Workplace Project Portfolio (WPP)

Sandra Carlson

FLUTRACKING SURVEILLANCE: A comparison of 2007 NSW symptom rates with laboratory confirmed influenza notifications and an examination of factors affecting community receipt of H1N109 vaccination

Table of Contents

Preface

Student role

At the time of completion of part A of the project, I was employed as a Biostatistical Trainee at New South Wales (NSW) Department of Health and was placed at Hunter New England Population Health (HNEPH) in the Flutracking team. My task was to validate the Flutracking survey using time series analysis techniques, with assistance from my statistical supervisor (Dr Frank Tuyl), my content supervisor (Dr Craig Dalton) and co-supervisor (Mr David Muscatello).

At the time of completion of part B of the project, I was employed by HNEPH to manage the Flutracking survey. In conjunction with Dr Craig Dalton (my line manager at HNEPH), we devised a research question, based on an existing gap in knowledge for the Flutracking survey program: understanding factors affecting participant vaccination against influenza.

For both parts of this project, I extracted multiple years of data from separate files, cleaned the data, and merged the data into a final data set. I also manipulated a number of variables in the dataset (such as adding a household identifier) to analyse the data accurately, and created a single symptom rate for each week of data. As there were no other colleagues at HNEPH with an understanding of the Flutracking dataset at this detailed level, this data preparation work was performed unassisted.

For part A of this project I extracted laboratory notifications for influenza from the NSW Department of Health notifiable diseases database, using the Health Outcomes Statistical Toolkit (HOIST). Counts were aggregated into weeks based on the date of specimen collection. The extraction of laboratory notifications was performed with the assistance of Mr David Muscatello.

My role also involved conducting all statistical analyses, with guidance from my statistical supervisors, Dr Frank Tuyl and Dr Patrick Kelly, and with assistance from my course notes from the Categorical Data Analysis (CDA) subject and Linear Models (LMR) subject. Guidance from my supervisors was based on my communication of the subject matter and data available to analyse.

The value added to the Flutracking project was the finding that Flutracking data correlated as expected with laboratory data (providing reassurance that Flutracking is measuring influenza-like illness) and a better understanding of factors that affect

vaccination. The findings of a correlation with laboratory data were published in the journal of Communicable Diseases Intelligence, and I have presented these findings at The Public Health Association Australia (PHAA) annual conference in 2009 and at the NSW Department of Health Epidemiology Special Interest Group meeting in 2010. I also presented these results internally to HNEPH staff at the annual Population Health Symposium held in 2008.

The information on factors affecting vaccination was presented to HNEPH staff at a staff meeting on 1 June 2010, and will be distributed to stakeholders of the survey, such as the Department of Health and Ageing, to better inform methods of promoting vaccination against influenza.

Reflections on learning

Communication skills

Throughout this project I had face-to-face meetings with Dr Craig Dalton and Dr Frank Tuyl, and mostly corresponded by telephone meetings and email contact with Dr Patrick Kelly and Mr David Muscatello. Dr Kelly and Mr Muscatello did not have access to the dataset that was being analysed. Therefore, I learnt how to clearly communicate the subject matter relating to the dataset, the scope of the data, the data layout, the data limitations, and analytical findings already known from the dataset. In addition, I communicated my results and any issues I had in performing statistical analyses for the project. This process required communicating clearly and specifically (especially via email) to ensure my supervisors and I had the same understanding of the project and statistical results. In addition, my content supervisor (Dr Dalton) was based in Bhutan for part B of this project, and I was required to communicate the statistical methods used and results from these methods to this supervisor via email and telephone. The write up of the Workplace Project Portfolio (WPP) report also required similar communication skills. These communication exercises taught me how to clearly articulate data issues and statistical concepts and issues to an audience with detailed knowledge of statistical procedures, as well as an audience without statistical expertise.

Work patterns/planning

Completing a workplace project is a very different process than completing a coursework subject - coursework subjects are broken down into discrete modules with assessments due at pre-defined deadlines, whereas the WPP does not contain pre-defined tasks due at set intervals (rather, one very large sometimes overwhelming

task), and therefore, required more discipline to ensure that the overall project would be completed on time. The WPP also differs, in that there is no guarantee that the planned statistical analysis will be appropriate, and that model checking may reveal additional analyses required to be performed. In addition, time must be allowed for feedback from multiple supervisors and revisions. I learnt to break down the WPP project into chunks (for example preparing data for analysis, performing statistical analyses, and writing different sections of the report) to submit to my supervisors at pre-defined times, as well as factoring time for feedback from multiple supervisors and revisions.

Throughout this project I had regular meetings arranged in advance mutually by my statistical supervisors and myself. I always ensured that I called/arrived on time for each meeting. Before the end of each meeting I summarised what tasks I would have completed before the next meeting to ensure that I could meet task deadlines contributing to completion of the project on schedule.

Statistical principles and methods

This project provided several challenges in applying statistical principles. Part A of this project involved application of time series analysis techniques to autocorrelated data series – statistical techniques to control for autocorrelation had not been taught in any Biostatistics Collaboration of Australia (BCA) subject I have completed. Therefore, this provided a challenge beyond the scope of my Master of Biostatistics course. There were many challenges within the time series analysis component of the project. For example, we had originally planned to analyse 2007 and 2008 Flutracking and laboratory data. However, as Flutracking data were only collected from May to October for each of these years, understanding how autoregressive integrated moving average (ARIMA) analysis treated missing data (data between October 2007 and May 2008) provided a challenge. Also, other timeseries analysis techniques were explored such as fitting spline curves to the data, and so additional learning was required to understand the application of this to the dataset, and the advantages and disadvantages over ARIMA analysis.

The original plan for part B of this project was to apply logistic regression techniques learnt in CDA to the Flutracking dataset. However, the nature of the dataset required intracluster correlation to be accounted for. I have not completed any BCA subjects that provide training on logistic regression modelling, controlling for intracluster correlation. Therefore, this project not only provided me with an opportunity to apply methods learnt in CDA, but also extended my knowledge of logistic regression modelling to the use of

a mixed effects logistic regression model, including how to specify and run this model in Stata, how to assess the fit of this type of model, and how to test model assumptions in this type of model.

Another challenge in this project was to understand how the scope of the Flutracking data set might affect the statistical modelling results. For example, I learnt the impact that missing data can have on results, and the difficulties with quantifying this impact. Unvaccinated participants who did not complete all Flutracking surveys from October to December 2009 may have actually been vaccinated with Panvax after dropping out of the survey, and there may be some bias in the Panvax vaccination data. However, it is possible that the participants who dropped out of the survey early may be less concerned with influenza prevention, and therefore less likely to be vaccinated with Panvax than those participants who completed the final survey in 2009. With the data available, it was difficult to gain a further understanding of these possible effects.

A further challenge was to apply model assumption testing to logistic regression, using my learnings from the LMR course. Assumption testing was not discussed in detail in the CDA course (and reference was made to refer to LMR notes), and I found some differences in the requirements for assumption testing between linear regression and logistic regression modelling.

Statistical computing

Data were cleaned using SAS version 9.1.3 and version 9.2, and statistical analyses were performed using SAS version 9.1.3 and version 9.2 and Stata version 10.0. In the BCA subjects, example data sets provided were generally in a layout that was ready to apply Stata commands. However, the Flutracking data required a significant amount of manipulation to achieve a format ready to apply commands for statistical modelling. For example, the original data set (once merged from multiple files) contained multiple survey records per participant. For the ARIMA analysis, this needed to be reduced to a single summary symptom rate for all participants for each week. For the logistic regression, the dataset needed to be reduced to a single record per person. These data manipulation processes were performed in SAS, and required application of some commands not previously taught in any BCA subjects (for example, ‘last.’). In addition, for the logistic regression analysis a cluster identification variable needed to be added to the dataset using an existing participant identifier and a variable specifying whether a participant was a respondent or a household member and who their primary respondent was. Although the SAS code used to create this identifier did not involve complex commands, it did require thought regarding the correct logic, and producing

several summary tables of final data to test the output produced. All data manipulations in SAS were performed without any guidance/advice. I solely relied on the resources of the 'SAS help' menu.

For the statistical analyses performed in SAS and Stata, I learnt several new commands. For example, in SAS I learnt the 'proc arima' command used for time series analysis, and in Stata I learnt the 'xtmelogit' command used for mixed effects multilevel models and the 'VCE' option in the 'logit' command used for intraclass correlation. Performing logistic regression in Stata also reinforced the commands learnt in CDA, such as 'logit', 'estat gof', and 'lrtest'.

Team work

Communication with other team members

I worked with several other team members on this project: My statistical supervisors (Dr Patrick Kelly and Dr Frank Tuyl) and my content supervisors (Dr Craig Dalton and Mr David Muscatello). I met with each of these supervisors regularly (either by phone or in person), and liaised by email as required. As Dr Dalton was located overseas for part B of this project, most contact was via email. I arranged for a teleconference to occur between myself and my supervisors near the beginning of the project so that any differences in the understanding of the direction of the project could be discussed early on. Consequently, there were no communication issues identified throughout this process.

