Evaluating the Use of Physician Billing Data for Age and Setting Specific Influenza Surveillance

Emily Chan

Master of Science

Epidemiology, Biostatistics and Occupational Health

McGill University

Montreal, Quebec

January 2009

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Master of Science

ABSTRACT

Syndromic surveillance has emerged as a novel, automated approach to monitoring diseases using pre-diagnostic but often non-specific data sources. However, there is little consensus about the best data sources. Using physician billing data from community-based care settings and emergency departments in Quebec, Canada during 1998-2003, we evaluated the lead-lag relationship between ambulatory medical visits for influenza-like illnesses (ILI) and pneumonia and influenza (P&I) hospitalizations by age-group, visit setting, and influenza season. To do so, we applied ARIMA modeling methodology and computed the cross-correlation function (CCF) using the residuals. ILI visits in community settings by children aged 5-17 years tended to provide the greatest lead times (at least 2 but up to 3 weeks) over P&I hospitalizations. Lead times varied each season, possibly due to the circulation of different strains each season. These findings have important implications for syndromic surveillance of influenza, as well as epidemic control strategies such as vaccination and school closure policies.

ABRÉGÉ

La surveillance syndromique a émergé comme une nouvelle approche automatisée pour le contrôle des maladies avec des sources de données pré-diagnostic, mais qui sont souvent non-spécifiques. Pourtant, il y a peu de consensus concernant les meilleures sources de données. En utilisant des factures médicales émises entre 1998 et 2003, et provenant de centres communautaire et de services d'urgence au Québec, Canada, nous avons évalué par tranche d'âge, le cadre des visites, et la saison de la grippe la relation d'avance-décalage entre les visites médicales ambulatoires pour le syndrome d'allure grippale (SAG) et les hospitalisations pour la pneumonie et la grippe. Pour ce faire, nous avons appliqué la méthodologie des modèles d'ARIMA et calculé la fonction de contre-corrélation (CCF) avec les résidus. Les visites communautaires reliée au SAG par des enfants âgés de 5-17 ans ont eu tendance à pourvoir les plus grandes avances (au moins 2 semaines, mais quelques fois jusqu'à 3 semaines) contre des hospitalisations pour la pneumonie et la grippe. Les avances ont varié chaque année, peut-être à cause de la circulation des souches différentes chaque saison. Ces résultats ont des implications importantes pour la surveillance syndromique de la grippe, ainsi que pour des stratégies de lutte contre l'épidémie, comme la vaccination et la fermeture d'écoles.

CONTRIBUTIONS OF AUTHORS

As first author, I contributed to the planning of the study design, conducted all data processing and statistical analyses, interpreted the results, and wrote both manuscripts. David Buckeridge, as my primary supervisor, conceived of the study, and was involved in supervising all stages of the research, from fine-tuning the study design and assisting with the implementation of the study to providing thoughtful discussion on the results and making editorial revisions of the manuscripts. Robyn Tamblyn, as my co-supervisor, provided helpful feedback on the study design and interpretation of the results, and made editorial revisions of the manuscripts.

ACKNOWLEDGEMENTS

As I near the end of my Masters degree, there are a several individuals I would like to thank for their assistance and encouragement along the way. First and foremost, I extend my sincerest thanks to my primary supervisor, David Buckeridge, who took me in when I was a rookie student and knew nothing about the field, and enthusiastically and patiently guided me as I learned both the joys and pains of the process of conducting research, and who moulded for me an extraordinary Masters project, allowing me to mine a rich data set at which I am still in awe at. I also hold great gratitude for my co-supervisor, Robyn Tamblyn, who never failed to provide insightful discussion and feedback on my project, kept us on track, and served as a true mentor. To both my supervisors, I would like to thank them for their patience, for providing stimulating and interesting challenges, for sharing their remarkable expertise and wisdom from which I have learned so much in the past year, and for their encouragement, in particular their heroic and amusing efforts in their never-ending quest to push me where I've been reluctant. I am amazed by their creativity, energy, and constant thirst for new routes of investigation (even in the hours before I submit my thesis) which have been an inspirational reminder of how research should be. I count myself blessed to have learned from such stellar supervisors and I hold my highest respect for them, both as researchers and as individuals.

I would also like to thank my closest colleagues with whom I share not just an office but knowledge and laughter as well: Aman Verma, who from day one taught me the ropes of SQL and who has patiently put up with all my questions on all things SQL (and any other computer issue for that matter) related; Katia

Charland, who has always kindly and thoroughly helped me in solving my biostatical problems; Masoumeh Tabaeh Izadi, without whom I probably would not have even attempted trying LaTeXon my own; and Anya Okhmatovskaia, for her R savvy assistance. Other individuals at the McGill Clinical and Health Informatics Research Group I would like to thank for their assistance and input include Genevieve Cadieux, Armel Kelome, Patricia Plouffe, and members of the Indicators meetings. I would also like to thank Julie Lainesse and David's wife Stephanie for providing last minute French help, and the McGill Centre for Bioinformatics for the funding they provided me.

Last but not least, I thank my friends (both old ones back home and new ones in Montreal) for their support and friendship, and my family - my mom, my dad and my brother - for their constant love and encouragement, even from afar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACF: autocorrelation function

AIC: Akaike information criterion

AR: autoregressive

ARIMA: autoregressive integrated moving average

CCF: cross-correlation function

ED: emergency department

GP: general practitioner

HA: hemagglutinin

ICD-9: International Classification of Diseases, Ninth Revision

ILI: influenza-like illness

MA: moving average

NA: neuraminidase

P&I: pneumonia and influenza

PACF: partial autocorrelation function

RAMQ: Régie de l'Assurance Maladie du Québec

RSV: respiratory syncytial virus

CHAPTER 1
Introduction

1.1 Background

1.1.1 Traditional Disease Surveillance

Infectious disease surveillance by public health has traditionally depended on the routine reporting of cases by physicians and laboratories, although the discovery and reporting of unusual clusters of cases by astute clinicians has often been the key to the detection of outbreaks[1] (Ashford et al., 2003; Dato et al., 2004; Morse, 2007). However, this mostly manual and passive approach to surveillance can be insensitive, inflexible, and slow as it relies on the voluntary participation of the reporters. Furthermore, there are inherent delays associated with obtaining laboratory test results for diagnostic confirmation, and reports are still sent by mail or facsimile in many areas (Morse, 2007). Consequently, it can be days or even weeks before health departments become aware of an ongoing outbreak (Birkhead et al., 1991; Jajosky and Groseclose, 2004). Sensitivity and timeliness of surveillance are further reduced when cases are spread out over a wide area.

[1] In the context of this thesis, the distinction between the terms "outbreak" and "epidemic" is not relevant and therefore these two terms have been used interchangeably in this thesis.

1.1.2 Syndromic Surveillance

In recent years, emerging infectious diseases, including influenza, and bioterrorism have been important concerns for public health. These concerns, in conjunction with the rapid evolution of travel behaviour and social contact patterns over the past few decades, have raised a pressing need for an updated approach toward disease surveillance if the monitoring of illness in the population is to be effective and efficient (Glezen, 1996; John et al., 2001; Koplan, 2001). With the major advances in information technology, automated approaches such as syndromic surveillance have emerged as novel ways to monitor and better understand the dynamics of disease (Heffernan et al., 2004; Henning, 2004). In syndromic surveillance, data are collected automatically in real-time or near real-time and continuously monitored using advanced statistical and computational methods. It typically makes use of non-traditional but convenient and abundant data sources. These include administrative data, especially health care encounter data such as emergency department (ED) chief complaints and discharge diagnoses, billing data for medical services, as well as emergency medical system (911) and health hotline calls, over-the-counter medication sales and school/work absenteeism. These pre-diagnostic data sources tend to be timelier than traditional data sources such as laboratory diagnostic test results and practitioner reports, which increases the potential for disease clusters to be identified at an earlier stage than through routine surveillance. Furthermore, since data are collected by an automated system, it poses minimal burden on the participating networks.

Although syndromic surveillance systems are promising, there is still much dispute as to whether these systems are useful additions to existing methods (Heffernan et al., 2004). Critics have expressed their doubt over the use of non-specific

and non-validated data, especially since these data are often not initially collected for the purpose of surveillance. For example, International Classification of Diseases, Ninth-Revision (ICD-9) diagnostic codes assigned to billing claims are commonly used for syndromic surveillance. However, since diagnostic code assignment is not linked to health care provider payment, they are not audited routinely, as is often done for the code assigned for services performed, and the studies to validate code assignment have tended to focus on chronic illnesses or acute injuries and hospital settings (Bazarian et al., 2006; Golomb et al., 2006; Wilchesky et al., 2004) with few assessing their use for infectious disease and in primary care settings (Cadieux and Tamblyn, 2008; Marsden-Haug et al., 2007; Schneeweiss et al., 2007). In addition, ICD-9 codes are usually mapped to a broad syndrome (such as influenza-like illnesses (ILI) in the case of influenza surveillance) to increase sensitivity among these non-validated data sources. However, this high capture comes at the cost of a loss of specificity. Finally, these data can be collected from a variety of settings and patient populations (among other variable factors) and this heterogeneity may obscure patterns of disease clusters, which in turn may obscure the utility of these types of data for disease surveillance. The accuracy and timeliness of syndromic surveillance may be limited by the uncertainty over such data quality issues and much work still needs to be done to evaluate the utility of these alternative data sources and to optimize their potential.

1.2 Rationale and Objectives

While there has been some work to evaluate different data sources for syndromic surveillance of influenza, most studies have been limited to the ED, where most of the currently operating syndromic surveillance systems are based, especially in the United States (Beitel et al., 2004; Fleischauer et al., 2004; Irvin et al., 2003;

Lemay et al., 2008; Muscatello et al., 2005; Olson et al., 2007; Suyama et al., 2003; Zheng et al., 2007). Although the potential of automated syndromic surveillance in primary care has previously been alluded to by a few (Lazarus et al., 2001; Smith et al., 2007), there has been little research to evaluate its implementation in primary care (Lazarus et al., 2002; Marsden-Haug et al., 2007; Miller et al., 2004; Sloane et al., 2006; Smith et al., 2007; van den Wijngaard et al., 2008; Yang et al., 2008). There is also evidence that children are the earliest indicators of an influenza epidemic (Brownstein et al., 2005; Lemay et al., 2008; Olson et al., 2007; Sebastian et al., 2008). However, to our knowledge, no researcher has compared subsets simultaneously stratified by different visit settings and specific patient age-groups. Moreover, while a few studies have compared different visit settings, none have compared community-based settings and hospital EDs using data derived from a single source population, which avoids the potential confounding biases that might arise in a comparison of two different populations.

There is also much annual variation in influenza epidemics due to the influenza virus's ability to constantly evolve, which results in the regular appearance of new influenza strains. However, most studies have conducted analyses using data aggregated across several influenza seasons with few taking a year-by-year approach (Lemay et al., 2008; Quenel et al., 1994). It is still unclear as to whether annual variations have implications for the potential timeliness of the data used in the syndromic surveillance of influenza.

Finally, only a few of these studies (Brownstein et al., 2005; Lemay et al., 2008; Yang et al., 2008) have used rigorous classical time series methodology to control for the autocorrelation often exhibited in these data. Autocorrelation refers to the lack of independence between different time series data points in that the value

at one time point depends on the values at previous time points. Controlling for autocorrelation is required to make valid inferences (Box and Newbold, 1971).

In this thesis, applying time series methodology to analyze ICD-9 coded medical billing data drawn from a single source population, the objectives were:

1) to describe the lead-lag relationships, by age-group and visit setting, between ILI visits in ambulatory care (community-based settings and hospital EDs) and influenza viral circulation, as represented by pneumonia and influenza (P&I) hospitalizations, and
2) to describe the year-to-year variation in these lead-lag relationships.

CHAPTER 2
Literature Review

2.1 Brief Overview of the Evaluation of Syndromic Surveillance

Although syndromic surveillance systems have the potential to improve the current state of disease surveillance, work still needs to be done to evaluate their utility. Evaluation of syndromic surveillance can be conducted from several perspectives. For example, some studies have evaluated the use of alternative data sources, while others have evaluated the performance of different statistical algorithms for the detection of epidemics, both naturally occurring and simulated. When evaluating data sources, individual records can be validated, by matching them one-to-one against a gold standard such as a laboratory diagnostic test or medical chart to determine whether they accurately represent illness. Alternatively, an aggregated data set can be compared as a whole against a similarly aggregated gold standard to examine whether there is a similarity in population trends over time. In these evaluation studies, different parameters that serve as hallmarks of utility, ranging from sensitivity, specificity, predictive value and timeliness to cost-effectiveness, can be assessed.

2.2 Timeliness of Ambulatory Visits for Influenza-Like Illnesses (ILI) for the Syndromic Surveillance of Influenza

In this literature review, we include studies that evaluated the timeliness of ambulatory visit data for the syndromic surveillance of influenza. The inclusion criteria were: the use of medical visits for influenza-like illnesses (ILI) or other general respiratory syndromes to measure influenza activity; the use of data from

community-based primary care settings, or hospital emergency departments (EDs); the reporting of measures of timeliness (lead time); and the use of influenza viral isolates, pneumonia and influenza (P&I) hospitalizations, or P&I mortality as comparison reference data. The exclusion criteria were: studies that used simulated outbreaks; studies that used aberration detection algorithms to assess timeliness; and studies that used one year or less of data. Using these criteria, we identified eleven studies (Table 2–1). This review aims to identify some of the consistent findings among these studies to synthesize the existing knowledge regarding the use of ambulatory visit data for syndromic surveillance of influenza. The methodology used, especially with respect to time series methods, will also be compared in the context of the results. Finally, this review will address some of the gaps in the literature to identify open research questions.

Table 2–1: Studies assessing the timeliness of ambulatory visit data for syndromic surveillance of ILI.

