect of age on infection rates as part of a regression analysis.

References

[1] Gelman A, Carlin J B, Stern H S, Rubin D B (1995) *Bayesian Data Analysis* London: Chapman & Hall.

[2] Zambon M, Hays J, Webster A, Newman R, Keene O (2001). Diagnosis of Influenza in the Community: Relationship of Clinical Diagnosis to Confirmed Virological, Serologic, or Molecular Detection of Influenza. *Arch Intern Med* 161:2116–22.

[3] Chen MIC, Lee VJM, Lim W-Y, Barr IG, Lin RTP, et al (2010). 2009 influenza A(H1N1) seroconversion rates and risk factors among distinct adult cohorts in Singapore. *J Am Med Assoc* 303:1383–91.

[4] Ong JBS, Chen MIC, Cook AR, Lee HC, Lee VJ, et al (2010). Real-Time Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-Like Illness from General Practice and Family Doctor Clinics in Singapore. *PLoS One* 5(4):e10036.

[5] Cutter JL, Ang LW, Lai FYL, Subramony H, Ma S, James L (2009). Outbreak of Pandemic Influenza A (H1N1-2009) in Singapore, May to September 2009. *Ann Acad Med Singapore* 39:273–82.