Negotiating roles and responsibilities

In initial face-to-face meetings between each supervisor and myself, the expectations of the project were discussed. We discussed modes of communication, analyses I would perform, and write up of the project. Therefore, roles were clear from the first meetings that I would be performing all analyses, with advice on statistical methods and feedback on all written drafts of the project to be provided by each of my supervisors. No issues were identified throughout the course of the project regarding roles.

Working within timelines

At the end of each meeting with my supervisors, I had set tasks to be completed prior to the next meeting. I always ensured that I completed the tasks that were assigned on time (or advised of any delays for completion of tasks), to ensure that the overall

deadline was met for completing the project. The WPP was submitted on time, demonstrating my ability to work within timelines.

Helping others to understand statistical issues – teaching

There were two scenarios in this project where I was required to communicate my findings to a non-statistical audience: 1) conveying results to my content supervisor, who is a public health physician. Despite not being a statistician, this supervisor has very good statistical knowledge, and so many concepts could be assumed. However, in communicating early findings from the project I ensured that I included a plain English description of results to reduce any confusion; 2) In communicating my findings to the public (for example, the journal of Communicable Diseases Intelligence, presentations of my findings to health professionals, the WPP report). This has required me to have a very clear understanding of the statistical techniques applied, so that I could convey a clear understanding for a non-statistical audience.

Ethical considerations

NHMRC ethics guidelines/confidentiality issues/professional responsibility

The Flutracking project was approved by the Hunter New England Research Committee (06/04/22/4.03). All data collected up to 2 October 2009 were subject to ethics approval. After 2 October 2009, the Flutracking project was incorporated into routine national influenza surveillance. Therefore, it was no longer considered a research project and ethics approval was no longer required. Although most of the data used in the logistic regression project was not monitored by the ethics committee, I was still careful to protect the participants who consented to participate in the study and ensure that all results were reported objectively, consistent with the original aims of the Flutracking project.

All data included in the analyses were stored on a secure password protected network within the HNEPH building. Only staff directly working on the Flutracking survey have access to this data. The data file containing participant details such as email addresses, usernames, age, and postcode of residence was purposely stored separately to the data file containing survey results. These two files were only linked via a common identifier variable. I was very careful when merging these two files, not to include irrelevant variables that might identify participants (such as email address and usernames). The resulting files that I produced for analyses only contained the

participant identifier variable and variables necessary for analysis. These data were only shown to my supervisors for the purposes of statistical advice.

The results in this report were communicated carefully to provide objective findings. All assumptions and limitations associated with the data to the best of my knowledge were reported.

Cover page

Project title

FLUTRACKING SURVEILLANCE: A comparison of 2007 NSW symptom rates with laboratory confirmed influenza notifications and an examination of factors affecting community receipt of H1N109 vaccination.

Location and dates

This project used data from the Flutracking surveillance system (a joint initiative from Hunter New England Population Health and the University of Newcastle). Part A of this project was undertaken from February 2008 – January 2009 and part B of this project was undertaken from March 2010 - June 2010.

Context

This project explored an application of time series analysis techniques and logistic regression techniques to surveillance data on influenza-like illness. Part A of this project was conducted as a placement as part of the NSW Health Biostatistical Officer Training Program, and the research question was devised by Dr Craig Dalton. The statistical supervisor for Part A was Dr Frank Tuyl, and the content supervisors were Dr Craig Dalton and Mr David Muscatello. Part B of this project arose as a result of completing the Categorical Data Analysis (CDA) subject, and identifying analysis techniques from CDA that may be applied to particular research questions for the Flutracking data. The statistical supervisor for Part B was Dr Patrick Kelly, and the content supervisor was Dr Craig Dalton. The main objectives of parts A and B of the study were to: 1) to measure the correlation between Flutracking symptom rates and another reliable measure of influenza; and 2) understand the influence of participant characteristics on vaccination for H1N109 pandemic influenza.

Student contribution

- Identified the research question and type of statistical analysis to be applied for part B of the project, in conjunction with content supervisor
- Prepared data for analysis, including data cleaning, merging multiple datasets, deriving applicable variables, and adding socio-economic status to the data
- Conducted all statistical analyses
- Prepared all manuscripts for submission

Statistical issues involved

- Understanding and applying ARIMA time series analysis techniques
- Identifying the type of regression model applicable to variables of analysis, and applying the logistic regression model
- Understanding and checking the assumptions for ARIMA analysis and logistic regression
- Adjusting for intraclass correlation

Student declaration

I declare this project is evidence of my own work, with direction and assistance provided by my project supervisors. This work has not been previously submitted for academic credit.

.....
Sandra Carlson

.....
Date

Supervisors' statements

Dr Patrick Kelly:

Sandra completed part A of this WPP while employed as a Biostatistical Trainee at NSW Department of Health. Although she was allowed to use this project as part of her WPP, it did not include a multivariable analysis, which is a requirement for a WPP. Hence, Sandra undertook a second analysis, which met this requirement (part B). Both parts of this project have involved analysing data from FLUTRACKING, but involve very different analyses. The work from both parts together is much more than what I would expect from a single WPP – I believe that part B is sufficient for a single WPP. However, since Sandra had already completed the work from part A, I agreed with Sandra that it was appropriate to include this as part of her WPP.

Sandra has been a conscientious student who has worked independently and has been able to learn new statistical concepts quickly.

.....
Dr Patrick Kelly

.....
Date

Dr Frank Tuyl:

As for Part A of this project, I can confirm that this is Sandra's own work. Sandra did an excellent job of keeping this project on track, learning about ARIMA time series modelling etc, while developing, for example, an excellent framework to assist the week-to-week running of the Flutracking survey.

.....
Dr Frank Tuyl

.....
Date

Introduction

The H1N109 influenza A virus that spread internationally in 2009 was a subtype of influenza that had not circulated previously in human beings. Therefore, the human immune system generally had little or no prior immunity, and it was possible that those who contracted the virus would experience more serious disease than that caused by normal seasonal influenza. In addition, transmission of this influenza subtype occurred globally. Therefore the H1N109 influenza A virus met criteria to be classified by the World Health Organisation as a 'pandemic'[1].

In Australia, at a community level, a number of surveillance systems confirmed that the number of people infected with H1N109 was comparable to 2008, and not as high as 2007 infection levels (for example, Australian Google Flu Trends data[2] and data from the Australian Sentinel Practices Research Network). Although H1N109 was a mild illness in most, in 2009 there were 4,992 people hospitalised (13% of confirmed cases) and 681 (14%) of these admitted to intensive care units[3]. There were 191 deaths due to H1N109[3]. The demographic characteristics of those seriously affected by H1N109 were different to a normal influenza season, with younger age groups being more seriously affected, and the elderly being spared[4].

To prevent a second more virulent wave of the pandemic returning, and ensure protection of those most vulnerable to the disease, a new influenza vaccine (Panvax) was made available to the Australian population for free on 30 September 2009[5].

The Flutracking surveillance system was the only Australian influenza surveillance system monitoring the uptake of the Panvax vaccine. Flutracking is a weekly online survey of influenza-like illness (ILI) completed by community members that integrates participants' ILI symptom information with their influenza vaccination status. The Flutracking surveillance system collects information on symptoms of fever, cough, and absence from work or normal duties due to fever or cough. In addition, Flutracking captures information on participant age, postcode of residence, whether or not the participant works face-to-face with patients, and seasonal influenza vaccination status. Flutracking aims to help fill the gap between laboratory and syndromal surveillance systems because it uniquely combines information on influenza symptom rates and vaccination status of participants. In 2009, as soon as the Panvax influenza vaccine was released, Flutracking began monitoring of the uptake of this vaccine in participants.

The purpose of part A of this study was to use time series methods to validate Flutracking as a method of influenza surveillance, through comparison of 2007 New South Wales (NSW) Flutracking data with NSW data for laboratory confirmed influenza. It is acknowledged that while laboratory confirmed influenza surveillance data may be biased by testing activity, it is usually considered the most reliable indicator of the onset and peak of influenza activity. Therefore, laboratory data are often used as the default measure for comparing the performance of syndromal (or 'syndromic') influenza surveillance. Zheng et al compared emergency department visits assigned a clinical diagnosis of influenza to NSW influenza laboratory data to determine whether the former could offer earlier warning of an increase in influenza incidence in the NSW population[6]. Lau et al defined the start of peak influenza activity using laboratory isolation rates for their analysis of multiple streams of influenza surveillance data[7].

The purpose of part B of this study was to use Flutracking data to investigate the relationship between receipt of the vaccination for H1N109 Pandemic influenza (vaccination with Panvax) and receipt of the seasonal influenza vaccination, as well as other demographic influences on receipt of the Panvax vaccination.

Part A of this study has been accepted and published in the journal of Communicable Diseases Intelligence[8]. An expanded methods section is included in the current report. The results and discussion sections were directly obtained from the published report.