Author, Year	Study Period	Location	Methodology	Comparison Reference	Yearly Analysis?	Syndromic Data	Setting	Lead Time (weeks) [age-group][§]
Lazarus et al., 2002	1996-1999	Eastern Massachusetts	CCF (not pre-whitened)[†]	hospitalizations (lower respiratory illness syndrome)	no	ambulatory care encounters for lower respiratory illness (ICD-9)	community	2
Ivanov et al., 2003	1998-2001	Salt Lake City	CCF (not pre-whitened)[†]	hospitalizations (P&I, bronchiolitis)	yes (but reported mean)	free-text chief complaints of respiratory illness	pediatric ED	1.1[*] (mean) [<5 yr]
Miller et al., 2004	1999/04-2000/12	Minneapolis-St Paul	time interval between increases	P&I mortality	no	ambulatory care encounters for ILI (ICD-9)	community	1-2
Brownstein et al., 2005	2000-2004 (different dates for each data source)	Eastern Massachusetts	Cross-spectral analysis (with Fourier transform)	P&I mortality	no	ICD-9 diagnosis in: ambulatory care (not ED)	community	3.6[*] [5-10, 11-17 yr]-5.3[*] [3-4 yr]
						respiratory presenting complaints in: pediatric ED	ED	4.0[*] [11-17 yr]-7.1[*] [3-4 yr]
						adult ED	ED	0.6[*] [40-64 yr]-1.6[*] [>64 yr]
						general ED	ED	-1.9[*] [40-64 yr]-4.6[*] [11-17 yr]
						community ED	ED	0.4[*] [18-39 yr]-3.7[*] [3-4 yr]
Marsden-Haug et al., 2007	2000/10-2004/12	U.S. military treatment facilities worldwide	CCF (not pre-whitened)[†]	virological	no	outpatient visits	community	0
Olson et al., 2007	2003-2004	New York City	Serfling model, then CCF of excess counts	virological	no	ED fever and respiratory chief complaints	ED	-1 [≥65 yr]-1 [2-4, 5-12, 13-17 yr]
Zheng et al., 2007	2001-2005	New South Wales, Australia	Poisson model, then CCF of residuals	virological	yes (calendar year)	ED visits for influenza (ICD-9, 10)	ED	0.4[*] (all years) 0.4-2.6[*] (mean 1.1) (by year)
Lemay et al., 2008	1998-2003	Ottawa, Canada	ARIMA model, then CCF of residuals	virological	yes	ED chief complaints for ILI	ED	2-4 (mean 2.8) [<5 yr]
Sebastian et al., 2008	1998/09-2004/08	British Columbia, Canada	time interval between peaks	P&I hospitalizations	yes (but reported mean)	P&I medical visits	mixed[‡]	-0.3[*] [≥65 yr]-0.8[*] [5-9, 10-19 yr] (mean)
				P&I mortality		P&I medical visits	mixed[‡]	-0.8[*] [≥65 yr]-0.3[*] [5-9, 10-19 yr] (mean)
van den Wijngaard et al., 2008	1999-2004	Netherlands	1) CCF (not pre-whitened)[†]	hospitalizations (respiratory)	no	GP consultations registry (respiratory syndrome)	community	-1
			2) linear regression with lagged pathogen counts	virological		GP consultations registry (respiratory syndrome)	community	1 (influenza A) 2 (influenza B)
Yang et al., 2008	1998/01-2006/05	Hong Kong	wavelet analysis	virological	no	ILI consultation rates to: GPs	community	4
						general outpatient clinics (GOPC)	community	2

Abbreviations: ARIMA = autoregressive integrated moving average, CCF = cross-correlation function, ED = emergency department, GP = general practitioner, ICD-9 = International Classification of Diseases, Ninth Revision, ILI = influenza-like illness, P&I = pneumonia and influenza

[§] If no age groups are given, it means the analysis was not stratified by age

[*] Original unit of measurement in the study was days but lead times have been converted to weeks here for comparability across studies

[†] Pre-whitening refers to the application of methods that remove autocorrelation (explained in text)

[‡] Mixed refers to the fact that different settings were not distinguished (aggregated over a variety of settings), but the data came from both community and ED settings

Five (Ivanov et al., 2003; Brownstein et al., 2005; Olson et al., 2007; Lemay et al., 2008; Sebastian et al., 2008) of the eleven studies restricted their analyses to specific age-groups. Age is an obvious factor to consider since the impact of influenza is known to be highly age-specific. Five (Ivanov et al., 2003; Brownstein et al., 2005; Olson et al., 2007; Zheng et al., 2007; Lemay et al., 2008) of the eleven studies used data from the ED. Although not obvious from the selection of studies presented in this review, the ED was actually the predominant setting for studies evaluating syndromic surveillance for influenza in the literature but many of these ED-based studies were excluded from this review because they were one year pilot studies or evaluated algorithm performance. The predominance of ED-based studies is probably due to the fact that, especially in the United States, many of the already existing syndromic surveillance systems for influenza are based on ED data such as chief complaints and discharge diagnoses (Fleischauer et al., 2004; Irvin et al., 2003; Lemay et al., 2008; Muscatello et al., 2005; Olson et al., 2007), which is not surprising considering the ease of obtaining ED data compared to data from other settings. Furthermore, there is a belief that the ED may provide an earlier warning of an infectious disease outbreak than in other settings because the ED is often the place where those in need of urgent medical care will seek care first (Lemay et al., 2008). Six studies (Lazarus et al., 2002; Miller et al., 2004; Brownstein et al., 2005; Marsden-Haug et al., 2007; van den Wijngaard et al., 2008; Yang et al., 2008) used data from community-based settings (including ambulatory care and general practitioner (GP) data). One study (Sebastian et al., 2008) used data derived from a mixture of both community-based and ED settings. However, only two studies (Brownstein et al., 2005; Yang et al., 2008) compared different settings, and only one of these studies (Yang et al., 2008) assessed the utility of

restricting data to specific visit settings using data from a single source population, although none have compared community and ED settings in this manner.

There is a wide range in lead times observed, depending on the comparison reference, age-group, visit setting, and probably methodology as well. Using a mortality standard, lead times ranged from -1.9 weeks (reported as -13 days) for 40 to 64 year olds in a general ED (Brownstein et al., 2005) to 7.1 weeks (50 days) for 3 to 4 year olds in a pediatric ED (Brownstein et al., 2005). Using P&I hospitalizations as a reference, the average lead times ranged from -0.3 weeks (-1.8 days) for ≥65 year olds in ED and community settings combined (Sebastian et al., 2008) to 1.1 weeks (7.4 days) for children aged <5 years in the ED (Ivanov et al., 2003). Finally, using an influenza virological gold standard, the lead times ranged from -1 week for ≥65 year olds in the ED during the 2003-2004 season (Olson et al., 2007) to 4 weeks for children aged <5 years in the ED during the 2000-2001 season (Lemay et al., 2008) and for GP visits (Yang et al., 2008).

The most consistent finding across all settings, comparison references, and methodologies was the importance of age as an influential factor in the timing of influenza outbreak detection among administrative data. In particular, children provided an earlier warning of an influenza outbreak than adults although the exact lead times varied across different studies. However, the implication of visit setting for syndromic surveillance of influenza is not as clear from theses studies due to a lack of comparable studies within many of the age-group and setting specific categories. This review will first give an overview of the different methodologies that can be applied to assess timeliness, and then next discuss each of the studies in greater detail in the context of three standards representing influenza circulation.

2.2.1 Methodologies to Assess Timeliness

A range of methodologies has been used to assess the timeliness of candidate data sources against a gold standard for viral circulation. One simple method that is often used to define lead time measures the distance between peaks (or another point of interest) in the time series being compared, as in the studies of Sebastian et al. (2008) and Miller et al. (2004). However, there is a debate over whether the peak is an ideal hallmark of an epidemic since an epidemic may have a long and varying onset period during which many cases may occur before it actually peaks. The larger issue in this respect is that by comparing just the two peak points, this method does not make full use of all the data that are available, including the points between the two peaks. Another approach involves the use of detection algorithms to generate alerts when a certain threshold is exceeded. Dates of detection can then be compared to assess timeliness. However, we excluded studies using this approach for the reason that the assessment of timeliness can be confounded by its dependence on algorithm performance. For example, the threshold is a subjective choice and merely changing the threshold may alter the estimation of lead time. The majority of the studies in this review (Ivanov et al., 2003; Lazarus et al., 2002; Lemay et al., 2008; Marsden-Haug et al., 2007; Olson et al., 2007; van den Wijngaard et al., 2008; Zheng et al., 2007) assessed timeliness with the use of a cross-correlation function (CCF) which measures the cross-correlation between two time series that have been lagged for various units of time. Lead time can then be defined as the lag at which the peak correlation in the CCF occurs. However, the key distinction separating the studies that applied this approach is whether the temporal autocorrelation in the time series was modeled and removed prior to cross-correlating the time series. Failure to control for autocorrelation in time series data can result in high cross-correlation estimates by

chance alone (Bartlett, 1935; Box and Newbold, 1971; Helfenstein, 1996). Removing the autocorrelation from the data before analyzing the data, a process known as "pre-whitening", is one way of controlling for autocorrelation. Only six of the eleven studies applied methods that control for temporal autocorrelation. Three of these studies used methods belonging to a larger group of methodologies that are generally more appropriate time series methods. These methods can be categorized by the domain in which they analyze time series patterns. Perhaps the most popular method is that used in the study of Lemay et al. (2008), the Box-Jenkins autoregressive integrated moving average (ARIMA) modeling approach, which analyzes patterns in the temporal domain by modeling the temporal autocorrelation structure of the data. The residuals obtained after fitting ARIMA models represent the time series with the autocorrelation removed, and the CCF can then be computed using the residuals. Another study (Brownstein et al., 2005) used cross-spectral analysis, which analyzes patterns in the frequency domain. A third study (Yang et al., 2008) used wavelet analysis, which analyzes patterns in both the temporal and frequency domains. A few other studies used non-time series methods that nonetheless operate on a similar idea in that their data were used to fit models, such as the Poisson model (Zheng et al., 2007) and the Serfling model Olson et al. (2007). As with ARIMA modeling, residuals were checked for autocorrelation and then the CCF was computed using the residuals.

2.2.2 A Comparison of the Effect of Age and Setting, by Gold Standard Pneumonia and Influenza Mortality

Since the impact of influenza in terms of mortality is substantial for the elderly (Menec et al., 2003), pneumonia and influenza (P&I) mortality is often used as an indicator of viral circulation. P&I mortality was the comparison standard

of choice for three of the eleven studies of this review (Table A–1 in appendix A). The study of Brownstein et al. (2005) was the only one in this review that assessed a variety of settings and distinct age-groups simultaneously. Visits to a pediatric emergency department (ED), an adult ED, a general ED, a community ED, and a large ambulatory care group practice were analyzed by age to assess their timeliness against P&I mortality. The use of cross-spectral analysis time series methodology adds methodological rigor to this study as it appropriately controls for the autocorrelation in the data, and it compares fluctuations throughout the entire time series overall, making full use of the data. They found that pediatric age-groups generally arrived the earliest. In particular, children aged 3 to 4 years consistently presented the earliest, across all settings except for one (general ED), with a 34 day lead on average and up to a 50 day lead in the pediatric ED. Their results also indicate that among adults, community-based ambulatory care visits may provide an earlier lead than ED visits. Contrarily, among children, visits to pediatric EDs were the earliest indicators although their visits to general EDs and community EDs provided lead times that were comparable to those of ambulatory care settings.

However, the data for each setting in their study came from five distinct health-care seeking populations rather than a single source population. If the five populations were very different in other ways (for example, by socioeconomic class), these factors may confound the interpretation of the effect of visit setting. Unfortunately, few researchers have conducted a similar comparison of the influence of age and setting using either P&I mortality or the other two comparison references, disallowing a synthesis of the findings for the interaction between the effects of visit setting and age.

Sebastian et al. (2008) also used a P&I mortality reference and examined separate age-groups, although their data for P&I medical visits were aggregated across a variety of general practitioner (GP) consultations, home visits and emergency visits and therefore represent a mix of setting types in British Columbia, Canada. They found that visits by school-aged children (5 to 19 years) were the timeliest indicators of influenza activity (average lead time of 0.3 weeks (1.8 days) over P&I mortality). This pediatric population is older than the pre-schoolers identified by Brownstein et al. (2005). However, the lead times observed by Sebastian et al. (2008) were generally much lower than those observed by others, even when compared to studies that used the same comparison reference (Brownstein et al., 2005; Miller et al., 2004). In fact, P&I medical visits in the Sebastian et al. (2008) study lagged P&I mortality for certain age-groups (the two youngest: $<$6 and 6 to 23 months, and all three adult age-groups: 20 to 49, 50 to 64 and $\geq$65 years). The use of different methodology may be a possible reason for this difference. Sebastian et al. (2008) defined lead time as the time interval between the first peaks in each data set for each influenza season, an approach that does not fully analyze the patterns of these complex data, as discussed earlier. Other possible reasons include the different countries from which the study populations were derived (populations belonging to countries with structurally different health care systems may demonstrate different health care seeking behaviour) and the use of a mix of settings in the Sebastian et al. (2008) study, which may have led to the dilution or obscuring of potentially different lead times.

Miller et al. (2004) cross-correlated patient visits for influenza-like illnesses (ILI) to a large ambulatory care network in Minnesota against P&I mortality and found the correlation to be 0.41 at both the 0 and 1 week lag. Although autoregressive models were used for another part of their analyses, it was unclear from

the reported methods whether methodology controlling for temporal autocorrelation was applied in this correlation calculation, and it must be kept in mind that correlating two time series without first pre-whitening the data to remove the autocorrelation within each time series can lead to the impression of a significant correlation where there actually is none (Bartlett, 1935; Box and Newbold, 1971; Helfenstein, 1996). Furthermore, if timeliness is to be assessed by comparing cross-correlations, the cross-correlation must be calculated across various lags, thus obtaining the cross-correlation function (CCF). However, the correlation at only lags of 0 and 1 week were reported in this study, making it difficult to assess timeliness. Although the dates of alerts generated by a detection algorithm were reported and compared to the dates of the first positive influenza isolates, no formal algorithm-independent assessment of timeliness was conducted. Rather, timeliness was inferred from a visual inspection of the time between initial signs and symptoms for ILI visits and an increase in P&I mortality. This was noted to be 1 to 2 weeks, which is less than the 29 day lead found by Brownstein et al. (2005) for their ambulatory care population averaged across all ages. This approach is similar to the difference between peaks approach of the Sebastian et al. (2008) study, and again ignores a large proportion of the data that is available. Moreover, the study period is relatively short, covering roughly one and a half influenza seasons and the results may reflect the particularities of a single season.

There are limitations to the use of a P&I mortality as a reference for estimating timeliness. Since influenza is fatal mainly among the elderly and rarely among healthy children and adults, P&I mortality misses a large proportion of influenza cases. Furthermore, as a rather late-occurring outcome compared to P&I hospitalizations or influenza isolates, it can result in estimates of lead time that appear

inflated, which likely explains why those observed by Brownstein et al. (2005) were distinctively larger than those noted in the other studies in this review.

Pneumonia and Influenza Hospitalizations

Four studies compared the timing of visits for influenza-like illnesses (ILI) relative to pneumonia and influenza (P&I) hospitalizations (Table A–2 in appendix A). Three of them (Ivanov et al., 2003; Lazarus et al., 2002; van den Wijngaard et al., 2008) assessed timeliness via a cross-correlation function (CCF), but none report applying pre-whitening methods to remove autocorrelation first. Ivanov et al. (2003) calculated the CCF between free-text chief complaints of respiratory illness among children <5 years of age presenting to pediatric EDs in Salt Lake City and hospitalizations for P&I or bronchiolitis. The peak correlation was observed to be 1.1 weeks (7.4 days) on average over three influenza seasons.

In a community-based setting, different results were obtained. Lazarus et al. (2002), who were perhaps among the first to point out a gap in disease surveillance that may be addressed by an automated ambulatory care record system (Lazarus et al., 2001), correlated episodes of lower respiratory illness from an ambulatory care practice (excluding ED visits) in Massachusetts against hospital admissions with the same discharge diagnosis (Lazarus et al., 2002). This study took into consideration that multiple encounters can often be associated with a single episode of illness in ambulatory care by mandating a minimum of six weeks between new episodes of illness. A peak correlation was found at the 2 week lag, but the extremely high correlation observed (0.92) suggests that the autocorrelation within the two time series may not have been controlled for before they were correlated.

The same may have occurred for van den Wijngaard et al. (2008), who computed the CCF between a general practitioner (GP) consultation registry for

a respiratory infectious disease syndrome and a hospitalization registry for general respiratory symptoms and diagnoses in the Netherlands. The correlations are high ($\sim$0.8 peak correlation), especially given that non-specific syndromes (as opposed to pathogen-specific symptoms such as ILI) were used. Furthermore, surprisingly, respiratory GP consultations were found to lag respiratory hospitalizations (by 1 week), contrary to the consistent tendency for medical visits to lead (or at worst, coincide with) hospitalizations in other studies.

Both the van den Wijngaard et al. (2008) and Lazarus et al. (2002) studies aggregated community setting visits across age. As discussed earlier, there has been consistent evidence that children provide the earliest warning of an influenza season. Therefore, if visits to community settings are truly no better, or even worse, than ED settings in forewarning an influenza outbreak, one might expect to see a relatively better lead than what was observed for the pediatric (< 5 years) ED data in the study of Ivanov et al. (2003) who found a smaller lead than the study of Lazarus et al. (2002). However, Ivanov et al. (2003) did restrict their study to young children and while at least one study has identified young children (3 to 4 years) as the earliest sentinels of infection (Brownstein et al., 2005), at least one other study has placed this role among school-aged children (5 to 19 years) (Sebastian et al., 2008). This uncertainty regarding the earliest signaling age-group, in addition to the use of methods that do not account for autocorrelation and the lack of comparable studies with respect to age and setting specific data, make it impossible to draw any interpretations about the role of setting for syndromic surveillance of influenza from these three studies using hospitalizations as a reference for comparison.

The fourth study using a P&I hospitalizations standard was that of Sebastian et al. (2008). They drew data from across a variety of settings and measured lead

time by the time difference between peaks, an approach that also has limitations as described earlier. As with the results they obtained for a P&I mortality standard, the timeliest age-group was school-aged children of 5 to 19 years (average 5.3 day lead over P&I hospitalizations). The lead times found for children in the age-groups <5 years among these mixed setting data were shorter by a few days on average than those found by Ivanov et al. (2003) for the same age-group in the ED.