PART A: Validating Flutracking as a method of influenza surveillance

Flutracking recruitment and data collection

A number of different methods were used to recruit participants for the Flutracking survey. An invitation to participate in the online survey was sent to approximately 7,000 email addresses on the Hunter New England area health service network with a clickable link to the survey. There was a national media release sent to all major newspapers and radio stations. A short domain name (www.flutracking.net) was used to assist the memory of people hearing or reading the recruitment messages. Emails were sent to colleagues and friends of investigators and participants were able to forward the invitation email on to acquaintances to consider joining the study. Potential participants were directed to a web page providing information about the study and an online consent form. A confirmatory email response from the participant's email address was required to complete enrolment in the study.

The study was approved by the Hunter New England Area Health Service Human Research Ethics Committee. Participants were allowed to join (or exit) at any time during the surveillance period. Further information on recruitment methods can be obtained from Dalton et al[9].

Each Monday during the typical influenza season (May/June to October each year), participants received an automatically generated weekly email link to the online survey. In the first online survey participants were asked about their usual postcode of residence; whether they work face-to-face with patients in hospitals, nursing homes, doctors' surgeries or as community health workers; their month and year of birth; and whether they received an influenza vaccination in the previous or current year.

For each subsequent survey, participants were asked whether during the prior week (ending Sunday) they had experienced fever and/or cough and/or muscle aches on any specific day/s, and whether they had been absent from usual activities on any specific day/s. Participants who reported not being vaccinated against influenza in the current season were asked if they had received vaccination in the prior week during each weekly survey. If they responded in the affirmative the question was automatically deleted from their subsequent weekly surveys, and their status was updated to ‘vaccinated’ for all following weeks.

.

Methods

At the time of analysis, data were available for 2006 – 2008. However, data from 2006 were pilot data, and there were too few participants to include in the time series analyses (394 participants completed one or more surveys). Data from 2008 were considered for analysis, however missing data between October 2007 and April 2008 posed difficulties for ARIMA analysis. Therefore, NSW data for 2007 only (for the week ending 3 June 2007 to the week ending 14 October 2007) were included in the analysis. NSW data accounted for 76% of all participants in Australia who completed at least one survey during 2007. For the purpose of this analysis, the laboratory data was classified as the independent variable, and each of the Flutracking symptoms were classified as dependent variables.

For each of the vaccinated and unvaccinated groups, a time series of the proportion of respondents reporting any of five possible case definitions was created. The case definitions were:

- fever only;
- cough only;
- absence from work or normal duties;
- fever and cough; or
- fever, cough and absence from work or normal duties.

A time series of weekly counts of positive influenza antigen tests (polymerase chain reaction and direct immunofluorescence) were created from the NSW Department of Health notifiable diseases database[10]. Unit record data were extracted from the Health Outcomes Statistical Toolkit (HOIST). Counts were aggregated into weeks based on the date of specimen collection.

We used autoregressive integrated moving average (ARIMA) time series analysis techniques and cross correlation analysis to determine whether there was an association between the laboratory time series and weekly proportions for each Flutracking case definition. The ARIMA method was chosen because, in time series modelling, the assumption that model residuals are independent is typically violated due to the residuals being autocorrelated (i.e. the current values of a series correlate with past values of the same series)[11]. If autocorrelation is not removed, then the

relationship between two time series could be overestimated[12]. Any comparisons made between laboratory data and Flutracking data potentially require correction for autocorrelation. ARIMA modelling is a well established time series analysis technique that can be used to model an autocorrelated variable[11]. Adding an independent variable to the usual ARIMA model (called transfer function analysis)[13] allows the relationship between two time series to be measured, while correcting for autocorrelation.

As the Flutracking data used for analysis were proportions, the variance stabilising transformation for binomial data was applied[14]. This is an arcsine transformation, Y_a = arcsin sqrt(Y), where Y_a is the transformed Flutracking data, and Y is the proportion of participants with the particular Flutracking symptom/s specified by each case definition. Similarly, the laboratory data were counts, and the variance stabilising transformation for a Poisson distribution was applied: X_a = sqrt(X) where X is the original laboratory data, and X_a is the transformed laboratory data[14].

For the vaccinated and unvaccinated groups, we calculated raw correlations with the laboratory notifications. These estimates were produced to compare to the estimates with autocorrelation controlled for, to determine whether there was a difference in results. We then used ARIMA models to estimate the association between weekly proportions of respondents reporting each case definition and weekly counts of positive influenza isolates. The SAS ARIMA[15] procedure was used to compute cross correlations between the two data series at various time differences, after both series had been 'prewhitened' (that is, filtered by an ARIMA model that was originally fitted to the independent variable).

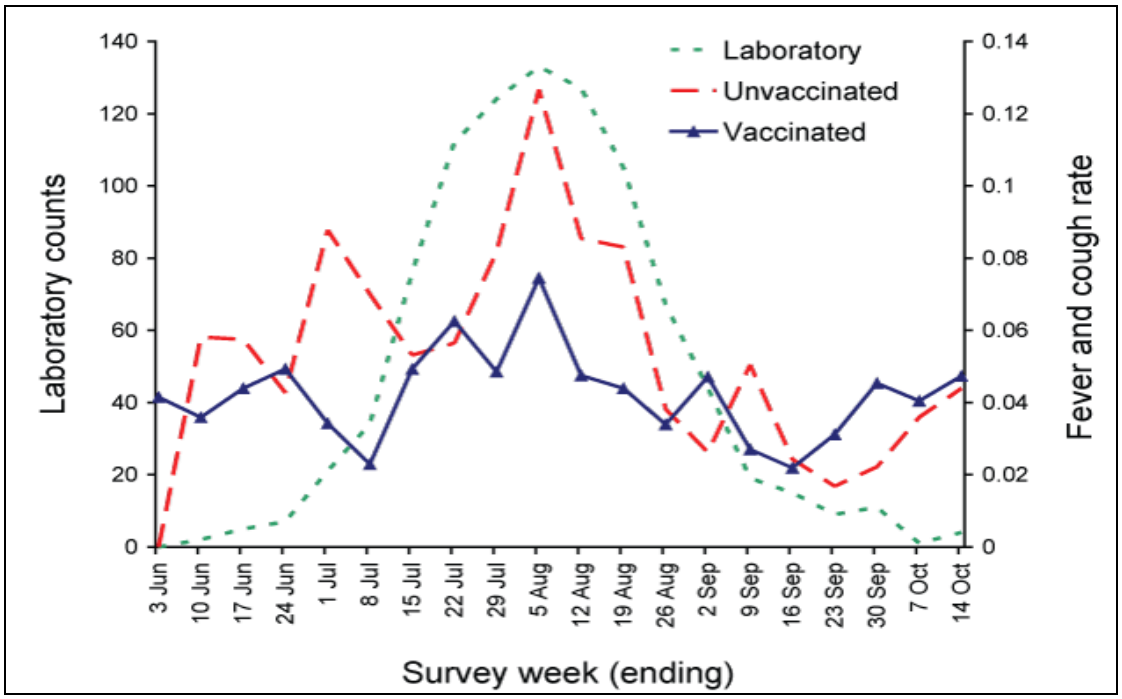
Results

Descriptive statistics

In NSW, for the 20 week period between 3 June and 14 October 2007, there was an average of 502 participants per week who completed the survey. Over that period, a weekly average of 65% of participants reported being vaccinated.

Visual inspection of the time series of each Flutracking case definition against laboratory data suggested that the peaks in laboratory data corresponded to periods of high Flutracking symptom rates for the unvaccinated group compared with the vaccinated group. A graph for the 'fever and cough' case definition is shown in Figure 1.

Figure 1: Flutracking symptom rates for 'fever and cough' case definition, compared with influenza laboratory notification counts, NSW, 2007, by influenza vaccination status.



Raw correlation analysis

Using raw correlation analysis (i.e. without autocorrelation correction), we found that the correlation values were generally highest when Flutracking symptom rates and laboratory data were compared in the same week (i.e. a lag of zero), but similar values also occurred at other differences in time (or lags).

Each Flutracking case definition in both the vaccinated and unvaccinated groups showed a statistically significant relationship with the laboratory data at a lag of zero (all *P* values for the correlation coefficients were less than 0.05). However, it was important to further analyse the relationship between the two time series using ARIMA analysis.

Autoregressive integrated moving average analysis

Results from an autocorrelation check for white noise using ARIMA analysis indicated that laboratory data showed significant autocorrelation (at the level of $P = 0.05$), and that the model that fitted this data best was $Y_t = 1.6\ Y_{t-1} - 0.6\ Y_{t-2} + e_t$ where Y_t is the laboratory data at time t (in weeks), and e are the residuals from the model. This model was used to pre-whiten both the Flutracking and laboratory data.

Cross correlations for the residuals from the ARIMA model applied to the laboratory data and each of the Flutracking data series are summarised in Table 1. Only cross correlation values at a lag of zero for each case definition related to laboratory data are reported.

Table 1: Cross correlation and corresponding probability values from the ARIMA analysis for each Flutracking case definition symptom rate compared with influenza laboratory notifications, NSW, 2007, by vaccination status.

Vaccination status	Case definition	Cross correlation value	Probability value for cross correlation (using a one-tailed t statistic)
Vaccinated	Fever	–0.006	1
Vaccinated	Cough	0.302	0.097
Vaccinated	Absence	–0.054	1
Vaccinated	Fever and cough	0.203	0.188
Vaccinated	Fever, cough and absence	–0.072	1
Unvaccinated	Fever	0.654	0.005
Unvaccinated	Cough	0.623	0.006
Unvaccinated	Absence	0.442	0.032
Unvaccinated	Fever and cough	0.640	0.005
Unvaccinated	Fever, cough and absence	0.652	0.005

In the unvaccinated group, all cross correlations at a lag of zero weeks were statistically significant at a level of $P = 0.05$. The cross correlation analysis did not provide evidence of a substantive difference between the case definitions, except for 'absence from work or normal activities,' which at 0.442, did not have as high a cross correlation as the other symptoms. In the vaccinated group no case definitions at a lag of zero were statistically significant at a level of $P = 0.05$. The results from the ARIMA analysis for the vaccinated group were not consistent with results from raw correlation analysis, where there were statistically significant relationships between every case definition for the vaccinated group and the laboratory data.

Discussion

There was a statistically significant correlation between time series of laboratory confirmed influenza and Flutracking data for unvaccinated participants in NSW for all five case definitions (fever; cough; absence; fever and cough; fever, cough and absence) at a lag of zero weeks. This indicates that Flutracking responds contemporaneously with laboratory surveillance of disease caused by influenza that leads to a specimen being collected. For the vaccinated group who should have at least some protection against influenza infection, cross correlations were not statistically significant after correction for autocorrelation, indicating that Flutracking can discriminate between influenza and other causes of ILI disease.

For vaccinated participants, the change in statistical significance between raw correlation results and ARIMA modelling results demonstrates the importance of adjusting for autocorrelation, and using appropriate analysis techniques for time series data. Without controlling for autocorrelation, spurious results were obtained. However, after correcting for autocorrelation the 'true' relationship between the two data series could be seen.

A limitation when quantifying the relationship between the Flutracking and laboratory data was that there were only 20 continuous time points in the weekly Flutracking data series, when usually at least double that number are recommended for ARIMA analysis[11]. However, we confirmed by Monte Carlo simulation that a model of the type found for the laboratory data, nearly always generates data that are clearly autocorrelated, even when there are only 20 time points, based on checking by time series analysis.

In conclusion, this analysis of Flutracking results has provided support for its value in providing alerts of influenza activity. Distinguishing between vaccinated and unvaccinated participants offers further potential to determine the value of Flutracking in assessing the effectiveness of the annual influenza vaccine composition in real-time.

Acknowledgements for publication

Sandra Carlson drafted the manuscript and performed the statistical analysis. Craig Dalton conceived and designed the Flutracking project, contributed to the statistical analysis and contributed to the manuscript. Frank Tuyl oversaw the statistical analysis and contributed to the writing of the manuscript. David Durrheim contributed to the design of the Flutracking project and writing of the manuscript. John Fejsa contributed

to the design of the Flutracking project and had primary responsibility for the online software and database development, as well as questionnaire design. David Muscatello contributed to the statistical analysis and writing of the manuscript. Lynn Francis contributed to the statistical analysis and writing of the manuscript. Edouard Tursan d'Espaignet contributed to the design of the Flutracking project, statistical analysis, and writing of the manuscript.

The authors would like to thank Robin Gilmour for providing New South Wales influenza laboratory data, and to acknowledge the University of Newcastle, NSW Department of Health, and the Hunter Medical Research Institute for their continued support. We would also like to acknowledge the thousands of Flutracking participants who give their time freely each week to contribute to influenza surveillance.

PART B: Factors associated with H1N109 vaccination

Data

Data were obtained from Hunter New England Population Health. The primary dataset used was October - December 2009 national Flutracking data. Only data from October onwards were included for analysis as this is when the Panvax vaccine was made publicly available to persons 10 years of age and over. The Panvax vaccine was made available to children 6 months to 9 years of age on 3 December 2009[16]. Due to the late availability of the vaccine to children under the age of 10 years, the data included for analysis in part B were restricted to participants 10 years of age or older.

In addition to the 2009 Flutracking data, the 2007 and 2008 datasets were also accessed to obtain data on years of participation in the Flutracking survey for those who participated from October to December 2009. Table 2 shows the relevant variables from the 2009 Flutracking data files. Note that not all variables are listed from each of the 2009 data files – only those relevant to the analysis are listed.

Table 2. Variables used for analysis from Flutracking 2009 data.

Variables used	Format	Description
Participant file (.xls file)		
ParticipantID	Numeric (up to 4 digits)	Unique identifier for each participant in the survey (retained over each year of participation)
BirthMonth	Numeric (mm)	Month participant was born
BirthYear	Numeric (yyyy)	Year participant was born
Postcode	Numeric(4 digit)	Postcode of residence of participant
WorkWithPatients	Numeric (1,2,3)	Whether participant works face-to-face with patients (1 = yes, 2 = no, 3 = don't know)
MasterRecordID	Numeric (up to 4 digits)	Identifier for whether participant responded for themselves or on behalf of other household members (0 recorded if responding for self, otherwise ParticipantID of the person responding on behalf of household member is recorded)
Survey file (.xls file)		
ParticipantID	Numeric (up to 4 digits)	Unique identifier for each participant in the survey (retained over each year of participation)
H1N1	Numeric (1,2,3)	Whether received Panvax (H1N109 vaccination) (1 = yes, 2 = no, 3 = don't know)
SurveyWeek	Numeric (ddmmyyyy)	Week beginning date for survey reference period - this was not in date format
Seasonal vaccination file (.xls file)		
ParticipantID	Numeric (up to 4 digits)	Unique identifier for each participant in the survey (retained over each year of participation)
VaccinationYear	Numeric (yyyy)	Year of vaccination with seasonal influenza vaccine
Vaccinated	Numeric (1,2,3)	Whether received seasonal influenza vaccination (1 = yes, 2 = no, 3 = don't know)
RecordEntryDate	Date time (dd/mm/yyyy hh:mm)	Date participant changed vaccination status - participants could change vaccination status throughout the year, therefore multiple records existed in this file per participant and year)

The procedure for obtaining the Flutracking data to analyse in part B follows:

1) The participant, survey and vaccination files were imported into SAS version 9.2. Data files were checked for duplicate records to ensure that there was a

unique record for each participant identification number and survey week. There were no duplicate records identified.

2) An age variable was created on the participant file using the formula: =2009 minus year of birth. There was one participant with an implausible age value of 2009 – this record was deleted from the file.

3) The three data files (Participant file, Surveyfile, and Seasonal vaccination file) were merged in SAS (by Participant ID) to create one file with ParticipantID and SurveyWeek as the unique identifiers. Each participant had multiple records – one record for each survey completed from October to December 2009.

4) Participants with a seasonal vaccination status or Panvax (H1N109) vaccination status of 'don't know' were excluded from the data.

5) SurveyWeek was converted from week beginning to week ending and to a date format (dd/mm/yy) to allow the last survey completed per participant to be easily identified and retained.

6) This data file was appended to Flutracking 2007 and 2008 data files (the 2007 and 2008 data files were prepared in a similar manner to the 2009 data – however some variables were not available on 2007 and 2008 datafiles). As the SAS procedure to append data requires variable names to be consistent on all files, some adjustments were required to the 2007 and 2008 files: 1) Dummy variables were created for Panvax vaccination status and MasterRecordID (MasterRecordID was only unavailable in 2007). In addition, a variable 'Year' was created to indicate year of survey completion, and finally, a variable for years of participation was created. This required data manipulation, including the use of the 'proc transpose' procedure to obtain a new dataset with this variable, and then remerging this new dataset with the original dataset.

7) After a file was created containing all three years of data, records without a value for Panvax vaccination status were excluded (only records from October 2009 to December 2009 were included), and only the last survey per participant within this timeframe was kept (therefore, there was only one record per participant, as opposed to one record per participant and survey week).

8) Records for participants under the age of 10 years were excluded from the file.

9) Socio-economic status (SES) scores were added to the file. These scores were obtained from the Australian Bureau of Statistics website[17]. The index of relative socio-economic advantage and disadvantage was used to assess socio-economic status in the current analysis. In this index a lower score indicates that an area is relatively disadvantaged compared to an area with a higher score. The index for advantage and disadvantage was used rather than the index of disadvantage only because vaccination with Panvax was expected to be affected by both advantage and disadvantage.

   As the difference between scores may not be equal (for example the difference in disadvantage between scores of 500 and 600 is not equal to the difference between scores of 900 and 1000) this variable could not be used as a continuous variable in regression analysis. Therefore, it was decided to use quintile categories for analysis. Postal areas were ordered from lowest to highest score (a low score indicating the postal area was relatively disadvantaged compared to an area with a high score). Quintiles were formed from these scores with the lowest 20% of postal areas given a code of 1, and so on up to the highest 20% of postal areas given a code of 5.

   The index of relative socio-economic advantage and disadvantage for postal areas was imported into SAS as a .csv file and merged (by postcode) with the data file.

10) The MasterRecordID variable was used to create a unique household identifier, as well as a variable for the number of participants per household. This was necessary to adjust for any intraclass correlation for groups of participants in each household. Failure to adjust for any intraclass correlation, could underestimate standard errors in regression analysis, and therefore artificially overestimate any relationships found.