Influenza Viral Isolates

The majority of the studies in this review (six of eleven) used influenza viral isolates as a reference for comparison (Table A–3 in appendix A). Although they can be an untimely indicator due to the delays inherent to diagnostic testing, and although they may miss many cases of influenza as diagnostic confirmation is not performed for every suspected case, they remain the gold standard for influenza circulation since no other method can definitively confirm influenza infection.

In the second of two approaches used by van den Wijngaard et al. (2008) to assess the timeliness of general practitioner (GP) consultation registry data corresponding to a respiratory infectious disease syndrome, these data were used to fit multiple linear regression models with lagged pathogen counts from laboratory data as explanatory variables. Residuals were checked for autocorrelation and timeliness was assessed by identifying the lag that resulted in an optimal fit. These lags were determined to be 1 week for influenza A and 2 weeks for influenza B.

Marsden-Haug et al. (2007) validated individual International Classification of Diseases, Ninth Revision (ICD-9) codes for the detection of influenza-like illnesses (ILI) among outpatient visits to military treatment facilities within an automated syndromic system by matching individual visits to respiratory virus laboratory test results. Individual codes as well as aggregated groupings of these

codes were also correlated to positive specimens. For the individual codes, lagged correlation analysis determined that every code tended to peak at the same time as positive specimens (i.e. no lead), although there was an indication that the less frequently used (and probably more specific) codes may be more likely to lag positive specimens. However, again, it is unclear from the reported methods whether autocorrelation was taken into consideration.

In contrast to the other community-based studies of van den Wijngaard et al. (2008) and Marsden-Haug et al. (2007), the study of Yang et al. (2008) applied classical time series methodology. Using wavelet analysis, they assessed coherence between ILI consultation rates to both general outpatient clinics and GP settings, and influenza virus activity. For GPs, they found that the oscillation of the ILI consultation rate led virus isolation by 4 weeks on average. However, it should be noted that the study setting was in a tropical region, where influenza activity presents differently from and is less predictable than temperate regions, which was the authors original motivation for their study.

None of these community setting studies using a virological gold standard looked at specific age-groups. In contrast, the emergency department (ED) studies using a virological gold standard were generally more in depth, looking into other factors such as age and annual variations as well. For example, Olson et al. (2007) analyzed fever and respiratory chief complaints collected from EDs in New York City for different age-groups. To determine the number of ED visits, hospitalizations and deaths attributable to influenza, Serfling cyclical regression models were fit using these data to obtain expected counts. The cross-correlation function (CCF) between excess counts and positive viral isolates was then computed. The greatest lead time was found among school-aged children (5 to 12, and 13 to 17 years) with a 1 week lead. Although the CCF for preschool-aged children (2 to 4

years) also peaked at 1 week, the CCF lagged slightly behind that for school-aged children. In contrast, the timing of ED visits for ILI by adults and the youngest of children (< 2 years) coincided with that of viral data while the elderly lagged viral data by 1 week. Although the evidence is weaker, this study further confirms the potential utility of pediatric visits for syndromic surveillance of influenza. However, it must be kept in mind that these results reflect only one influenza season, the only one during their study period with significant excess estimates across all age-groups and therefore the only one whose data was used for the CCF. Unfortunately, no study has examined visits by specific pediatric age-groups to community settings using a virological gold standard, otherwise the effect of setting among the pediatric population would have made an informative comparison. It is interesting to note that the magnitude of the lead times demonstrated by the children in this study, relative to those of the adults, was smaller than in the Brownstein et al. (2005) study, which also similarly examined different age-groups in the ED. However, unlike the Olson et al. (2007) study, the Brownstein et al. (2005) data came from different EDs for the adults than for the children.

Unlike the other ED studies using a virological gold standard, Lemay et al. (2008) applied classical time series methodology to control for autocorrelation. In their study, ARIMA models were fit to laboratory-confirmed influenza cases for each of five influenza seasons and these models were then applied to ED consultations with a chief complaint of ILI. The CCF was computed using the residuals. In four seasons, the correlation for ILI chief complaints among children <5 years peaked 2 to 4 weeks earlier than viral isolations. These lead times are greater than those for the <5 years age-groups among the Olson et al. (2007) study, perhaps due to the different methodological approaches or study country. Lemay et al. (2008) did not report the lead times observed for other age-groups, although

they did mention that ED consultations for ILI preceded viral isolates in three of five seasons among those aged 6 to 18 years, and one of five seasons among adults. This provides additional support for the role of children as timely indicators of influenza epidemics.

The methodological approach of Zheng et al. (2007) was similar in some sense to classical methods, although it is somewhat unconventional. In this study, long-term trend and autocorrelation were first removed from daily counts of ED visits for influenza, and laboratory-confirmed cases of influenza using cubic smoothing splines. The residuals were then used to fit Poisson regression models. By computing the CCF between the Poisson model residuals and viral isolates, the lead time was determined to range from 0.4 to 2.3 weeks (reported as 3 to 18 days) with a mean of 1.1 weeks (8 days).

2.2.3 Year-by-Year Analysis

The majority of the studies covered in this review analyzed data aggregated across a span of several years (or analyzed by year but reported only the mean), but two of the eleven studies did compare the results for individual influenza seasons. Considering that each influenza season can be quite distinctive, varying in circulating strains and the extent of morbidity and mortality, this approach is worthy of investigation to determine if timeliness might be affected by the individual parameters characteristic to each influenza season. When Zheng et al. (2007) compared ED visits to virological data aggregating across all years, a 3 day lead time was found, but when individual years were used, lead times ranged from 0.4 to 2.3 weeks (reported as 3 to 18 days). However, it seems that they were analyzing by calendar year, as opposed to influenza seasons, but each calendar year actually often represents segments of two different seasons. For ED consultations for

ILI among children <5 years of age, Lemay et al. (2008) found leads of 2 to 4 weeks over virological data for four of five influenza seasons but no significant correlation was detected in the fifth season. Based on these results, lead times seem to be generally consistent for most years, but there can occasionally be major deviations in certain years, which may mean the subpopulations identified as optimal for influenza surveillance will likely only be rough generalizations and should not be expected to be reliable for every influenza season.

2.3 Conclusions and Future Directions

Looking across several studies, there is generally consistent evidence that ambulatory visits by children provide the earliest indication of an influenza epidemic. However, the importance of visit setting for syndromic surveillance of influenza, and whether visits by children to the ED may differ from their visits to community settings in terms of timeliness is a little less clear. Tables A–1, A–2, and A–3 (in appendix A) reveal the prominent gaps in the literature for such studies assessing visits by children in community settings. Considering the community setting leads obtained by Brownstein et al. (2005) for ambulatory care visits, and by Yang et al. (2008) for GP and general outpatient clinic visits, this should be a priority for future studies. The earlier lead times exhibited among community setting visits has a plausible basis as well. Since mild initial symptoms often do not require ED care (Heffernan et al., 2004), patients in the initial stages of illness often seek primary care first, which may increase the potential for an earlier ILI signal in community settings compared to the ED (Lazarus et al., 2001).

Another general observation gleaned from this literature review is that there may be a slight tendency for studies that used time series methods (Brownstein et al., 2005; Lemay et al., 2008; Yang et al., 2008) to measure greater lead times

than those that used simpler methods, although again, more studies are required to make the appropriate comparisons within each setting and age-group restricted category. Finally, an approach that analyzes these data by individual influenza seasons may be a beneficial supplement to an aggregated analysis considering that some studies have found exceptional influenza seasons.

A limitation that was common to all studies in this review was the inability to account for the contribution of respiratory syncytial virus (RSV), another major viral respiratory pathogen in the community, especially among children (Jansen et al., 2008) but among the elderly as well (Falsey et al., 2005; Fleming and Cross, 1993). In fact, RSV is one of the most important causes of lower respiratory illness among young children, its impact exceeding that of influenza (Lee et al., 2005; Schanzer et al., 2006). Since RSV often co-circulates with influenza, and since their clinical symptoms are similar, they can be difficult to distinguish without laboratory confirmation (Mathur et al., 1980; Zambon et al., 2001). As a result, it would be expected that a respiratory syndrome such as ILI would capture many cases of RSV as well, and the impact of this problem on the evaluation of these data sources for syndromic surveillance of influenza is not known.

In conclusion, there have been some informative results from these studies demonstrating the ability of automated records of medical visits to lead indicators of influenza circulation, which suggests their potential utility for syndromic surveillance of ILI. However, some questions still remain unanswered, pointing to interesting routes of investigation for future research, particularly with respect to children and community settings. With a better understanding of these patterns, improvements can be made to surveillance systems for the timelier detection of influenza outbreaks.

CHAPTER 3
Methods

3.1 Overview

The objectives of this thesis were:

1) to describe the lead-lag relationships, by age-group and visit setting, between visits for influenza-like illnesses (ILI) in ambulatory care (community-based settings and hospital emergency departments (EDs)) and the circulation of the influenza virus, as represented by pneumonia and influenza (P&I) hospitalizations, and
2) to describe the year-to-year[1] variation in these lead-lag relationships.

For a study period that ran from the first week of 1998 (starting January 4, 1998) to the last week of 2003 (ending December 27, 2003) inclusive, these analyses were split into two stages:

1) an *overall* analysis using data for all weeks across the entire study period, and
2) an *annual* analysis for each individual influenza season during our study period, using only epidemic weeks, as defined by viral data.

For both stages of the analysis, we (1) defined a study population of all patients seen by a cohort of physicians in Quebec, Canada during our study period,

[1] In this thesis, the use of the term "year" or "annual" refers to each influenza season and not each calendar year unless otherwise noted. The use of the term "season" has been limited to avoid confusion with the seasons referring to winter, spring, summer, or fall.

(2) measured medical visits for ILI made by these patients to community-based care settings and hospital EDs that were billed on a fee-for-service basis as counts by age group per week, (3) measured Quebec-wide hospitalizations for pneumonia and influenza (P&I) as counts of admissions per week, and (4) used time series methods to compare ILI visits by setting and age to population-wide P&I hospitalizations.

3.2 Context

The Régie de l'Assurance Maladie du Québec (RAMQ) is the agency responsible for the health insurance program in the province of Quebec, covering the cost of hospital and physician services for all residents. In Quebec, 99% of all residents are covered by RAMQ and between 1993 and 2003, 85-95% of physicians billed RAMQ for services conducted on a fee-for-service basis (RAMQ, 1995).

3.3 Data Sources

3.3.1 Fee-for-Service Billing Data

For our study population of patients seen by a cohort of Quebec physicians (as will be described below), we obtained from the RAMQ all fee-for-service billing claims for medical visits they made between 1993 and 2003, including those to physicians outside the original cohort of physicians. Therefore, we had complete ascertainment of healthcare delivered on a fee-for-service basis for our patient study population. Each billing claim contains information such as anonymized unique identifiers for the physician and patient, an International Classification of Diseases, Ninth Revision (ICD-9) diagnostic code, a code for the setting type, and the date of visit. Demographic data on the patients (e.g. age) were also available separately.

3.3.2 Hospitalization Data

Our Quebec-wide hospitalization data were based on records from the Quebec hospitalization database (MED-ECHO). These records included the date of admission, date of discharge, and the discharge diagnosis.

3.3.3 Viral Isolates Data

Viral testing data were obtained from the Laboratoire de Santé Publique du Québec (Quebec Public Health Laboratory). These data included weekly counts of positive specimens and the total number of specimens tested for three types of influenza diagnostic tests (culture, antigen-detection, and polymerase chain reaction). For the first two tests, positive counts for each of influenza A and B were available, while for the latter test, only the combined (influenza A and B) positive count was available.

3.4 Study Population

In a previous study (Tamblyn et al., 2007), a cohort of 3424 new physicians who took the Medical Council of Canada clinical skills examination between 1993 and 1996 and who were licensed to practice in either Ontario or Quebec was assembled. From this cohort, we then identified those who were in practice in Quebec by 1998 as indicated by the presence of at least one claim with their identifier before January 1, 1998 in the fee-for-service billing database.

The study population included all patients seen on a fee-for-service basis between 1998 and 2003 by this cohort of "in-practice" study physicians. Our study population in each year of our study period represented approximately 35%-36% of the total source population of all RAMQ beneficiaries that had received at least one medical service in the same year (RAMQ, 2008). Our patient cohort was similar

to the total provincial population by age and sex distributions except for a slight overrepresentation of the elderly and females in our study population.

3.5 Outcome Measures

3.5.1 Medical Visits for Influenza-Like Illnesses (ILI)

We generated multiple weekly time series of ILI visits by counting the number of fee-for-service billing claims with an ICD-9 coded diagnosis for ILI per week (starting Sunday, ending Saturday) for each age-group and for each of two types of outpatient settings: (1) community-based care setting (including private offices, private clinics, and local community health and social services centres), and (2) hospital ED. For the ICD-9 code set used for ILI, see Table 4-4 in Appendix A of the first manuscript or Table 5-4 of the second manuscript. The ILI code set is based on ILI groupings validated by Marsden-Haug et al. (2007) against respiratory virus laboratory test results. The RAMQ establishment codes by which we classified settings as community-based or hospital ED are defined in Table A–4 in Appendix A of the thesis. We excluded hospital-based outpatient clinics from our study. Since we expected the data to exhibit a strong weekly pattern, aggregating by week eliminated day of the week effects. Due to the fact that there were often multiple claims submitted per unique visit for multiple services rendered during a visit, each unique patient was counted no more than once per day for each subset to reduce overcounting of episodes of illness. The resulting time series of weekly counts summing daily prevalence reflects a combination of both the incidence of ILI and disease intensity.

3.5.2 Pneumonia and Influenza (P&I) Hospitalizations

We also generated a weekly time series of the total number of short-term hospitalizations in all of Quebec with a primary discharge diagnosis of P&I (ICD-9

codes 480-487) per week to serve as a common reference against which the time series of ILI visits will be compared. P&I hospitalizations data are commonly used for influenza surveillance and for measuring the impact of influenza because they provide a sensitive and representative measure of the burden of influenza morbidity within a population Perrotta et al. (1985); Upshur et al. (1999).

3.5.3 Epidemic Period Definition

Since viral culture does remain the gold standard for identifying periods of influenza circulation, it is viral isolates data that were used to define the start and end of the epidemic periods. We pooled the results of the three diagnostic tests together and defined the start of each influenza season as four weeks before the first two consecutive weeks during which the total number of positive specimens (for either influenza A or B) was greater than or equal to five. We shifted the start week back to accommodate both our expectation that an increase in positive viral tests will be preceded by an increase in ILI visits, as well as the fact that we would later be shifting the time series in the cross-correlation function (CCF) computation. The end week was defined as the week before two consecutive weeks during which the total count was under five. These epidemic periods are described in Table 5–1 in the second manuscript.

3.6 Data Analysis

3.6.1 Overview

In the application of time series methodology to assess timeliness, the goal is to determine whether changes in one time series precede changes in another time series spanning the same period of time. One approach to assessing this lead-lag relationship is to shift one time series backward or forward by a certain unit of time (for example 1 week) so that one lags the other and then computing their

correlation at this lag. This is repeated for various lags of time to obtain the cross-correlation function (CCF). The lag at which the peak correlation occurs is an indication of the duration of time by which one time series leads (or lags) the other, which we refer to as the lead time.

However, before computing the CCF between two time series, it is important to determine whether the two time series are each autocorrelated within themselves. Autocorrelation refers to the relationship in a time series that consists of observations that are not independent of each other, meaning the value of one variable depends on previous value(s) Diggle (1990). If such autocorrelation exists, the application of the usual regression approaches which require independence of observations would result in both invalid and inefficient inferences (Box and Jenkins, 1970; Zeger et al., 2006). Correlating two autocorrelated time series can lead to high correlations by chance alone, and one may come to the conclusion that an association exists where there may be none (Bartlett, 1935; Box and Newbold, 1971; Helfenstein, 1996). Furthermore, the correlation estimates at different lags may also be correlated. However, this problem can be resolved by first removing the autocorrelation in each time series before computing their cross-correlation. An excellent illustrative example is given by Bowie and Prothero (1981) who demonstrated that without controlling for autcorrelation, one would come to the conclusion that the number of deaths due to ischemic heart disease each month is related to the tonnage of imported oranges each month which logically would be difficult to believe. However, they showed that if the trend and seasonal components were first removed from each time series before being correlated, the correlation coefficients would no longer be significant.