11) Non-essential variables were removed from the final file. All variables in the final file that were used in analyses are described in Table 3.

12) The final file was then exported as a .csv file, and this .csv file was imported into Stata 10.0 for further analysis. Stata was used for all regression analysis as this was the software package practiced most during the Master of Biostatistics course.

13) A check was performed of potential bias due to missing data for participants who were vaccinated with Panvax after dropping out of the survey. There were 2453 participants who dropped out of the survey prior to the last week of participation. Of these participants 1977 (80.6%) were not yet vaccinated with Panvax. It is possible that some of these participants may have actually been vaccinated with Panvax after dropping out of the survey, and there may be some bias in the data. However, it is possible that the participants who dropped out of the survey early may be less concerned with influenza prevention, and therefore less likely to be vaccinated with Panvax than those participants who completed the final survey in 2009.

Table 3. Variables in final data file for analysis.

Variable	**Description**
ParticipantID	Unique identifier for each participant
Yrspart	Number of years participated in the survey (ranges from 1 to 3)
WorkWithPatients	Whether participant works face-to-face with patients (1 = yes, 2 = no, 3 = don't know)
Age	Participant age in years
H1N1	Whether received Panvax vaccination (1 = yes, 2 = no)
Vax09	Whether received seasonal influenza vaccination (1 = yes, 2 = no)
Score_adv_disadv	Socio-economic status score
Ses_quintile	Socio-economic status quintile
Hhid	Unique identified for each household
hhmember	Whether participant responded for themselves (primary respondent) or had another household member respond on their behalf (other respondent) (1 = other respondent and 0 = primary respondent)
Cluster	Number of participants per household
Weekending	week that last survey of 2009 was completed (dd/mm/yy)

Methods

The relationship between Panvax vaccination status and the covariates: seasonal vaccination status; whether works face-to-face with patients; age; socio-economic status; and number of years participated in the survey was assessed using binary logistic regression. Binary logistic regression was chosen because the explanatory variable (Panvax vaccination status) had values of 0 and 1 only, and so followed a binary distribution. From the Categorical Data Analysis (CDA) subject, binary logistic

regression is suitable, widely used, and easy to implement in common statistical packages such as SAS and Stata.

The logistic regression model is a type of generalised linear model that relates a linear regression model to a response/dependent variable via the logit link function ($logit(p) = ln\left(\frac{p}{1-p}\right)$). Binary logistic regression is used when the outcome variable is dichotomous. The equation used to describe binary logistic regression is: $logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$. The independent variables in the model can be continuous or categorical.

The dependent variable in the binary logistic regression model for the current analysis was Panvax vaccination status (H1N1). The independent variables were: Years of participation (yrspart); whether participant worked face-to-face with patients (workwithpatients); age (age); seasonal vaccination status (vax09); and socio-economic status quintile (ses_quintile).

All independent variables were categorical variables, except for age. To allow easy interpretation of results, age was categorised into five groups: 10-19 years, 20-34 years, 35-49 years, 50-64 years, and 65 years and over. These categories were chosen based on ages more or less likely to get immunised against influenza.

The equation used to describe the logistic regression model was: $logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{14} x_{14}$

where:

the outcome variable (H1N1 vaccination status) has two categories

p is the probability of being vaccinated with Panvax

β represents the parameters in the model

$$x_1 = \begin{cases} 1 \; for \; yrspart \; code \; 2 \\ 0 \; for \; otherwise \end{cases}$$

$$x_2 = \begin{cases} 1 \; for \; yrspart \; code \; 3 \\ 0 \; for \; otherwise \end{cases}$$

$$x_3 = \begin{cases} 1 \; for \; workswithpatients \; code \; 1 \\ 0 \; for \; otherwise \end{cases}$$

$$x_4 = \begin{cases} 1 \; for \; workswithpatients \; code \; 3 \\ 0 \; for \; otherwise \end{cases}$$

$$x_5 = \begin{cases} 1\ for\ age\ code\ 10-19\ years \\ 0\ for\ otherwise \end{cases}$$

$$x_6 = \begin{cases} 1\ for\ age\ code\ 20-34\ years \\ 0\ for\ otherwise \end{cases}$$

$$x_7 = \begin{cases} 1\ for\ age\ code\ 35-49\ years \\ 0\ for\ otherwise \end{cases}$$

$$x_8 = \begin{cases} 1\ for\ age\ code\ 50-64\ years \\ 0\ for\ otherwise \end{cases}$$

$$x_9 = \begin{cases} 1\ for\ age\ code\ 65\ years\ and\ over \\ 0\ for\ otherwise \end{cases}$$

$$x_{10} = \begin{cases} 1\ for\ seasonal\ vaccination\ status\ code\ 1 \\ 0\ for\ otherwise \end{cases}$$

$$x_{11} = \begin{cases} 1\ for\ socio-economic\ status\ quintile\ 2 \\ 0\ for\ otherwise \end{cases}$$

$$x_{12} = \begin{cases} 1\ for\ socio-economic\ status\ quintile\ 3 \\ 0\ for\ otherwise \end{cases}$$

$$x_{13} = \begin{cases} 1\ for\ socio-economic\ status\ quintile\ 4 \\ 0\ for\ otherwise \end{cases}$$

$$x_{14} = \begin{cases} 1\ for\ socio-economic\ status\ quintile\ 5 \\ 0\ for\ otherwise \end{cases}$$

Before fitting the multivariate logistic regression model, univariate statistics were produced for each individual covariate as well as the dependent variable (Panvax vaccination status). The purpose of the univariate analyses was to describe the data. Tables of counts of participants in each covariate category were produced, as well as the percentage of participants vaccinated with Panvax in each covariate category.

Univariate logistic regression analyses were performed to show the relationship between each covariate and the dependent variable (vaccination with Panvax). However, measuring relationships between independent and dependent variables in univariate analysis cannot control for the effects of other covariates, or provide any information on the relationships between covariates. Therefore, multivariate logistic regression analysis was also performed.

In the Flutracking survey a participant can respond on behalf of other household members, therefore there may be potential positive correlation between observations from the same household. This correlation needs to be taken into account in order to obtain correct variances of regression coefficients[18] and avoid incorrect statistical inferences. In the current analysis, this was achieved by empirically adjusting the standard errors using the sandwich estimator in the logistic regression models. This was achieved using the Stata option 'VCE(cluster *variable*)' for the 'logit' command.

The goodness of fit of the model was assessed overall. The Hosmer Lemeshow test for goodness of fit was chosen because it has the advantage over the chi squared goodness of fit test of being able to split observations into groups of approximately equal size, so there are less likely to be groups with low observed and expected frequencies[19].

The following assumptions for the final logistic regression model were also checked:

1) The outcome variable (Panvax vaccination status) was checked to ensure it was linearly related to the log odds of the combination of the dependent variables.

2) Outliers/influential observations were assessed.

3) Multicollinearity between covariates was assessed (covariates should be independent from one another). Multicollinearity occurs when two or more variables in the model are approximately equivalent to a linear combination of other variables in the model (from LMR course notes). When there is perfect collinearity, Stata notifies this in the model. To measure less severe multicollinearity, the Stata command 'collin' was used. The VIF (variance inflation factor) was used to assess the level of multicollinearity in the model.

Results

Descriptive statistics

There were 7145 participants greater than or equal to 10 years of age who completed at least one Flutracking survey in 2009 and responded to the survey question asking about Panvax vaccination status. Table 4 provides summary statistics for these participants.

Table 4 shows that just over one third of participants (33.1%) were vaccinated against H1N109. Of those participants who were vaccinated against H1N109, 82.1% were vaccinated in either October or November.

More than half of participants were vaccinated with the seasonal vaccine (60.9%) and, of these participants, almost half (47.0%) were vaccinated against H1N09. Whereas only 11.4% of those who did not receive the seasonal vaccine, received the Panvax vaccine.

There were 49.0% of participants who participated in Flutracking for more than one year (completed at least one survey in 2009, as well as either 2007, 2008 or both years). As the number of years of participation increased, the percent of participants vaccinated with Panvax also increased (from 30.5% to 40.6%).

As age increased, the percent of participants vaccinated with Panvax also increased from 12.7% in the 10-19 years group to 55.2% in the 65 years and over age group.

There were 27.6% of participants who worked face-to-face with patients. Of these participants, less than half (44.3%) were vaccinated against H1N109. However, the percent of participants vaccinated against H1N109 was higher in this group than those who did not work face-to-face with patients (28.7%).

The number of participants in each socio-economic status quintile increased for each increase in quintile. However, there did not appear to be any difference in vaccination uptake between socio-economic status quintiles, with approximately one third of participants being vaccinated in each quintile.

More than two thirds of participants (73.4%) were primary respondents (participants who responded for themselves, and may or may not have responded for other participants in their household). There was a higher proportion of primary respondents vaccinated against H1N109 (37.4%), as compared to other household members in the survey (21.2%). However, the average age of a primary respondent was 46.4 years (s.d. = 11.4) and the average age of other household members was 36.5 years (s.d. = 19.0). Therefore, this higher vaccination uptake rate may be confounded by age.