3.6.2 Removal of Autocorrelation with ARIMA Modeling

Perhaps the most influential method for removing autocorrelation from a time series (referred to as "pre-whitening" the data) has been that of statisticians George Box and Gwilym Jenkins (Box and Jenkins, 1970). The Box-Jenkins approach fits the time series to an autoregressive integrated moving average (ARIMA) model to model the dependence between consecutive observations (Box and Jenkins, 1970; Helfenstein, 1996). Mathematically, the autoregressive moving average model ARMA(p, q) is defined by:

$$z_t = \phi_0 + \phi_1(z_{t-1}) + \ldots + \phi_p(z_{t-p}) + a_t - \theta_1(a_{t-1}) - \ldots - \theta_q(a_{t-q})$$

where $\{\ldots z_{t-1}, z_t, z_{t+1}, \ldots\}$ is a series of observations at equally spaced time intervals, $\{\ldots a_{t-1}, a_t, a_{t+1}, \ldots\}$ is a white noise series of independent and identically distributed random variables whose distribution is approximately normal with mean zero and variance σ^2, p and q are the order of the autoregressive (AR) and moving average (MA) components respectively, and $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the AR and MA parameters respectively.

The Box-Jenkins modeling procedure consists of a preliminary data preparation step, and then three main steps that are repeated as many times as necessary until an adequate model has been found. This methodology has been elaborated extensively elsewhere (Box and Jenkins, 1970; Brockwell and Davis, 2002; Diggle, 1990; Helfenstein, 1996) but a brief overview will be given here:

1. *Data preparation*

   ARIMA modeling can only be applied to a stationary time series, which is characterized by a constant mean and variance over time. A graph of the autocorrelation function (ACF) can be checked to assess stationarity. The ACF is similar to the CCF except that it represents a single time series

cross-correlated against itself. If the series is not stationary, stationarity can be achieved by "differencing" the data by subtracting the value at each time point by the value at the previous time point to detrend the series, and/or transforming the data (e.g. logarithm, square root) to stabilize the variance. If differencing is required, then the model becomes an autoregressive integrated moving average ARIMA(p, d, q) model, where d represents the order of differencing.

An example with a non-stationary time series is illustrated in Figure 4–3 of Appendix B of the first manuscript. In the ACF plot of the raw data shown in panel C, it can be seen that the autocorrelation is still very high out to large lags. However, in the ACF plot after first order differencing shown in panel D, the autocorrelation decays rapidly (it is within the 95% confidence interval of the correlation about 0 by a lag of 2).

2. *Model identification*

   In this step, graphs of the ACF and the partial autocorrelation function (PACF) of the stationary time series are assessed to attempt to identify a provisional order of AR and MA terms. The PACF is similar to the ACF in that it measures the autocorrelation at a particular lag for multiple lags except that it removes the effects of the intervening observations at the intermediate lags (Shumway and Stoffer, 2006). The details of this process of choosing the order of AR and MA terms are lengthy and therefore will not be elaborated upon here but one should be able to find an explanation in any time series methodology reference (Box and Jenkins, 1970; Helfenstein, 1996; Brockwell and Davis, 2002; Diggle, 1990).

If the time series exhibits a seasonal pattern (determined by visual inspection), differencing can be applied at the seasonal level if necessary as well, and an additional set of AR and MA terms representing the seasonal component should also be added to the model. An ARIMA$(p, d, q)\times(P, D, Q)$ model represents a model with seasonal terms, where P, D, and Q are the seasonal level equivalents of p, d, and q. If there are other external independent variables believed to have an important effect on the time series observations, they may also be added to the model.

3. *Parameter estimation*

   Values of the AR and MA coefficients which provide the best fit to the data are determined using computational algorithms (for example, maximum likelihood estimation) via standard statistical software.

4. *Model checking*

   The adequacy of a model can be checked through various diagnostic tests. If an adequate model is chosen, the residuals should be independent of each other and resemble a white noise process (with constant mean of 0 and variance σ^2 over time). If the model is inadequate, steps 2 to 4 are repeated to identify another potential model.

For our data, the ACF plots of our time series of visits for influenza-like illnesses (ILI) and its subsets by visit setting and age group indicated that they were not stationary. Therefore, before applying ARIMA modeling, we made each ILI visits time series stationary with first a log transformation and then first order differencing at the non-seasonal level. We then proceeded to identify appropriate ARIMA models for the time series of the overall data set and for each subset

using 312 weeks worth of data available for each time series to fit each model. For parsimony, we limited models to no more than one seasonal term (looking at the observation 52 weeks ago), and no higher than eighth order non-seasonal terms. We also included two indicator variables to account for winter holiday effects (for any week containing Christmas Day, Boxing Day or New Year's Day), and other holiday effects (for any week containing any other statutory holiday). See either Table 4-5 in Appendix A of the first manuscript or Table 5-5 in the Appendix of the second manuscript for a list of all statutory holidays included. Two separate holiday variables were used since the winter holidays tended to have a much more pronounced effect on the weekly count of ambulatory medical visits than did other holidays throughout the year (see Figure 4–1 in the first manuscript). Models were fit using conditional-sum-of-squares to find initial parameter values, and then using maximum likelihood estimation to refine to a more precise estimate of the parameter values. We examined the ACF, PACF, and the Ljung-Box plot of the residuals to assess the presence of autocorrelation in the residuals. A histogram of the residuals and a normal quantile plot was used to assess the normality of the residuals. The Akaike information criterion (AIC) was used to assess the goodness of fit of each model

3.6.3 Analysis of Timeliness and Correlation through the Cross-Correlation Function (CCF)

The cross-correlation function (CCF) describes the correlation between two time series shifted by various lags of time. In this way, the timeliness of one series with respect to the other can be evaluated by identifying the lag at which their correlation is the highest. However, as mentioned earlier, two autocorrelated time series must first be pre-whitened to remove their autocorrelation before their CCF can be computed to avoid high correlation estimates due to chance alone.

In accordance with Box-Jenkins methodology, an ARIMA model is first fit to the explanatory (or "input") time series (Box and Jenkins, 1970; Helfenstein, 1996). The same model must then be applied to the target (or "output") series. The residuals that result for each time series represent what remains of each time series after removing the autocorrelation. The CCF can then be computed between these residuals to obtain valid estimates of the correlation between the two time series.

Figure 4–4 in Appendix B of the first manuscript illustrates an example of a CCF between a reference time series (labeled A) and two other time series (labeled B and C) and the determination of the lead time. For simplicity and clarity, the time series being correlated are essentially lagged versions of each other. Moreover, to maintain simple looking graphs, ARIMA modeling was not first applied in order that it can be obvious from looking at the time series plots shown in panel A that time series A lags time series B by 1 time unit, and time series C by 3 time units. This lead time is more formally assessed with a CCF plot, which shows that the maximum cross-correlation between time series A and B occurs when they are lagged by 1 time unit (panel B) and for time series A and C, by 3 time units (panel C).

In our analyses, the overall ILI visits time series or its various subsets constituted our input series, while the P&I hospitalizations time series constituted our reference output series. The CCF was computed between the residuals obtained after applying the pre-whitening procedure to the input and output series as described above. For the overall analysis, the CCF was computed using the residuals for all weeks of the data. However, for the yearly analysis, the CCF was computed using the residuals restricted to the weeks that fell within the epidemic period only, repeating for each individual influenza season from 1998-1999 through 2002-2003. Examining the CCF, we assessed lead time in two ways: we identified (1) the lags

at which the peak correlation occurred, as well as (2) the lags at which other significant correlations occurred. In this way, we aimed to understand the *overall pattern* of correlation across various lags. Significance was assessed with respect to whether the correlation fell outside the 95% confidence interval about a correlation of 0 (calculated using Fisher's transformation). To better highlight these patterns, we created "heat-maps" of the CCFs, where the degree of correlation was represented on a colour gradient. For the heat-maps, the correlations were first standardized (with respect to each subset) by centering and then scaling by dividing by their root-mean-square.

Data extraction and processing were carried out using Oracle Database 10g (Release 10.2.0.1.0; Oracle Corp., Redwood City, CA) and all statistical analyses were carried out using the R statistical software (version 2.6.2; R Foundation for Statistical Computing, Vienna, Austria). This study was approved by the Faculty of Medicine Institutional Review Board at McGill University (certificate provided in appendix B).

CHAPTER 4
Preface to Manuscript #1

Many studies have implicated children as the primary vectors in influenza transmission in the community and in fact, their role as the earliest indicators of an influenza epidemic among syndromic data for influenza-like illnesses (ILI) has already clearly been established. However, the implication of visit setting in these data has not yet been widely investigated, especially for community-based settings. Furthermore, to our knowledge, no researcher has assessed the impact of age and visit setting factors simultaneously in a direct comparison using data drawn from a single source population. The use of appropriate time series methodology to control for autocorrelation when analyzing such seasonal data has also been inconsistent among the studies evaluating syndromic data in the literature.

In this manuscript, using International Classification of Diseases, Ninth Revision (ICD-9) coded medical billing data derived from a single sample of the ambulatory care seeking population in Quebec, Canada between 1998 and 2003, we assessed the timeliness of subsets of ILI visits restricted by specific age-groups and visit settings (community-based and hospital emergency department (ED)). We did so by first applying ARIMA modeling to the data to control for autocorrelation, and then computing the cross-correlation (CCF) of the residuals against a pneumonia and influenza (P&I) hospitalizations standard across various lags. We intend to submit this manuscript to *Emerging Infectious Diseases* and this manuscript has been formatted to that journal's specifications.

Manuscript #1

Title: Identifying age and setting specific leading indicators of influenza activity in physician billing data

Authors: Emily H. Chan, BSc,[1 2] Robyn Tamblyn, PhD,[1 2] Katia Charland, PhD,[1 2] and David L. Buckeridge, MD, PhD[1 2 3]

[1] Clinical and Health Informatics Research Group, McGill University, Montreal, Quebec, Canada

[2] Department of Epidemiology, McGill University, Montreal, Quebec, Canada

[3] Corresponding author

Address for Correspondence: Clinical and Health Informatics Research Group, McGill University, 1140 Pine Ave West, Montreal, Quebec, H3A 1A3, Canada; tel: 514-934-1934 ext 32991; fax: 514-843-1551; e-mail: david.buckeridge@mcgill.ca

Article Summary Line: This study demonstrated that among outpatient physician billing data, visits to community-based healthcare settings by children aged 5 to 17 years for influenza-like illnesses (ILI) were the earliest and strongest indicators of an influenza season and therefore may be a subgroup of interest for the syndromic surveillance of influenza.

MeSH Keywords: Influenza, Human; Population Surveillance; Ambulatory Care; Epidemiology; Syndrome; Medical Records Systems, Computerized; Emergency Service, Hospital; Community Health Services; Age Factors; Immunization Programs

4.1 Abstract

Although syndromic surveillance has emerged as a promising, automated approach to monitoring disease occurrence, there is little consensus about the best data sources to use. To address this problem, using physician billing data from Quebec, Canada, we assessed the timeliness of medical visits for influenza-like illnesses (ILI) by age-group at two types of outpatient healthcare settings. We computed the cross-correlation function (CCF) between time series of multiple subsets of ILI visits, and a common reference time series of hospitalizations for pneumonia and influenza (P&I). ILI visits by children aged 5-17 years to community-based care settings were more strongly correlated with P&I hospitalizations at greater lags than adult or hospital emergency department ILI visits and therefore may be the earliest and strongest indicators of an influenza season. These findings have important implications as they identify potential targets for public health strategies for controlling influenza epidemics including surveillance, vaccination and school closure.

4.2 Introduction

Influenza is an infectious respiratory disease with an annual epidemic cycle associated with high mortality among the elderly (1, 2), hospitalizations among both the elderly and the very young (1–4), and substantial economic consequences (5). Concerns of an impending influenza pandemic and potential epidemics of other emerging diseases such as SARS have fueled efforts to improve disease surveillance systems to better detect and limit the spread of infectious disease outbreaks.

Syndromic surveillance has emerged as a promising approach to monitoring disease occurrence by using advanced statistical and computational methods to continuously monitor large streams of data that are automatically collected from clinical and other non-traditional settings in real-time or near real-time (6). Since pre-diagnostic data are used, syndromic surveillance can provide an earlier indication of an outbreak than traditional laboratory or sentinel physician based surveillance. However, there is little consensus about the best data sources for this type of surveillance.

Although analyses of administrative data suggest that children are early sentinels of influenza infection in the population (2, 7–9), there is debate about which specific age-groups provide the earliest signal. It is also not clear which setting provides the earliest signal as the majority of these studies have used emergency department (ED) data (7–11) and few researchers have compared the timing of signals from visits to EDs to signals from visits to other settings such as community settings including private offices and community clinics.

To our knowledge, only one study has compared data from EDs to data from community-based settings (12), but the ED data came from one health population while the community-based data came from a different population. It must be

kept in mind that a comparison of two different populations may potentially be confounded by differences between the two populations.

In this study, we sought to clarify the timing of visits for influenza-like illnesses (ILI) by age-group and setting using physician billing data for a single cohort of patients in Quebec, Canada. To determine the age-groups and outpatient visit settings presenting the earliest signals, we cross-correlated time series of ILI visits to community settings and to EDs against a common reference time series of hospitalizations for pneumonia and influenza (P&I) across various lags of time. Identifying specific leading age-groups and settings is important for improving the timeliness of detecting an influenza epidemic by syndromic surveillance and ensuring limited resources that may be allocated for surveillance are used efficiently.

4.3 Methods

4.3.1 Overview and Study Design

We obtained data on 1) all fee-for-service medical billing claims for patients seen by a cohort of physicians in Quebec, Canada, and 2) admissions for P&I at all hospitals in Quebec during our study period running from January 4, 1998 to December 27, 2003 (inclusive). To assess the extent by which changes in outpatient visits for ILI preceded the changes in hospitalizations for P&I, we cross-correlated multiple time series of weekly counts of ILI visits to community-based care settings and to hospital EDs by different age-groups against a common reference time series of weekly counts of P&I hospitalizations after shifting the time series apart by various lags of time.

4.3.2 Context

The Régie de l'Assurance Maladie du Québec (RAMQ) is the agency responsible for the health insurance program in the province of Quebec, covering the cost of hospital and physician services for all residents. In Quebec, 99% of all residents are covered by RAMQ and between 1993 and 2003, 85-95% of physicians billed RAMQ for services conducted on a fee-for-service basis (13).

4.3.3 Data Sources

Fee-for-Service Billing Data

In a previous study, we assembled a cohort of 3424 new physicians who were licensed to practice in Ontario and/or Quebec and then requested RAMQ to identify all patients seen by these physicians between 1993 and 2003 and provide us all fee-for-service billing claims submitted for these patients by any physician in Quebec (inside or outside the cohort of study physicians) during this

period (14). Therefore, we had complete ascertainment of healthcare delivered on a fee-for-service basis for this cohort of patients. Each billing claim contains information such as anonymized unique identifiers for the physician and patient, an International Classification of Diseases, Ninth Revision (ICD-9) diagnostic code, a code for the setting type, and the date of visit.

Hospitalization Data

Our Quebec-wide hospitalization data were based on records from the Quebec hospitalization database (MED-ECHO). These records included the date of admission, date of discharge, and the discharge diagnosis.

4.3.4 Study Population

We identified all physicians as "in practice" in the fee-for-service system by January 1, 1998 if they had at least one billing claim among our fee-for-service data by this date. The study population included all patients seen by these physicians between 1998 and 2003. Our study population in each year of our study period represented approximately 35%-36% of the total source population of all RAMQ beneficiaries that had received at least one medical service in the same year (15). Our patient cohort was similar to the total provincial population by age and sex distributions except for a slight overrepresentation of the elderly and females in our study population.