Just under half of participants (44.8%) had more than one respondent per household. As household size for participation increased (from 1 to 8), percent vaccinated against

H1N109 generally decreased. However, this result may also be confounded by age – in general, age decreased as size of household increased (see Figure 2).

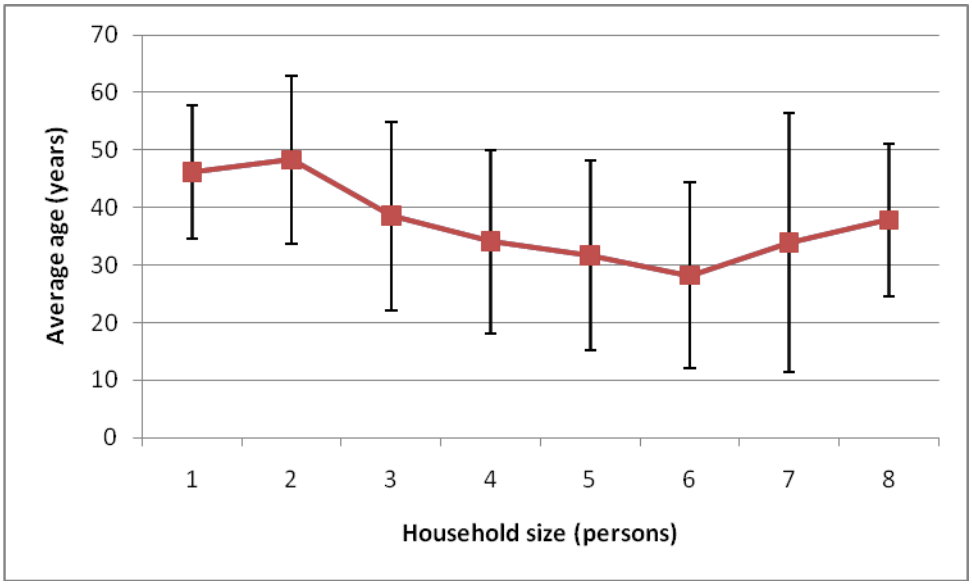
Most participants last completed a survey in December 2009 (80.2%).

Table 4. Description of participant characteristics, Flutracking data, Australia, October – December 2009.

Variable	Categories	Total	Number vaccinated with Panvax	% vaccinated with Panvax
Overall	N/A	7145	2363	33.1
Seasonal vaccination status	Vaccinated	4349	2044	47.0
	Unvaccinated	2796	319	11.4
Years of participation	1	3642	1109	30.5
	2	2910	1013	34.8
	3	593	241	40.6
Age (years)	10 – 19	584	74	12.7
	20 – 34	1214	333	27.4
	35 – 49	2496	795	31.9
	50 – 64	2534	986	38.9
	65+	317	175	55.2
Working face-to-face with patients	Yes	1974	875	44.3
	No	5149	1480	28.7
	Don't know	22	8	36.4
Socio-economic status	Quintile 1 (lowest quintile)	632	227	35.9
	Quintile 2	838	277	33.1
	Quintile 3	1557	529	34.0
	Quintile 4	1888	621	32.9
	Quintile 5 (highest quintile)	2230	709	31.8

NA=Not Applicable

Figure 2. Average age in years (and standard deviation) for each household size.



Logistic regression

All variables included in the univariate and multivariable logistic regression analyses were classified as categorical variables. Table 5 shows the results from the univariate logistic regression models and Table 6 shows the results from the multivariate model. Both tables show the standard errors, unadjusted and adjusted for clustering within household. As expected, the adjusted standard errors are slightly larger.

Table 5 shows that (without controlling for any other covariates) having the seasonal influenza vaccine was strongly related to receiving the Panvax vaccine, with the odds of receiving the Panvax vaccine increasing by 589% for those who received the seasonal vaccine, as compared to those who did not.

The number of years a participant participated in the survey also showed a statistically significant relationship ($p < 0.001$) with uptake of the Panvax vaccine (without controlling for any other covariates). For each increase in the number of years participating in the Flutracking survey, the odds of receiving the Panvax vaccine increased.

Age of participants was also strongly related to receiving the Panvax vaccine. For each increase in age group, the odds of receiving the Panvax vaccine increased (without controlling for any other covariates). In particular, the odds of receiving the Panvax vaccine increasing by 749% for those participants aged 65 years and over compared to those in the 10-19 years age group.

Working face-to-face with patients was also strongly related to receiving the Panvax vaccine, with the odds of receiving the vaccine increasing by 97% for those who worked face-to-face with patients compared to those who did not work face-to-face with patients.

Only socio-economic status did not show a statistically significant relationship with receipt of the Panvax vaccine in the univariate analyses (p = 0.45). The odds of receiving the Panvax vaccine were not very different between each quintile.

Table 5. Results from univariate model.

Covariate	Reference level	Level	Odds ratio	SE (without adjusting for intracluster correlation)	SE	Lower CI	Upper CI	Overall P-value
Seasonal vaccination status	Unvaccinated	Vaccinated	6.89	0.46	0.51	5.95	7.97	<0.001
Years of participation	1	2	1.22	0.06	0.07	1.08	1.37	<0.001
		3	1.56	0.14	0.14	1.30	1.88	
Age (years)	10-19	20-34	2.60	0.36	0.44	1.87	3.62	<0.001
		35-49	3.22	0.42	0.50	2.38	4.36	
		50-64	4.39	0.57	0.70	3.21	6.00	
		65+	8.49	1.43	1.70	5.74	12.58	
Working face-to-face with patients	No	Yes	1.97	0.11	0.11	1.77	2.21	<0.001
		Don't know	1.42	0.63	0.63	0.59	3.38	
Socio-economic status	Quintile 1	Quintile 2	0.88	0.10	0.11	0.70	1.11	0.45
		Quintile 3	0.92	0.09	0.10	0.74	1.14	
		Quintile 4	0.87	0.08	0.09	0.71	1.08	
		Quintile 5	0.83	0.08	0.09	0.68	1.02	

From Table 6, after adjusting for age, working with patients status, socio-economic status and years of participation in the survey, as well as adjusting for the effect of some participants living in the same household and responding for other participants in that household, having the seasonal influenza vaccine was still strongly related to receiving the Panvax vaccine, with the odds of receiving the Panvax vaccine increasing

by 489% for those who received the seasonal vaccine, as compared to those who did not.

The number of years a participant participated in the survey still showed a statistically significant relationship ($p = 0.02$) with uptake of the Panvax vaccine in the multivariate model, however this relationship was not as strong as shown in the univariate analysis.

Age of participants also showed a statistically significant relationship with receiving the Panvax vaccine ($p<0.001$). However, this relationship was not as strong as shown in the univariate analysis. For each increase in age group, the odds of receiving the Panvax vaccine increased. In particular, the odds of receiving the Panvax vaccine increased by 200% for those participants aged 65 years and over compared to those in the 10-19 years age group.

Working face-to-face with patients also showed a statistically significant relationship with receiving the Panvax vaccine ($p<0.001$) in the multivariate model, with the odds of receiving the vaccine increasing by 48% for those who worked face-to-face with patients compared to those who did not work face-to-face with patients.

Socio-economic status now showed an even weaker relationship with receipt of the Panvax vaccine in the multivariate model ($p = 0.82$), as compared to the univariate model. The odds of receiving the Panvax vaccine were not very different between each quintile.

Table 6. Results from multivariate model.

Covariate	Reference level	Level	Odds ratio	SE (without adjusting for intracluster correlation)	SE	Lower CI	Upper CI	Overall P-value
Seasonal vaccination status	Unvaccinated	Vaccinated	5.89	0.41	0.45	5.07	6.84	<0.001
Years of participation	1	2	1.12	0.07	0.07	0.98	1.27	0.02
		3	1.31	0.13	0.13	1.08	1.60	
Age (years)	10-19	20-34	1.24	0.19	0.22	0.87	1.75	<0.001
		35-49	1.30	0.19	0.21	0.95	1.79	
		50-64	1.70	0.24	0.28	1.22	2.36	
		65+	3.00	0.54	0.62	2.00	4.51	
Working face-to-face with patients	No	Yes	1.48	0.09	0.09	1.31	1.67	<0.001
		Don't know	1.46	0.70	0.69	0.57	3.70	
Socio-economic status	Quintile 1	Quintile 2	0.92	0.11	0.12	0.72	1.19	0.82
		Quintile 3	0.95	0.10	0.11	0.75	1.20	
		Quintile 4	0.96	0.10	0.11	0.77	1.20	
		Quintile 5	0.89	0.09	0.10	0.71	1.11	