4.3.5 Outcome Measures

Medical Visits for Influenza-Like Illnesses (ILI)

We generated multiple weekly time series of ILI visits by counting the number of fee-for-service billing claims with an ICD-9 coded diagnosis for ILI (code set provided in Table 4–4 in Appendix A) per week for each age-group and for each of

two types of outpatient settings: 1) community-based care setting (including private offices, private clinics, and local community health and social services centres), and 2) hospital ED. We excluded hospital-based outpatient clinics from our study. Since there were often multiple claims submitted for multiple services rendered during a single visit, each unique patient was counted no more than once per day in each time series.

Pneumonia and Influenza (P&I) Hospitalizations

We also generated a weekly time series of the total number of short-term hospitalizations in all of Quebec with a primary discharge diagnosis of P&I (ICD-9 codes 480-487) per week to serve as a common reference against which the time series of ILI visits would be compared. P&I hospitalizations data are commonly used for influenza surveillance and for measuring the impact of influenza because they provide a sensitive and representative measure of the burden of influenza morbidity within a population (16, 17).

4.3.6 Data Analysis

Removal of Autocorrelation with ARIMA Modeling

We first used Box and Jenkins seasonal autoregressive integrated moving average (ARIMA) models (18, 19) to model and control for the autocorrelation structure within each age-group and setting specific subset of the ILI visits time series. All models included two indicator variables to account for the effects of the winter and non-winter holidays (listed in Table 4–5 in Appendix A). For each subset of ILI visits, we developed several potential models and chose a final model after a comparison based on several diagnostic tests to assess the presence of any remaining autocorrelation and model fit (19). The residuals from each fitted series constituted the "pre-whitened" time series, meaning the autocorrelation has been

removed in these time series. A theoretical overview of ARIMA modeling and more technical details of how we found appropriate ARIMA models for our data are provided in Appendix B.

Analysis of Timeliness and Correlation through the Cross-Correlation Function (CCF)

A cross-correlation function (CCF) was computed to assess the timeliness of ILI visits relative to P&I hospitalizations. See Appendix B for technical details of the CCF. After finding ARIMA models for each subset of the ILI visits time series as described above, we applied the same model to the P&I hospitalizations reference time series. We then computed the CCF (up to a lag of 4 weeks) between the residuals for each subset of the ILI visits time series and the residuals for the P&I hospitalizations time series, which served as a common reference. To assess timeliness, we noted the lags at which the peak correlation and other significant correlations occurred in the CCF to determine lead time. Significance was assessed by constructing the 95% confidence interval about a correlation of 0 (calculated using Fisher's transformation). We also created "heat-maps" of these CCFs, in which the degree of correlation was represented by a colour gradient after being standardized (to each subset) by centering and then scaling by dividing by their root-mean-square.

Data extraction and processing were carried out using Oracle Database 10g (Release 10.2.0.1.0; Oracle Corp., Redwood City, CA) and all statistical analyses were carried out using the R statistical software (version 2.6.2; R Foundation for Statistical Computing, Vienna, Austria). This study was approved by the Faculty of Medicine Institutional Review Board at McGill University.

4.4 Results

4.4.1 Descriptive Statistics

Between 1998 and 2003, out of a total of 2,541,926 unique study patients with at least one billing claim from a community-based healthcare setting or a hospital ED in Quebec, there were cumulatively 1,551,173 (61%) unique patients diagnosed with ILI (including non-specific diagnoses such as fever and cough) at least once over the six year time period. The annual prevalence of ILI patients out of all study patients ranged from 21% (2003) to 28% (1998, 1999). The ILI patient population had higher proportions of females and young children, and lower proportions of the middle-aged and elderly as compared to the total patient population (Table 4–1).

Collectively, the entire study population (including those who never made an ILI visit) made a total of 73,091,025 visits, of which 5,085,226 (7%) visits were given an ILI diagnosis during the study period. During the same period, there was a total of 104,571 short-term hospitalizations with a primary diagnosis of P&I. The proportion of ILI visits out of all visits was approximately the same (7%) among both community-based care settings and hospital EDs. The majority of ILI visits were to community settings (83%) as opposed to the ED (17%). However, while working-aged adults made up a larger proportion of community-based ILI visits than they did for ED ILI visits, the converse was true among the youngest children and the elderly (Table 4–2).

Weekly counts of both ILI visits and P&I hospitalizations exhibited a highly seasonal pattern, and a general downward trend across the study period, although the trend was more pronounced for the peaks than for the troughs (Figure 4–1).

4.4.2 Timeliness and Correlation

For each cross-correlation function (CCF) between each subset of ILI visits and P&I hospitalizations, Table 4–3 shows the lag at which the peak correlation was found, as well as the greatest lag at which a significant correlation was found (see Table 4–6 in appendix A for the full table of correlations at all lags and for the final ARIMA models chosen for each subset). While visits by adults and visits to EDs tended to be more strongly correlated with P&I hospitalizations, pediatric visits and visits to community settings tended to be correlated with P&I hospitalizations at earlier lags. All peak correlations were significantly different from 0 based on a type I error of 5%.

The CCF results indicate that each ILI visits subset either led or was concurrent with P&I hospitalizations. If lead time is defined by the lag at which the peak correlation occurs, ILI visits, when not restricted by age or setting, provided no lead over P&I hospitalizations. However, when restricted by setting, the community-based setting subsets of ILI visits provided lead times that were equivalent to or greater than the lead times provided by the ED subsets for every age-group (Table 4–3). Considering age as well as setting, ILI visits by children aged 2 to 17 years old to community settings were most strongly correlated with P&I hospitalizations at the greatest lags. The greatest lag at which peak correlations occurred was 2 weeks, although significant correlations occurred at a lag of 3 weeks as well. For example, the peak correlation between the 13 to 17 year olds community-based setting subset of ILI visits and P&I hospitalizations was 0.24 and this occurred when the two time series were lagged by 2 weeks, but their cross-correlation of 0.17 at a lag of 3 weeks was also significant. In the heat-map of the CCFs (Figure 4–2), community setting ILI visits by 5 to 12 year olds and 13 to 17 year olds are the most prominent leading subsets. The subset of visits by 5 to 12 year olds to community settings is

most noticeable (in red) in the heat-map at the 2 week lag, but there is another significant (though not peak) and prominent correlation at the greater lag of 3 weeks for the subset of ILI visits by 13 to 17 year olds to community settings.

4.5 Discussion

We found that among physician billing claims from outpatient care, ILI visits to community-based setting tended to provide an earlier indication of the influenza season than ILI visits to EDs. Confirming previous findings, we also found that ILI visits by children provide an earlier indication than ILI visits by adults or the elderly. When considering both visit setting and age, we found that community setting visits for ILI by children aged 2 to 17 years tended to provide the greatest lead times over P&I hospitalizations, with community setting visits by children aged 5 to 12 years standing out in particular due to the high peak correlation with P&I hospitalizations. However, community setting visits by children aged 13 to 17 years were also significantly correlated (though not peak correlation) at an even greater lag and at a greater magnitude than for the other subsets at the same lag.

A potential reason for the earlier timing of ILI visits to community settings compared to ILI visits to the ED may be that community-based primary care is often sought first among those who seek care early because mild symptoms in the early stages of illness generally do not require ED care (20). Most studies that have evaluated the potential utility of different data sources for syndromic surveillance of influenza have focused on the ED (7–11). Few have looked at community-based settings (21–25). To our knowledge, only one study has compared these two setting types (7) but that study did not use a single cohort of patients as we did to avoid potential confounders due to the possible variation in, for example, socioeconomic status or healthcare utilization behaviour, between different populations. Our findings have important implications for syndromic surveillance of ILI because they identify specific subgroups that may be sentinels of an influenza season.

Our findings highlight the role of children as sentinels of influenza infection. This role is not surprising given that children have immature immune systems

that render them susceptible to infection. They have the second highest rate of excess hospitalizations for P&I, after the elderly, and the highest rates of excess physician visits and ED visits for P&I (1). A compounding factor is the health-care seeking behaviour of this subgroup, as concerned parents may tend to bring their ill children to a doctor at earlier stages of an illness than adults would seek care for themselves (7). The incidence of influenza-related medical visits is greatest for infants and toddlers (6 to 23 months), followed by preschool-aged children (2 to 4 years), and to a lesser extent, older school children (2). Among young children, the rate of influenza-associated clinic visits is estimated to be 3 to 8 times greater than for the ED (26). Therefore, it is ambulatory care rather than hospital care that bears the brunt of an influenza epidemic among children (1), which may explain the earlier timing of their ILI visits to community settings compared to the ED in our results.

Despite the fact that influenza has the highest impact on the youngest of children, our results point to school-aged children (5 to 17 years) as the earliest and strongest indicators of an influenza season among ILI medical visits data. This discrepancy may be explained by the role of school-aged children as the primary vectors of influenza transmission (27–30). In one simulation study, it was found that immunizing a single individual in the 13 to 19 year old age-group results in a larger decrease in new infections than any other age-group (31). The efficiency of school-aged children as spreaders and mixers of influenza may be the result of an interplay between both their innate ability to shed the virus earlier and for a longer time (even after their clinical symptoms subside) compared to adults (32, 33), as well as their extensive social contact patterns, especially among children their own age at school, sports activities, and other social activities (31, 34). School closure can lead to a significant reduction in influenza transmission (35). In contrast, the

youngest children (<2 years) not yet attending school have much more limited social contact networks. In Quebec, our study location, most public daycares do not accept children <2 years old, although there are some private or family day cares that do. In light of such evidence for the role of school-aged children in influenza transmission, the Advisory Committee on Immunization Practices (ACIP) of the Centers for Disease Control and Prevention (CDC) in the United States recently updated their recommended target groups for annual vaccine vaccination to include all children aged 5 to 18 years, starting with the 2008-2009 season (36). In previous seasons, the ACIP targeted only children 6 months to 4 years among children who are healthy. Canada has not yet adopted this expanded vaccination policy (37).

Our findings for pediatric age-groups as sentinels of infection within administrative data are consistent with other studies (2, 7–9). Both one study analyzing fever and respiratory ED chief complaints in New York City (8), and another analyzing influenza medical visits in British Columbia, Canada (2) pointed to school-aged children as sentinels, as we did. A slightly younger age-group (3 to 4 years) was identified by a study analyzing visits to several EDs and one ambulatory care setting for respiratory illness (7), while another study identified children <5 years of age to be most frequently correlated with laboratory-confirmed influenza cases among ED chief complaints of ILI (9).

There are some limitations to our study. Respiratory syncytial virus (RSV) is another major viral respiratory pathogen whose impact exceeds that of influenza among young children (38) and the elderly (39). We could not distinguish RSV from influenza since RSV is clinically similar to and often co-circulates with influenza (40). Therefore, the patterns in our time series of ILI visits may reflect a combination of the patterns of both influenza and RSV, which could diminish the correlation between the ILI visits time series and the P&I hospitalizations time

series. However, RSV mainly affects the very young, and we did use an ILI code set that has been validated against influenza viral isolates (21). We also acknowledge that our ILI visit counts overestimated the actual number of episodes of infection, although we did attempt to reduce overcounting by limiting each patient to one visit per day in each subset we analyzed. We had tried other "windows" to define episodes of infection, such as a 7-day window (i.e. each patient counted no more than once during a 7 day interval), but we found that other artificial (weekly) patterns were introduced. Daily time series would have provided more precise estimates of lead time, but we used weekly counts to eliminate day-of-the-week effects and to smooth out some of the random variation. Finally, we acknowledge there are potential biases resulting from multiple testing. Potential future directions include analyzing each influenza season separately to see if annual variations in the patterns that we observed (for e.g. due to circulating strains) may be important considerations for surveillance.

In conclusion, we found that ILI visits by school-aged children to community-based care settings provided the earliest and strongest indication of influenza circulation in ambulatory care physician billing data. We recommend the implementation of syndromic surveillance in community-based primary care, with a specific focus on school-aged children, as a valuable complement to existing surveillance systems.

4.6 Acknowledgements

We thank Jean Gratton at the Direction de Santé Publique de Montréal for providing the hospitalizations data, Rodica Gilca at the Institut National de Santé Publique du Québec for the viral isolates data, and members of the PPWC meetings at our research group for their comments on the manuscripts. This research was supported by a scholarship from the McGill Centre for Bioinformatics.

4.7 Biographical Sketch

Emily Chan is currently a graduate research assistant at the Surveillance Lab led by David Buckeridge of the McGill Clinical and Health Informatics Research Group, and recently submitted her MSc thesis in the Department of Epidemiology at McGill University, Montreal, Canada. Her research interests include infectious disease, syndromic surveillance, and healthcare data analysis.

4.8 References

[1] Menec VH, Black C, MacWilliam L, Aoki FY. The impact of influenza-associated respiratory illnesses on hospitalizations, physician visits, emergency room visits, and mortality. Can J Public Health. 2003;94(1):59–63.

[2] Sebastian R, Skowronski DM, Chong M, Dhaliwal J, Brownstein JS. Age-related trends in the timeliness and prediction of medical visits, hospitalizations and deaths due to pneumonia and influenza, British Columbia, Canada, 1998-2004. Vaccine. 2008 Mar 4;26(10):1397–403.

[3] Neuzil KM, Mellen BG, Wright PF, Mitchel J E F, Griffin MR. The effect of influenza on hospitalizations, outpatient visits, and courses of antibiotics in children. N Engl J Med. 2000 Jan 27;342(4):225–31.

[4] Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. J Infect Dis. 2000 Mar;181(3):831–7.

[5] Molinari NA, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the US: measuring disease burden and costs. Vaccine. 2007 Jun 28;25(27):5086–96.

[6] Henning KJ. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004 Sep 24;53 Suppl:5–11.

[7] Brownstein JS, Kleinman KP, Mandl KD. Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. Am J Epidemiol. 2005 Oct 1;162(7):686–93.

[8] Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the impact of influenza by age: emergency department fever and

respiratory complaint surveillance in New York City. PLoS Med. 2007 Aug 7;4(8):e247.

[9] Lemay R, Mawudeku A, Shi Y, Ruben M, Achonu C. Syndromic surveillance for influenzalike illness. Biosecur Bioterror. 2008 Jun;6(2):161–170.

[10] 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.

[11] Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. Proc Natl Acad Sci U S A. 2003 Feb 18;100(4):1961–5.

[12] Brownstein JS, Mandl KD. Pediatric Population Size Is Associated With Local Timing and Rate of Influenza and Other Acute Respiratory Infections Among Adults. Ann Emerg Med. 2008 Mar 26;.

[13] Statistiques Annuelles. Québec, Québec: Régie de l'Assurrance-Maladie du Québec (RAMQ);1995.

[14] Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007 Sep 5;298(9):993–1001.

[15] SM.13 Nombre de participants, nombre de services médicaux et leur coût par participant selon le sexe, le groupe d'âge et le type de service, rémunération à l'acte, médecine et chirurgie. Search Tool for Statistical Information (STSI). [Online Tables]. Régie de l'assurance maladie du Québec (RAMQ);[Accessed: Oct. 23, 2008]. Available from: http://www.ramq.gouv.qc.ca/fr/statistiques/index.shtml.

[16] Perrotta DM, Decker M, Glezen WP. Acute respiratory disease hospitalizations as a measure of impact of epidemic influenza. Am J Epidemiol. 1985

Sep;122(3):468–476.

[17] Upshur RE, Knight K, Goel V. Time-series analysis of the relation between influenza virus and hospital admissions of the elderly in Ontario, Canada, for pneumonia, chronic lung disease, and congestive heart failure. Am J Epidemiol. 1999 Jan 1;149(1):85–92.

[18] Box G, Jenkins G. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1970.

[19] Helfenstein U. Box-Jenkins modelling in medical research. Stat Methods Med Res. 1996 Mar;5(1):3–22.

[20] Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health. 2001;1:9.

[21] Marsden-Haug N, Foster VB, Gould PL, Elbert E, Wang H, Pavlin JA. Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. Emerg Infect Dis. 2007 Feb;13(2):207–16.