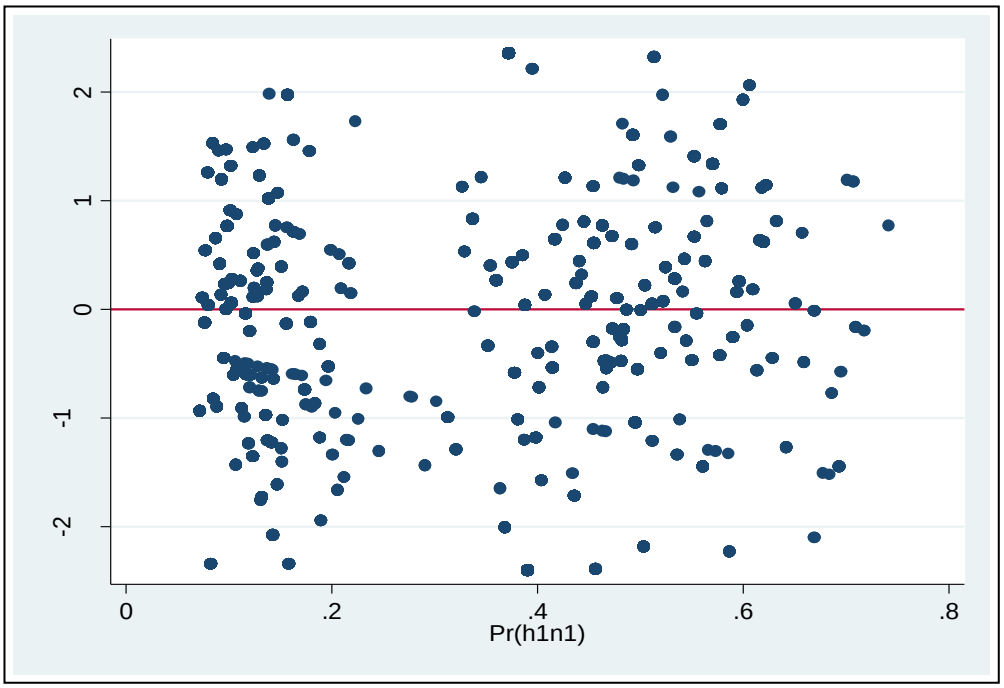
Goodness of fit and testing of model assumptions

According to the Hosmer-Lemeshow goodness of fit test, the multivariate model provided an adequate fit to the data ($\chi^2(8) = 9.71, p = 0.29$).

Pearson and deviance residuals are useful in identifying observations that are not explained well by the model[20]. According to the plot of standardised deviance residuals in Figure 3, there do not appear to be any outliers in the data affecting the fit of the model (any observations outside a value of +/- 2 or 3 would signal potential outliers requiring further investigation). The plot of Pearson residuals also did not reveal any potential outliers. Figure 3 also shows that there appears to be a linear relationship between the predicted values and the covariates of interest, and constant variance, as the plots do not show any obvious trends or curvature.

Please note that for Figures 3 - 5, each point represents multiple observations.

Figure 3. Scatter plot of deviance residuals against predicted values for H1N1 from final multivariate logistic regression model.



An observation was identified as having high leverage (data points had a large influence on the regression results) if it had a leverage value larger than 0.5. Moderate leverage was defined as any value between 0.2 and 0.5 (from LMR notes). Figure 4 shows that no participants had very high leverage.

Figure 4. Scatter plot of leverage against predicted values for H1N1 from final multivariate logistic regression model.

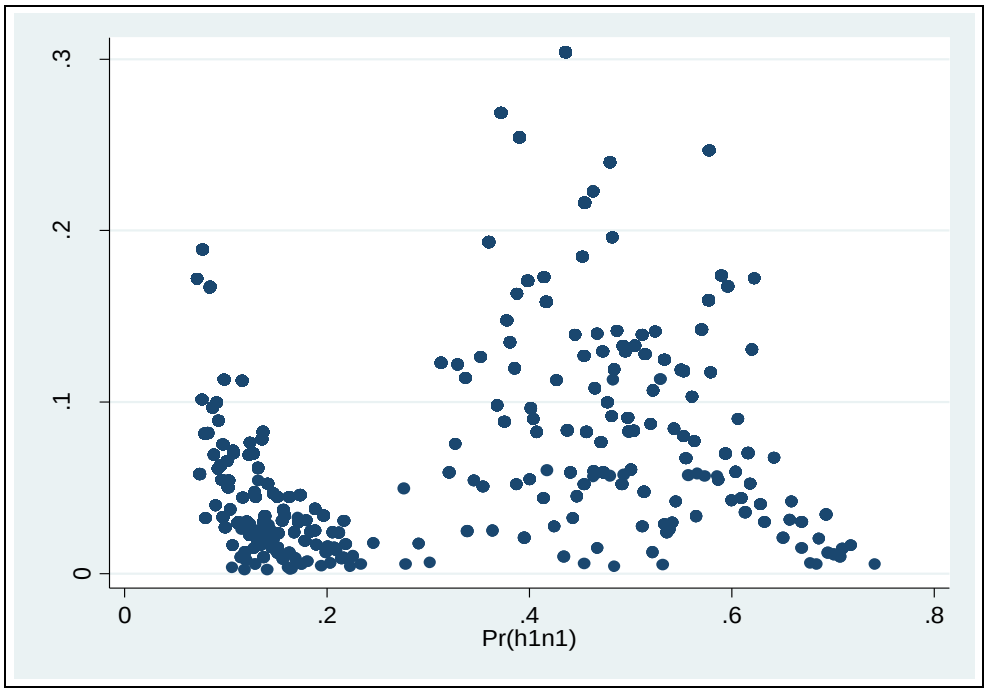
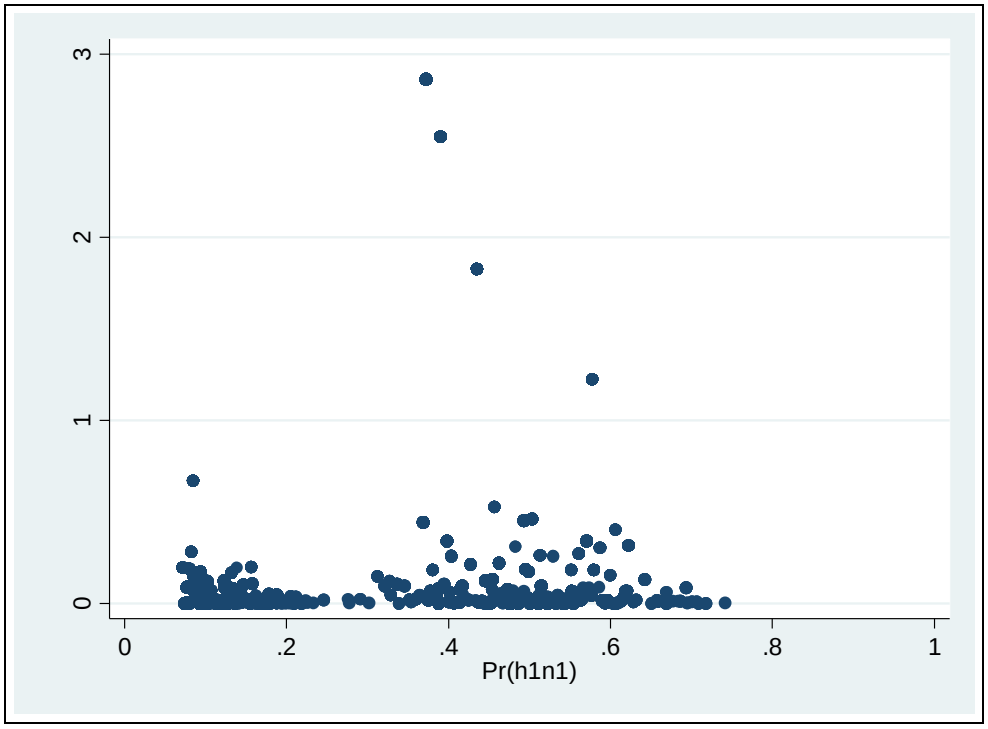


Figure 5 shows the Pregibon Delta-Beta influence statistic. This statistic indicates how much a regression coefficient would change if an observation was omitted (LMR course notes). From Figure 5 there were four points that had much larger Delta-Beta values than all other values, requiring further investigation. However, these four points in Figure 5 actually represented 653 observations. The largest Delta-Beta value of 2.86 was shared by 203 observations. These participants had high values for the Standardised Pearson residuals, deviance residuals, and leverage values. On further investigation, all of these participants participated in the survey for one year, were vaccinated with the seasonal influenza vaccine, were aged 49 – 64 years, did not work face-to-face with patients, and were in the highest quintile for socio-economic status. As these data values were plausible, and there were a large number of participants with these same values, it did not warrant further investigation.

Figure 5. Scatter plot of the Pregibon Delta-Beta influence statistic against predicted values for H1N1.



Using Kleinbaum, Kuppler, and Mullers (1988) guidelines[21], any variable with a VIF value over 10 was considered to exhibit a level of multicollinearity to be concerned about, and would warrant further investigation. There was no evidence of multicollinearity found in the variables included in the model (see Table 7).

Table 7. Results from multicollinearity check for each covariate in final logistic regression model.

Covariate	Reference level	Level	VIF value
Seasonal vaccination status	Unvaccinated	Vaccinated	1.13
Years of participation	1	2	1.10
		3	1.08
Age (years)	10-19	20-34	2.69
		35-49	3.75
		50-64	3.81
		65+	1.56
Working face-to-face with patients	No	Yes	1.10
		Don't know	1.00
Socio-economic status	Quintile 1	Quintile 2	2.06
		Quintile 3	2.73
		Quintile 4	2.99
		Quintile 5	3.20

No further model fitting was conducted as the model fit was adequate and all model assumptions appeared to be satisfied. As only socio-economic status was found to not be statistically significant, and this variable was of clinical relevance, it was retained in the model.

Discussion

The odds of receiving the Panvax vaccine increased by 489% for those who received the seasonal vaccine, as compared to those who did not, suggesting a common attitude to receipt of both vaccines. The seasonal influenza vaccine for 2010 contains the H1N109 influenza strain. Experts have predicted a return of H1N109 to Australia as the dominant strain of influenza during the 2010 influenza season. Therefore, these results imply that participants who have not been vaccinated with Panvax, are also unlikely to be vaccinated with the seasonal vaccine and, therefore, have no protection against a return of H1N109.

The older a Flutracking participant was, the more likely they were to receive the Panvax vaccination. It is now well known that H1N109 affected younger age groups more severely than older age groups, and protection against this strain of influenza in the young healthy population is therefore crucial. In Australia, the National Immunisation Program Schedule[22] (current from 1 July 2007) recommends the seasonal influenza vaccination for all persons aged 65 years and over. Therefore, persons over the age of 65 years are used to receiving a regular influenza vaccine (and

perhaps maybe more trusting of the safety of the vaccine or simply in the routine of receiving it). Younger healthy adults may not perceive themselves to be at great risk as they are generally healthier and have stronger immune systems than older age groups, or perhaps they are less trusting of the safety of the vaccine. These factors suggest that changes in attitudes/beliefs of the younger age groups is necessary to ensure sufficient uptake for herd immunity against H1N109.

Participants who worked face-to-face with patients were also more likely to receive the Panvax vaccination than those who did not work face-to-face with patients. This finding is reassuring, however, 56% of Flutracking participants who work face-to-face with patients were not yet vaccinated against H1N109. Barriers to receipt of the vaccine should be further explored with this group to ensure the safety of themselves and the potentially vulnerable patients they are treating.

As years of participation in the Flutracking survey increased, participants were more likely to receive the Panvax vaccination. This relationship (although statistically significant) was one of the weakest relationships in the model. However, this highlights an important characteristic of Flutracking participants - that loyalty of participants to the Flutracking survey may be an indicator of support for influenza vaccination.

Receipt of the Panvax vaccination was similar across each socio-economic status quintile. Given that the Panvax vaccination was offered free of charge to all Australians, this result is not unexpected.

The above results help to provide a better understanding of the influences in Australia on uptake of the influenza vaccination. This is important as vaccination campaigns are still encouraging Australians to receive the Panvax vaccination if they cannot/ choose not to receive the seasonal vaccine. Given the likelihood of a return of the H1N109 strain of influenza in 2010, it is crucial to understand attitudinal barriers and demographic differences in likelihood of vaccination.

There were several limitations identified in this study. Firstly, The Flutracking dataset may be viewed as a hierarchical dataset with potentially three levels: person level, household level, and geographic level (for example, postcode). Therefore, there may be positive correlation between observations from the same household and positive correlation between observations from the same postcode. Three approaches can be used to adjust for intracluster correlations: 1) fit a marginal model and empirically adjusting the standard errors for clustering; 2) apply a multilevel mixed effects logistic regression model, fitting a random effect or effects for clustering; or 3) fit a fixed effects

model by including the clustering variable or variables as covariates in the model. In the analysis presented, the model included a fixed effect variable for socio-economic status and empirically adjusted the standard errors for clustering by household. The socio-economic status variable was created assuming that each individual in a postcode had the same level of socio-economic status. It would be preferable to have a measure of socio-economic status at the individual level, or alternatively to consider postcode as either a fixed or random effect.

If a multilevel mixed effects model were applied, with a random effect for household and a random effect for postcode, then intracluster correlation could be calculated for both the household level and geographic level. Although not reported in the results section, a multilevel mixed effects model was fitted to the data. Assumption checking showed that the deviance residuals followed a bimodal distribution (rather than a normal distribution). The meaning of this result was not understood well enough to be confident in the model fit.

Another limitation with this study was that the results may be subject to missing data bias. We assumed that any missing data was random. However, participants who stopped completing surveys in October or November may not have had an opportunity to respond 'yes' to the Panvax vaccination. Alternatively, these participants may be less likely to be vaccinated than other participants, as they are less interested in influenza.

A further limitation with this study is that the Flutracking survey was not designed specifically for the research question at hand. Therefore, there may be other variables influencing the relationships shown in the model that have not been captured in the Flutracking data (for example, gender).

Interactions between covariates were not assessed in this study, as they were not part of the research question at hand. This may be an area of interest for future research (for example, the interaction between age and seasonal vaccination status, and this effect on Panvax vaccination uptake).

Additional research questions have been identified as a result of the current analysis. These include: 1) Assessing factors affecting the uptake of seasonal vaccination in Flutracking participants; 2) Assessing factors affecting participant retention Flutracking over years of participation; and 3) Assessing whether working face-to-face with patients is predictive of influenza-like illness.

References

1. World Health Organisation. What is a pandemic? [cited 9 June 2010]; Available from: http://www.who.int/csr/disease/swineflu/frequently_asked_questions/pandemic/en/index.html.

2. Google Flu Trends. Explore flu trends – Australia [cited 9 June 2010]; Available from: http://www.google.org/flutrends/au/#AU.

3. Department of Health and Ageing. Australian influenza report 2010. Report 18, week ending 7 May 2010. [cited 9 June 2010]; Available from: http://www.health.gov.au/internet/main/publishing.nsf/Content/cda-ozflu-no18-10.htm.

4. Bishop JF, Murnane MP, Owen R. Australia's winter with the 2009 pandemic influenza A (H1N1) virus. *NEngl J Med 2009: 361(27): 2591 – 2594.*

5. Department of Health and Ageing. H1N1 influenza latest news: Free pandemic flu vaccine available for all. [cited 9 June 2010]; Available from: http://www.healthemergency.gov.au/internet/healthemergency/publishing.nsf/Content/news-300909.

6. Zheng W, Aitken R, Muscatello DJ, Churches T. Potential for early warning of viral influenza activity in the community by monitoring clinical diagnoses of influenza in hospital emergency departments. *BMC Public Health* 2007;7: 250.

7. Lau EH, Cowling BJ, Ho LM, Leung GM. Optimizing use of multistream influenza sentinel surveillance data. *Emerg Infect Dis* 2008 Jul [Online] [cited 21 October 2008]; Available from: http://www.cdc.gov/EID/content/14/7/1154.htm.

8. Carlson SJ, Dalton CB, Tuyl FA, Durrheim DN, Fejsa J, Muscatello DJ, et al. Flutracking surveillance: Comparing 2007 New South Wales results with laboratory confirmed influenza notifications. Commun Dis Intell 2009;33(3): 323–326.

9. Dalton C, Durrheim D, Fejsa J, Francis L, Carlson S, Tursan d'Espaignet E, et al. Flutracking: A weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008. Commun Dis Intell 2009;33(3):316–322.

10. NSW Department of Health. Notifiable Diseases Database (HOIST), Centre for Epidemiology and Research.

11. Pankratz A. Forecasting with dynamic regression models. New York: John Wiley & Sons Inc; 1991.

12. Schwartz J, Spix C, Touloumi G, Bachárová L, Barumamdzadeh T, le Tertre A, et al. Methodological issues in studies of air pollution and daily counts of death or hospital admissions. *J Epidemiol Community Health* 1996;50 (Suppl 1): S3–S11.

13. Box GEP, Jenkins GM, Reinsal GC. Time series analysis: Forecasting and control. New Jersey: Prentice Hall; 1994.

14. Fisher LD, van Belle G. Biostatistics: A methodology for the health sciences. New York: John Wiley & Sons Inc; 1993.

15. SAS 9.1.3 Help and Documentation. Cary (USA): SAS Institute; 2002–2004.

16. Department of Health and Ageing. H1N1 influenza latest news: Pandemic flu vaccine approved for children. [cited 9 June 2010]; Available from: http://www.healthemergency.gov.au/internet/healthemergency/publishing.nsf/Content/news-0322009.

17. Australian Bureau of Statistics. 2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia - Data only, 2006. [cited 9 June 2010]; Available from: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2033.0.55.0012006?OpenDocument.

18. Graubard BI, Korn, EL. Regression analysis with clustered data. *Statistics in Medicine. 1994: 13: 509-522.*

19. Bewick V, Cheek L, Ball, J. Statistics review 14: Logistic regression. *Critical Care. 2005: 9(1): 112-118.*

20. SAS/STAT® 9.2 User's Guide, Second Edition. [cited 25 June 2010]; Available from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#/documentation/cdl/en/statug/63033/HTML/default/statug_logistic_sect042.htm.

21. Kleinbaum D.G, Kupper L.L, Muller K.E, *Applied regression analysis and other multivariate methods (2nd edition),* PWS-KENT Publishing, 1988. Pp 206-217.

22. Department of Health and Ageing. National immunisation program schedule. [cited 28 June 2010]; Available from http://www.immunise.health.gov.au/internet/immunise/publishing.nsf/Content/nips2.