[22] Miller B, Kassenborg H, Dunsmuir W, Griffith J, Hadidi M, Nordin JD, et al. Syndromic surveillance for influenzalike illness in ambulatory care network. Emerg Infect Dis. 2004 Oct;10(10):1806–11.

[23] Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria J A, et al. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg Infect Dis. 2002 Aug;8(8):753–60.

[24] Yang L, Wong CM, Lau EHY, Chan KP, Ou CQ, Peiris JSM. Synchrony of clinical and laboratory surveillance for influenza in Hong Kong. PLoS ONE. 2008;3(1):e1399.

[25] van den Wijngaard C, van Asten L, van Pelt W, Nagelkerke NJD, Verheij R, de Neeling AJ, et al. Validation of syndromic surveillance for respiratory pathogen activity. Emerg Infect Dis. 2008 Jun;14(6):917–925.

[26] Poehling KA, Edwards KM, Weinberg GA, Szilagyi P, Staat MA, Iwane MK, et al. The Underrecognized Burden of Influenza in Young Children. N Engl J Med. 2006 July 6, 2006;355(1):31–40.

[27] Fox JP, Hall CE, Cooney MK, Foy HM. Influenzavirus infections in Seattle families, 1975-1979. I. Study design, methods and the occurrence of infections by time and age. Am J Epidemiol. 1982 Aug;116(2):212–27.

[28] Esposito S, Marchisio P, Cavagna R, Gironi S, Bosis S, Lambertini L, et al. Effectiveness of influenza vaccination of children with recurrent respiratory tract infections in reducing respiratory-related morbidity within the households. Vaccine. 2003 Jul;21(23):3162–3168.

[29] Reichert TA, Sugaya N, Fedson DS, Glezen WP, Simonsen L, Tashiro M. The Japanese experience with vaccinating schoolchildren against influenza. N Engl J Med. 2001 Mar 22;344(12):889–96.

[30] Hurwitz ES, Haber M, Chang A, Shope T, Teo S, Ginsberg M, et al. Effectiveness of influenza vaccination of day care children in reducing influenza-related morbidity among household contacts. JAMA. 2000 Oct 4;284(13):1677–82.

[31] Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am J Epidemiol. 2006 Nov;164(10):936–944.

[32] Frank AL, Taber LH, Wells CR, Wells JM, Glezen WP, Paredes A. Patterns of shedding of myxoviruses and paramyxoviruses in children. J Infect Dis. 1981 Nov;144(5):433–441.

[33] Hall CB, Douglas RG, Geiman JM, Meagher MP. Viral shedding patterns of children with influenza B infection. J Infect Dis. 1979 Oct;140(4):610–613.

[34] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLoS Medicine. 2008 March 01, 2008;5(3):e74.

[35] Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from Sentinel data. Nature. 2008 Apr 10;452(7188):750–4.

[36] Fiore AE, Shay DK, Broder K, Iskander JK, Uyeki TM, Mootrey G, et al. Prevention and control of influenza: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2008. MMWR Recomm Rep. 2008 Aug;57(RR-7):1–60.

[37] Statement on influenza vaccination for the 2008-2009 season. An Advisory Committee Statement (ASC). National Advisory Committee on Immunization (NACI). Can Commun Dis Rep. 2008;34:1–46.

[38] Fleming DM, Pannell RS, Elliot AJ, Cross KW. Respiratory illness associated with influenza and respiratory syncytial virus infection. Arch Dis Child. 2005 Jul;90(7):741–746.

[39] Falsey AR, Walsh EE. Respiratory syncytial virus infection in elderly adults. Drugs Aging. 2005;22(7):577–587.

[40] Zambon MC, Stockton JD, Clewley JP, Fleming DM. Contribution of influenza and respiratory syncytial virus to community cases of influenza-like illness: an observational study. Lancet. 2001 Oct;358(9291):1410–1416.

4.9 Tables and Figures

Table 4–1: A sex and age-group distribution comparison between the population of study patients* who made at least one visit with an influenza-like illness (ILI) diagnosis, and the total population of all study patients* (all visits) during the first and last year of the study period (1998 and 2003).

	Proportion of Patients			
	Influenza-Like Illness Study Patients		All Study Patients	
	1998 (N = 576,123)	2003 (N = 427,292)	1998 (N = 2,082,141)	2003 (N = 2,042,214)
Characteristic				
Sex				
Female	0.59	0.58	0.56	0.56
Male	0.41	0.42	0.44	0.44
Total	1.00	1.00	1.00	1.00
Age Group				
<2 years	0.08	0.08	0.04	0.04
2-4 years	0.09	0.09	0.04	0.04
5-12 years	0.12	0.11	0.08	0.08
13-17 years	0.05	0.05	0.06	0.05
18-39 years	0.26	0.27	0.31	0.31
40-64 years	0.24	0.26	0.30	0.31
$\geq$ 65 years	0.14	0.15	0.17	0.17
Total	1.00	1.00	1.00	1.00

* Each of these study patients had at least 1 billing claim from a community-based care setting (i.e. private offices, private clinics and local community health and social services centres) or a hospital emergency department in Quebec between 1998 and 2003. People with missing demographic data (<0.2%) were omitted.

Table 4–2: The proportion of influenza-like illness (ILI) visits made between 1998 and 2003 by each age-group at community-based care settings (i.e. private offices, private clinics and local community health and social services centres) and at hospital emergency departments (ED).

	Proportion of ILI Visits By Each Age Group	
Age-Group	Visits to Community-Based Care Settings (N = 4,233,782)	Visits to Hospital Emergency Department (N = 868,072)
<2 years	0.01	0.14
2-4 years	0.12	0.12
5-12 years	0.13	0.11
13-17 years	0.04	0.04
18-39 years	0.23	0.21
40-64 years	0.25	0.20
$\geq$ 65 years	0.14	0.19
Total	1.00	1.00

Table 4–3: A cross-correlation function (CCF) was computed between subsets of influenza-like illness (ILI) visits (by age-group and by setting type) to community-based care settings and hospital emergency departments (ED), and pneumonia and influenza (P&I) hospitalizations occurring between 1998 and 2003. This table provides a summary of the key features (peak correlation and earliest occurring significant correlation) of each of these CCFs.

Subset		Peak Correlation		Earliest Significant Correlation	
Setting	Age-Group	Lag (weeks)	Correlation*	Lag (weeks)	Correlation*
Overall					
Both setting types	All ages	0	0.29	3	0.12
By Visit Setting					
Community-based settings	All ages	1	0.25	3	0.13
Emergency departments	All ages	0	0.45	2	0.26
By Visit Setting × Patient Age					
Community-based settings	<2 years	1	0.25	2	0.21
	2-4 years	2	0.25	3	0.13
	5-12 years	2	0.32	2	0.32
	13-17 years	2	0.24	3	0.17
	18-39 years	1	0.34	2	0.16
	40-64 years	0	0.31	2	0.12
	≥65 years	0	0.31	1	0.17
Emergency departments	<2 years	0	0.29	1	0.18
	2-4 years	0	0.29	2	0.22
	5-12 years	2	0.19	3	0.12
	13-17 years	2	0.19	2	0.19
	18-39 years	0	0.34	2	0.18
	40-64 years	0	0.48	1	0.27
	≥65 years	0	0.57	2	0.14

* All correlations shown here were significant ($\alpha = 0.05$)

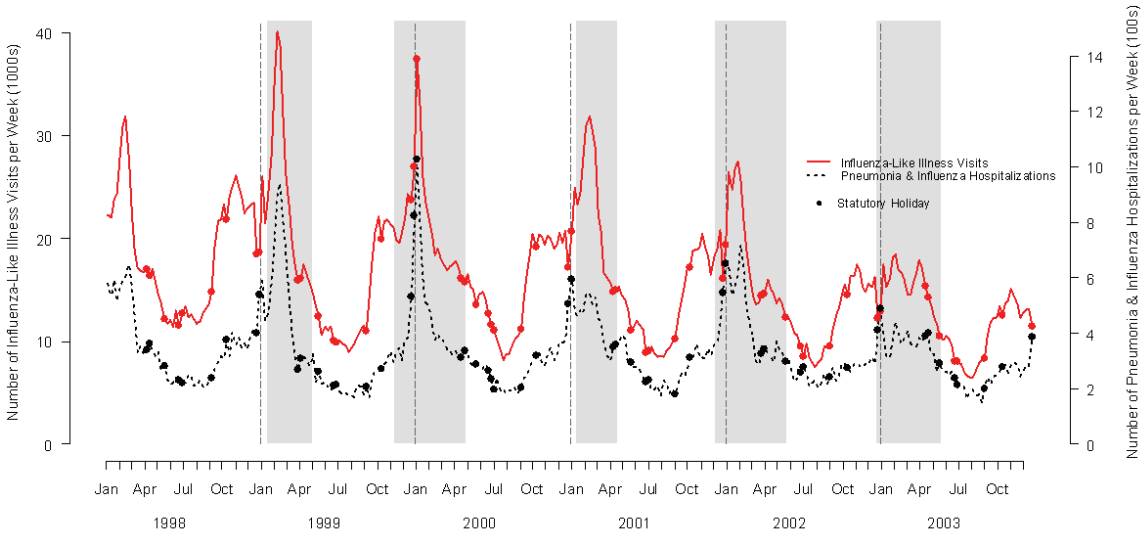

Figure 4–1: Time series plots of the weekly total counts of all influenza-like illness (ILI) visits to community-based care settings and hospital emergency departments (ED), and pneumonia and influenza (P&I) hospitalizations. Shaded regions indicate sustained periods of positive viral cultures (influenza A or B).

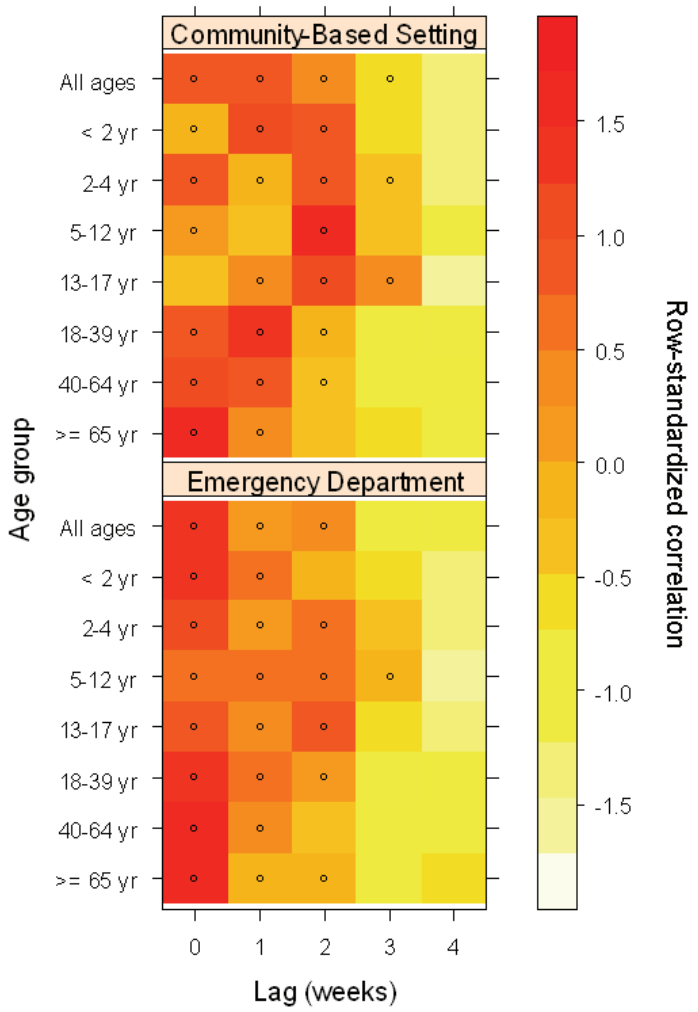

Figure 4–2: A heat-map representation of the cross-correlation functions (CCFs) between various age-group subsets of influenza-like illness (ILI) visits to community-based care settings and to hospital emergency departments (ED), and pneumonia and influenza (P&I) hospitalizations from 1998 to 2003. Correlations are represented on a colour gradient after having been standardized to each subset (row). Dots indicate correlations that were significant ($\alpha = 0.05$).

4.10 Appendix A

Table 4–4: The set of International Classification of Diseases, Ninth Revision (ICD-9) codes used in our influenza-like illness (ILI) syndrome categorization, based on ILI groupings validated by Marsden-Haug et al. (21).

ICD-9 Code	Description
079.9	Unspecified viral and chlamydial infections
382.9	Unspecified otitis media
460	Acute nasopharyngitis [common cold]
461.9	Acute sinusitis, unspecified
465.8	Acute upper respiratory infections of other multiple sites
465.9	Acute upper respiratory infections of an unspecified site
466.0	Acute bronchitis
486	Pneumonia, organism unspecified
487.0	Influenza with pneumonia
487.1	Influenza with other respiratory manifestations
487.8	Influenza with other manifestations
490	Bronchitis, not specified as acute or chronic
780.6	Fever
786.2	Cough

Table 4–5: Holiday weeks (during which statutory holidays fell*) were represented by one of two holiday indicator variables in our autoregressive integrated moving average (ARIMA) models.

Statutory Holiday	Date	Indicator Variable
New Year's Day	January 1	Winter holiday
Good Friday	Friday before Easter Sunday (varies)	Other holiday
Easter Monday	Monday after Easter Sunday (varies)	Other holiday
Victoria Day	Monday preceding May 25	Other holiday
St-Jean-Baptiste Day	June 24	Other holiday
Canada Day	July 1	Other holiday
Labour Day	First Monday of September	Other holiday
Thanksgiving Day	Second Monday in October	Other holiday
Christmas Day	December 25	Winter holiday
Boxing Day	December 26	Winter holiday

* If the holiday fell on a Saturday, both that week and the following week were treated as holiday weeks.

Table 4–6: The correlation between age-group and setting specific subsets of influenza-like illness (ILI) visits to community-based care settings and hospital emergency departments (EDs), and pneumonia and influenza (P&I) hospitalizations that occurred between 1998 and 2003, as represented by the cross-correlation function (CCF) across various lags.

Subset		Correlation					ARIMA Model
Setting	Age-Group	Lag: 0 weeks	Lag: 1 week	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	(p,d,q)†
Overall							
Both setting types	All ages	0.29*	0.26*	0.24*	0.12*	0.05	(2,1,2)
By Visit Setting							
Community-based settings	All ages	0.25*	0.25*	0.21*	0.13*	0.07	(2,1,2)
Emergency departments	All ages	0.45*	0.23*	0.26*	0.01	0.00	(4,1,4)
By Age-Group							
Both setting types	$<$ 2 years	0.15*	0.26*	0.23*	0.09	−0.03	(0,1,2)
	2-4 years	0.29*	0.16*	0.29*	0.13*	0.02	(2,1,1)
	5-12 years	0.12*	0.09	0.31*	0.11	0.01	(2,1,2)
	13-17 years	0.12*	0.17*	0.26*	0.16*	−0.06	(5,1,5)
	18-39 years	0.25*	0.33*	0.18*	0.08	0.05	(3,1,4)
	40-64 years	0.31*	0.30*	0.10	−0.01	0.00	(4,1,5)
	$\geq$ 65 years	0.41*	0.17*	0.12*	0.02	−0.03	(2,1,2)
By Visit Setting × *Age-Group*							
Community-based settings	$<$ 2 years	0.12*	0.25*	0.21*	0.08	0.00	(2,1,4)
	2-4 years	0.25*	0.17*	0.25*	0.13*	0.04	(0,1,2)
	5-12 years	0.13*	0.09	0.32*	0.09	−0.01	(2,1,2)
	13-17 years	0.07	0.16*	0.24*	0.17*	−0.05	(2,1,3)
	18-39 years	0.28*	0.34*	0.16*	0.08	0.04	(2,1,5)
	40-64 years	0.31*	0.27*	0.12*	0.06	0.04	(3,1,2)
	$\geq$ 65 years	0.31*	0.17*	0.08	0.03	−0.03	(2,1,2)
Emergency departments	$<$ 2 years	0.29*	0.18*	0.11	0.04	−0.06	(3,1,3)
	2-4 years	0.29*	0.18*	0.22*	0.10	−0.01	(0,1,4)
	5-12 years	0.19*	0.18*	0.19*	0.12*	−0.02	(3,1,4)
	13-17 years	0.19*	0.13*	0.19*	0.04	−0.05	(3,1,3)
	18-39 years	0.34*	0.25*	0.18*	0.04	0.05	(2,1,3)
	40-64 years	0.48*	0.27*	0.11	−0.02	0.01	(7,1,0)
	$\geq$ 65 years	0.57*	0.12*	0.14*	−0.03	−0.02	(1,1,0)

* Significant ($\alpha = 0.05$)

† All ARIMA models also included a first-order seasonal autoregressive term (looking back 52 weeks ago) and two holiday variables.

4.11 Appendix B

4.11.1 Autocorrelation of Time Dependent Data

A time series refers to a collection of multiple observations of the same variable over time. Time series data points are often not independent of each other in that the value at one time point can depend on the values at previous time points. This is known as autocorrelation and its consequence is that the application of the usual regression approaches which require independence of observations would result in both invalid and inefficient inferences (1, 2). For example, failure to control for temporal autocorrelation before correlating two time series can result in high correlations by chance alone (3-5).

4.11.2 Controlling Autocorrelation with ARIMA Modeling

Perhaps the most influential method for removing autocorrelation from a time series (referred to as "pre-whitening" the data) has been that of statisticians George Box and Gwilym Jenkins (1). In the Box-Jenkins approach, time series data are fit to an autoregressive integrated moving average (ARIMA) model to model the dependence between consecutive observations (1, 5). Mathematically, the autoregressive moving average model ARMA(p, q) is defined by:

$$z_t = \phi_0 + \phi_1(z_{t-1}) + \ldots + \phi_p(z_{t-p}) + a_t - \theta_1(a_{t-1}) - \ldots - \theta_q(a_{t-q})$$

where $\{\ldots z_{t-1}, z_t, z_{t+1}, \ldots\}$ is a series of observations at equally spaced time intervals, $\{\ldots a_{t-1}, a_t, a_{t+1}, \ldots\}$ is a white noise series of independent and identically distributed random variables whose distribution is approximately normal with mean zero and variance σ^2, p and q are the order of the autoregressive (AR) and moving average (MA) components respectively, and $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the AR and MA parameters respectively.

The Box-Jenkins modeling procedure consists of a preliminary data preparation step, and then three main steps that are repeated as many times as necessary until an adequate model has been found. This methodology has been elaborated extensively elsewhere (1, 5-7) but a brief overview will be given here:

1. *Data preparation*

   ARIMA modeling can only be applied to a stationary time series, which is characterized by a constant mean and variance over time. The autocorrelation function (ACF) represents the cross-correlation of a time series correlated against itself lagged by multiple units of time, and the ACF plot can be checked to assess the stationarity of a time series. If a time series is not stationary, its ACF plot would show that the autocorrelation is still positive and large out to a great time lag. On the other hand, the ACF plot of a stationary time series decays to zero or a negative autocorrelation fairly quickly (the general rule of thumb is that if the autocorrelation at a lag of 1 is close to 0 or negative, the time series is stationary). If a time series is not stationary, stationarity can be achieved by "differencing" the data by subtracting the value at each time point by the value at the previous time point to detrend a series with a non-constant mean, and/or transforming the data (e.g. logarithm, square root) to stabilize non-constant variance. Differencing can be repeated as many time as necessary (the number of times this is done is referred to as the order of differencing) to achieve stationarity although most ARIMA models generally do not exceed first order differencing. If differencing is required, then the model becomes an autoregressive integrated moving average ARIMA(p, d, q) model, where d represents the order of differencing.

An example with a non-stationary time series is illustrated in Figure 4–3. In the ACF plot of the raw data shown in panel C, it can be seen that the autocorrelation is still very high out to large lags. However, in the ACF plot after first order differencing shown in panel D, the autocorrelation decays rapidly (it is within the 95% confidence interval of the correlation about 0 by a lag of 2).

2. *Model identification*

   In this step, plots of the ACF and the partial autocorrelation function (PACF) of the stationary time series are assessed to attempt to identify a provisional order of AR and MA terms. The PACF is similar to the ACF in that it measures the autocorrelation at a particular lag for multiple lags except that it removes the effects of the intervening observations at the intermediate lags (8). The details of this process of choosing the order of AR and MA terms are lengthy and therefore will not be elaborated upon here but one should be able to find an explanation in any time series methodology reference (1, 5-7).

   If the time series exhibits a seasonal pattern (determined by visual inspection), differencing can be applied at the seasonal level if necessary as well, and an additional set of AR and MA terms representing the seasonal component should also be added to the model. An ARIMA$(p, d, q)\times(P, D, Q)$ model represents a model with seasonal terms, where P, D, and Q are the seasonal level equivalents of p, d, and q. If there are other external independent variables believed to have a significant effect on the values of the time series observations, they may also be added to the model.

In our study, the ACF plots of our time series of visits for influenza-like illnesses (ILI) and its subsets by visit setting and age group indicated that they were not stationary. Therefore, before applying ARIMA modeling, we first made each ILI visits time series stationary with first a log transformation and then first order differencing at the non-seasonal level. We then proceeded to identify ARIMA models for each subset of the ILI visits time series. For parsimony, we limited models to no more than one seasonal term (looking at the observation 52 weeks ago), and no higher than eighth order non-seasonal terms. We also included two indicator variables to account for winter holiday effects (for any week containing Christmas Day, Boxing Day or New Year's Day), and other holiday effects (for any week containing any other statutory holiday). Two separate holiday variables were used since the winter holidays tended to have a much more pronounced effect on the weekly count of ambulatory medical visits than did other holidays throughout the year

3. *Parameter estimation*

   Values of the AR and MA coefficients which provide the best fit to the data are determined using computational algorithms via standard statistical software. In this study, models were fit using conditional-sum-of-squares to find initial parameter values, and then maximum likelihood estimation to find more precise values.

4. *Model checking*

   The adequacy of a model can be checked through various diagnostic tests. For example, we assessed the presence of autocorrelation in the residuals with the ACF, PACF, and the Ljung-Box plot of the residuals, while we determined the normality of the residuals with a histogram and a normal

quantile plot of the residuals. If an adequate model is chosen, the residuals should be independent of each other and resemble a white noise process (with constant mean of 0 and variance over time). We assessed the goodness of fit of the model with the Akaike information criterion (AIC). If the model is inadequate, steps 2 to 4 are repeated to identify another potential model.

4.11.3 The Cross-Correlation Function (CCF)

The cross-correlation function (CCF) describes the correlation between two time series (one often being a standard reference time series) that have been shifted apart by different lags of time. However, as mentioned earlier, simply cross-correlating two time series which are themselves autocorrelated can result in high correlation estimates by chance alone (3-5). Therefore two autocorrelated time series must first be pre-whitened to remove their autocorrelation before their CCF can be computed. ARIMA modeling, as explained above, is one approach used for pre-whitening data. In accordance with Box-Jenkins methodology, an ARIMA model is first fit to the explanatory (or "input") time series (1, 5). The same model must then be applied to the target (or "output") series. The residuals that result for each time series represent what remains of each time series after removing the autocorrelation. The CCF can then be computed between these residuals to obtain valid estimates of the correlation between the two time series. The CCF can then be used to assess timeliness. The amount of time by which one time series leads the other (i.e. lead time) can be determined by identifying the lag at which their cross-correlation is the highest. Figure 4–4 illustrates an example of a CCF between a reference time series (labeled A) and two other time series (labeled B and C) and the determination of the lead time. For simplicity and clarity, the time series being correlated are essentially lagged versions of each other. Moreover, to maintain

simple looking graphs, ARIMA modeling was not first applied in order that it can be obvious from looking at the time series plots shown in panel A that time series A lags time series B by 1 time unit, and time series C by 3 time units. This lead time is more formally assessed with a CCF plot, which shows that the maximum cross-correlation between time series A and B occurs when they are lagged by 1 time unit (panel B) and for time series A and C, by 3 time units (panel C).

4.11.4 Appendix B References

1. Box G, Jenkins G. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1970.
2. Zeger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. Annu Rev Public Health. 2006;27:57-79.
3. Bartlett MS. Some aspects of the time-correlation problem in regard to tests of significance. J R Stat Soc. 1935;98(3):536–43.
4. Box GEP, Newbold P. Some comments on a paper of Coen, Gomme and Kendall. J R Stat Soc [Ser A]. 1971;134(2):229–40.
5. Helfenstein U. Box-Jenkins modelling in medical research. Stat Methods Med Res. 1996 Mar;5(1):3-22.
6. Brockwell PJ, Davis RA. Introduction to time series and forecasting. 2nd ed. New York: Springer; 2002.
7. Diggle PJ. Time series: a biostatistical introduction. Oxford: Clarendon Press; 1990.
8. Shumway RH, Stoffer DS. Time series analysis and its applications: with R examples: Springer; 2006.

4.11.5 Appendix B Figures

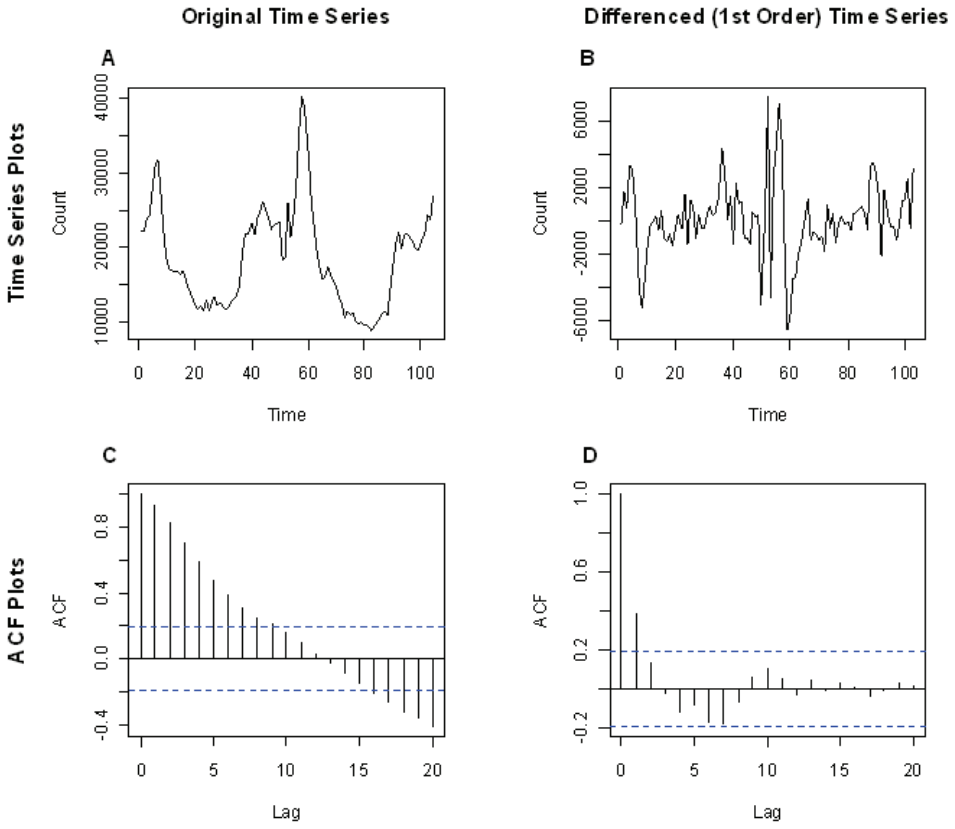

Figure 4–3: Time series plot and autocorrelation function (ACF) plot of an example non-stationary time series (panels A and C) and the same plots after first order differencing (panels B and D). Dotted lines in the ACF plots mark the 95% confidence interval about a correlation of 0.

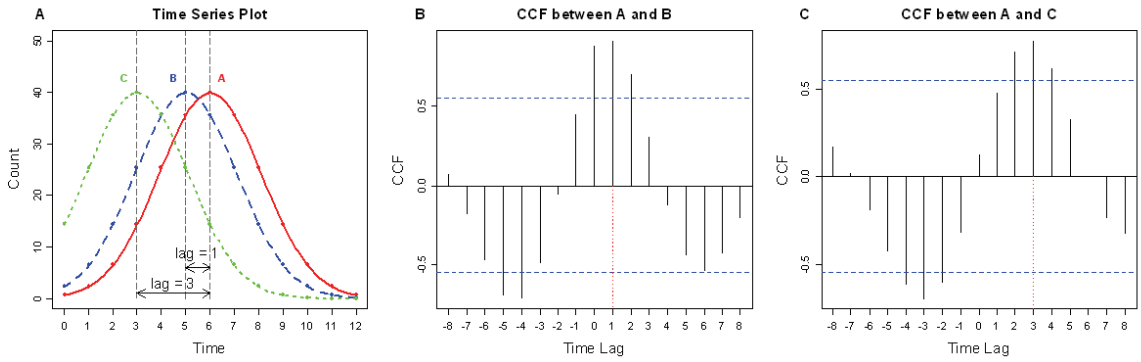

Figure 4–4: Time series plots of three example time series (panel A) that are lagged versions of each other. The cross-correlation function (CCF) can be used to identify the lag at which the maximum correlation between two time series occurs. The CCF between time series A and B (panel B) demonstrates that time series A lags time series B by 1 time unit while the CCF between time series A and C (panel C) demonstrates that time series A lags time series C by 3 time units.

CHAPTER 5
Preface to Manuscript #2

In the previous manuscript, we found that using ambulatory care physician billing data, visits for influenza-like illnesses (ILI) by school-aged children of 5 to 17 years to community settings tended to be more strongly correlated with P&I hospitalizations at earlier lags than other age-group and setting combinations. In that study, we computed cross-correlation functions (CCFs) using data spanning six years (covering five full influenza seasons), even though each influenza season on its own can be quite distinctive. Only a few studies have conducted a year-by-year analysis and none have examined age-group and setting year by year.

In this manuscript, we determined whether the timeliness of these same subsets varied annually. Such differences would imply that annual variations in influenza epidemics (for example, in circulating influenza strains) are influential factors in the timing of these subsets. In this study, using the same data, we computed separate CCFs for each annual influenza season from 1998-1999 to 2002-2003. Each season was restricted to epidemic weeks only, as defined by viral isolate data. We intend to submit this manuscript to *Emerging Infectious Diseases* and therefore this manuscript is formatted to that journal's specifications.

Manuscript #2

Title: Seasonal variation in leading indicators of influenza activity using physician billing data

Authors: Emily H. Chan, BSc,[1 2] Robyn Tamblyn, PhD,[1 2] Katia Charland, PhD,[1 2] and David L. Buckeridge, MD, PhD[1 2 3]

[1] Clinical and Health Informatics Research Group, McGill University, Montreal, Quebec, Canada

[2] Department of Epidemiology, McGill University, Montreal, Quebec, Canada

[3] Corresponding author

Address for Correspondence: Clinical and Health Informatics Research Group, McGill University, 1140 Pine Ave West, Montreal, Quebec, H3A 1A3, Canada; tel: 514-934-1934 ext 32991; fax: 514-843-1551; e-mail: david.buckeridge@mcgill.ca

Article Summary Line: Using physician billing data, this study demonstrated that annual variations inherent in each influenza season (for example, circulating influenza strains) may influence the timing of different age-group and visit setting specific subsets of outpatient visits for influenza-like illnesses (ILI), with consequences for the syndromic surveillance of influenza.

MeSH Keywords: Influenza, Human; Population Surveillance; Seasons; Epidemiology; Syndrome; Medical Records Systems, Computerized; Emergency Service, Hospital; Community Health Services; Age Factors; Antigenic Variation

5.1 Abstract

Certain subsets of patients and medical visits have been shown to be useful sentinels for syndromic surveillance of influenza. However, there has been little work done to determine if the utility of different subsets varies each year. Using outpatient physician billing data from community-based care settings and emergency departments in Quebec, Canada, we determined whether the timing of subsets of outpatient visits for influenza-like illnesses (ILI) by age-group and by type of visit setting varied from one influenza season to the next. We computed the cross-correlation function between multiple subsets of ILI visits and a common reference time series of pneumonia and influenza hospitalizations for each influenza season spanning 1998-2003. Both the earliest indicators and their lead time over P&I hospitalizations were not consistent across influenza seasons. This year-to-year variability suggests that syndromic surveillance of influenza should not focus on just a single subgroup but a combination of early-presenting subgroups.

5.2 Introduction

In recent years, the potential threats of bioterrorism and emerging infectious diseases have heightened the need for effective and efficient surveillance systems to monitor disease activity in the population (1, 2). Syndromic surveillance has emerged as a novel, automated approach to monitor as well as better understand the dynamics of disease in real-time or near real-time (1). Amidst the current climate of concern for pandemic influenza, the syndrome of influenza-like illnesses (ILI) has been one prominent target for syndromic surveillance (2–4).

In earlier work (5), we examined patterns in the timing of the presentation of patients to community-based care settings and hospital emergency departments (EDs) for ILI by age-group and by the type of visit setting. We did so using data that were aggregated across several influenza seasons and we found that ILI visits to community-based settings by 5 to 17 year olds tended to predict hospitalizations for pneumonia and influenza (P&I) earlier than ILI visits to hospital EDs or ILI visits by adults. P&I hospitalizations are a commonly used measure for tracking and measuring the impact of influenza (6, 7).

However, each influenza season can be quite different due to the constant evolution of the influenza virus, resulting in different circulating strains each year. The introduction of a new antigenic strain often leads to increased morbidity and healthcare utilization, as was the case with the B/Hong Kong/330/01 strain that spread to North America during the 2001-2002 season after a decade-long absence of the B/Victoria lineage on that continent (8) and the A/USSR/90/77 strain that arose during the 1977-1978 season after a 20 year global absence of A/H1N1 strains (9). We suspected that annual differences in the healthcare utilization in a population may modify the ability of surveillance systems to pick up signals of the onset of an influenza season each year. Therefore, in this study, we examined

whether patterns in the timing of the presentation of patients by age-group and by type of visit setting also varied by year. Understanding these patterns is important for forecasting applications and for implementing syndromic surveillance of influenza.

5.3 Methods

5.3.1 Overview and Study Design

In this study, we obtained data on 1) all fee-for-service medical billing claims for patients seen by a cohort of physicians in Quebec, Canada, and 2) admissions for P&I at all hospitals in Quebec during our study period running from January 4, 1998 to December 27, 2003 (inclusive). Additionally, viral data were used to define epidemic periods. Our objective in this study was to determine whether the timing of outpatient visits for ILI by age-group and by type of visit setting varied annually. The cross-correlation function (CCF) allows one to determine whether the changes in one time series precede the changes in another time series by shifting the two time series by various lags of time and comparing their correlation at each lag. For each of the five full influenza seasons during our study period, we computed the CCF between multiple time series of weekly counts of ILI visits to community-based care settings and to hospital EDs by different age-groups against a common reference time series of weekly counts of P&I hospitalizations to examine annual variations in the timeliness of ILI visits by each subgroup.

5.3.2 Context

The Régie de l'Assurance Maladie du Québec (RAMQ) is the agency responsible for the health insurance program in the province of Quebec, covering the cost of hospital and physician services for all residents. In Quebec, 99% of all residents are covered by RAMQ and between 1993 and 2003, 85-95% of physicians billed RAMQ for services conducted on a fee-for-service basis (10).

5.3.3 Data Sources

Viral Isolates Data

Viral testing data were obtained from the Laboratoire de Santé Publique du Québec (Quebec Public Health Laboratory). These data included weekly counts of the results of three types of diagnostic tests for influenza (culture, antigen-detection, and polymerase chain reaction).

Fee-for-Service Billing Data

In a previous study, we assembled a cohort of 3424 new physicians who were licensed to practice in Ontario and/or Quebec, and then requested RAMQ to identify all patients seen by these physicians between 1993 and 2003 and provide us all fee-for-service billing claims submitted for these patients by any physician in Quebec (inside or outside the cohort of study physicians) during this period (11). Therefore, we had complete ascertainment of healthcare delivered on a fee-for-service basis for this cohort of patients. Each billing claim contains information such as anonymized unique identifiers for the physician and patient, an International Classification of Diseases, Ninth Revision (ICD-9) diagnostic code, a code for the setting type, and the date of visit.

Hospitalization Data

Our Quebec-wide hospitalization data were based on records from the Quebec hospitalization database (MED-ECHO). These records included the date of admission, date of discharge, and the discharge diagnosis.

5.3.4 Study Population

We identified all physicians as "in practice" in the fee-for-service system by January 1, 1998 if they had at least one billing claim among our fee-for-service

data by this date. The study population comprised of all patients seen by these physicians between 1998 and 2003. Our study population in each year of our study period represented approximately 35%-36% of the total source population of all RAMQ beneficiaries that had received at least one medical service in the same year (12). Our patient cohort was similar to the total provincial population by age and sex distributions except for a slight overrepresentation of the elderly and females in our study population.

5.3.5 Outcome Measures

Epidemic Period Definition

Epidemic periods (Table 5–1) were identified using viral isolates data, the gold standard for viral circulation. We pooled the results of the three diagnostic tests together and defined the start of each epidemic period as four weeks before the first two consecutive weeks during which the total number of positive specimens (for either influenza A or B) was greater than or equal to five. We shifted the start week back to accommodate both our expectation that an increase in positive viral tests will be preceded by an increase in ILI visits, as well as the fact that we would later be shifting the time series in the cross-correlation function (CCF) computation. The end week was defined as the week before two consecutive weeks during which the total count was under five.

Medical Visits for Influenza-Like Illnesses (ILI)

We generated multiple weekly time series of ILI visits by summing the number of fee-for-service billing claims with an ICD-9 coded diagnosis for ILI (code set provided in Table 5–4 in the appendix) for each week of the study period (312 weeks total) for each age-group and for each of two types of outpatient settings: 1) community-based care setting (including private offices, private clinics, and local

community health and social services centres), and 2) hospital ED. We excluded hospital-based outpatient clinics from our study. Since there were often multiple claims submitted for multiple services rendered during a single visit, each unique patient was counted no more than once per day in each time series.

Pneumonia and Influenza Hospitalizations

We generated a weekly time series of the total number of short-term hospitalizations in all of Quebec with a primary discharge diagnosis of P&I (ICD-9 codes 480-487) for each week of the study period (312 weeks total) to serve as a common reference against which the time series of ILI visits would be compared.

5.3.6 Data Analysis

Removal of Autocorrelation with ARIMA Modeling

To control for temporal autocorrelation, we used Box and Jenkins autoregressive integrated moving average (ARIMA) models (13, 14) to first model the autocorrelation structure in the data, which we then removed by extracting and retaining just the residuals. Since we wanted to apply consistent ARIMA models from year to year, we decided to fit models using the entire data set of 312 weeks (6 years) worth of data for each subset of the ILI visits time series rather than obtaining different models for each year, although we extracted the residuals for epidemic weeks only and it is with these residuals that we later performed our subsequent analysis for each influenza season. Before fitting ARIMA models, we first made each time series stationary with a log transformation and non-seasonal differencing. In each model, we included a first order seasonal autoregressive term (defining a season as a 52 week period, first order looks back to the observation 52 weeks ago, second order 104 weeks ago, etc.) and two indicator variables representing winter and non-winter holidays (holidays are listed in Table 5–5 in the appendix). Models were

fit using conditional-sum-of-squares to find initial parameter values, and then using maximum likelihood estimation to refine to more precise estimates. Diagnostic tests used for model checking (14, 15) included the autocorrelation function (ACF), the partial ACF (PACF), the Ljung-Box plot, histogram and normal quantile plot of the residuals, and the Akaike information criterion (AIC). The final models chosen are listed in Table 5–6 in the appendix.

Analysis of Timeliness and Correlation through the Cross-Correlation Function (CCF)

After identifying appropriate ARIMA models for each subset of the ILI visits time series as described above, we applied the same models to the P&I hospitalizations time series (16), which will serve as the reference time series in the cross-correlation function (CCF). Again, the residuals for epidemic weeks were extracted. Using the epidemic week residuals from both the ILI visits time series and the P&I hospitalizations time series, we computed their CCF at lags of up to four weeks. We assessed lead time (the interval of time by which one time series leads or lags the other) by noting the lags at which the peak correlation and other statistically significant correlations occurred. Significance was assessed with respect to whether the correlation fell outside the 95% confidence interval about a correlation of 0 (calculated using Fisher's transformation). In order to understand the overall pattern of correlations across various lags, we also created "heat-maps" of the CCFs, where the degree of correlation was represented on a colour gradient. For the heat-maps, the correlations were first standardized (with respect to each subset) by centering and then scaling by dividing by their root-mean-square.

Data extraction and processing were carried out using Oracle Database 10g (Release 10.2.0.1.0; Oracle Corp., Redwood City, CA) and all statistical analyses were carried out using the R statistical software (version 2.6.2; R Foundation for

Statistical Computing, Vienna, Austria). This study was approved by the Faculty of Medicine Institutional Review Board at McGill University.

5.4 Results

5.4.1 Descriptive Statistics

Between 1998-2003, among the population of patients with ILI diagnoses, there was a slight decline across time in the proportion of young children <2 years, working adults aged 18-39 years but an increase for those 5-12 years (Table 5–2). Among the total patient population (any individual with at least one billing claim), there was a slight decline across time in the proportion of young children <2 years, but a slight increase for those 40 to 64 years (Table 5–2). There was also a general decline in the number of ILI visits and P&I hospitalizations over the study period (Figure 5–1).

5.4.2 Correlation and Timeliness

The CCFs between ILI visits and P&I hospitalizations demonstrated that although ILI visits generally led or coincided with P&I hospitalizations, the correlation at different lags varied widely from year to year (Table 5–6 in the appendix). Table 5–3 shows selected subsets that demonstrated the greatest peak correlation and other significant correlations at the earliest lags.

The heat-maps (Figure 5–2) also show that the overall utility of the ED in terms of timeliness of significant correlations is comparable to that of community settings in three seasons, better in one season (1998-1999) but worse in another (2001-2002). The most frequently appearing subsets in Table 3 were community setting visits by 13 to 17 year olds and ED visits by 5 to 12 year olds (appeared three times each), but the most frequently bolded subset (indicating it demonstrated the highest correlations among the subsets listed) was community setting visits by 13 to 17 year olds (appeared twice).

The 1998-1999 season was peculiar in that ED visits by 2 to 4, and 5 to 12 year olds provided leads of 2 and 3 weeks respectively, while, uncharacteristically, no community setting subset demonstrated a lead. There were closely matching lead times and correlations among community and ED subsets in a few other seasons as well. During the 2002-2003 season, visits by school-aged children 5 to 17 years to EDs and to community settings provided comparable lead times although community setting subsets maintained the higher correlations. During the 2001-2002 season, at community settings, visits by 18 to 39 year olds and pediatric age-groups demonstrated close peak correlations at the same lag except it was not statistically significant ($\alpha = 0.05$) for the adults. During the 1999-2000 season, although community setting visits by children 13 to 17 years demonstrated the earliest peak correlation (2 weeks), ED visits by adults aged 18 to 39 years were more strongly correlated at the same lag of 2 weeks, but its peak correlation was at the 1 week lag.

With subsets demonstrating at best a 2 week lead time in most seasons, the 2001-2002 season was particularly distinctive for the 3 week lead time demonstrated among community setting visits by those aged <2, 2 to 4, and 13 to 17 years. The peak correlations for those aged 5 to 12, and 18 to 39 years occurred at the same lag but they fell just below statistical significance ($\alpha = 0.05$). Even adults aged 40 to 64 years, unlike any other season, had a peak correlation at the 1 week lag.

5.5 Discussion

We found a degree of year-to-year variation in the timing of ILI visits to outpatient care settings by age-group and by type of visit setting relative to P&I hospitalizations, a reference standard representing influenza circulation. The lack of both consistently optimal subsets and consistent lags at which the two time series were most highly correlated has important implications for setting and age-group focused influenza surveillance. It may also make it difficult to construct accurate statistical forecasting models because such models require stable indicators of the onset or peak of an outbreak, as argued by one study (17). This study modeled influenza incidence and found that the parameters of models changed substantially between different years in order to maintain optimal fit of the models.

The year-to-year variation in the timing of age-group and setting specific subsets of ILI visits reflects the distinctiveness of each influenza season, primarily driven by the ability of the influenza virus to constantly evolve, resulting in the regular emergence of new strains. If the mutation rate is fast, fewer individuals will have had the opportunity to gain immunity to circulating strains through exposure, with the consequence of lower levels of herd immunity and thus more severe epidemics. For example, A/H3N2 viruses are believed to have faster rates of antigenic mutation (antigenic drift) than A/H1N1 and B viruses (18). It is likely for this reason that years predominated by A/H3N2 strains have long been associated with more severe epidemics, especially for young children and the elderly (19–21). Rates of excess P&I hospitalizations have been estimated to be twice as high during A/H3N2 years than in A/H1N1 or B years (21). During the 1998-1999 season, an A/H3N2 strain was the predominant strain in circulation, and we found that ILI visits to the ED were significantly correlated with P&I hospitalizations at lags of 1 week or more for all age-groups except those <2 years. On the other hand, no

community setting subset provided any lead during this season. In contrast, for the 2000-2001 season, during which no laboratory-confirmed cases of subtype A/H3N2 were reported in Quebec (22), no ED subset provided any lead. It is interesting to note that since the ED typically sees more severe cases than community settings, the contrast in lead times perhaps reflect heavier ED utilization during the more severe A/H3N2 seasons.

A relationship between age and influenza subtype has been established as well. Fox found that young school children (5 to 9 years) had the highest infection rates and were the main introducers of influenza during A/H3N2 seasons, but implicated teenagers (10 to 19 years) during A/H1N1 and B seasons (23, 23). Another study pointed to younger age-groups: those aged 1 to 4 years during type A outbreaks and those aged 5 to 9 years during mixed or type B outbreaks (24). Our results found an older pediatric population as the earliest indicators of an influenza epidemic. Community visits by 13 to 17 year olds provided the earliest leads in two of the three A/H3N2 predominant seasons (1999-2000, 2001-2002). In contrast, community visits by 5 to 12 year olds provided the best lead during the single A/H1N1 and B predominant season (2000-2001) during our study period. However, especially in light of the variation in the findings across these different studies, definitive conclusions about the dependence of timeliness on age and influenza subtype cannot be made since the study periods of all of these studies were too short (5 to 6 years) to generate enough data for each subtype.

Mismatches between vaccine strains and circulating strains often lead to reduced vaccine effectiveness, which is sometimes paralleled by a severe outbreak. For example, the 2001-2002 season marked the first time in a decade that a B/Victoria-lineage virus was circulating in North America (25). Since the B component of the 2001-2002 influenza vaccine was of the B/Yamagata lineage and

therefore a mismatch, the impact of influenza during this season was considerable, notably among school-aged children, and on ED visits (8). Reflecting the link between age and immunity, when a strain re-emerges after a long absence, children will often be particularly vulnerable to infection, in contrast to adults who often already have immunity if the strain last circulated within their life time.

Our results for the 2001-2002 season provide compelling evidence for the impact of a re-emerged strain on the timing of ILI visits by children to community settings as well. For our study location (Quebec), until the 2001-2002 season, the B/Victoria lineage had not been identified since the 1988-1989 season (26). Therefore it would be expected that most of the younger children (<13 years) had no or limited immunity. Among community setting subsets, we found a remarkable 3 week lead for most of the pediatric age-groups during this season. However, no lead was observed for the two oldest age-groups (40 to 64 and 65 years), which is consistent with their presumed immunity gained from prior exposure to this lineage.

The new A/H1N2 subtype, seemingly the product of the genetic reassortment between the circulating A/H1N1 and A/H3N2 viruses, emerged during the 2002-2003 season. Since the hemagglutinin and neuraminidase proteins of the A/H1N2 strain resembled those of the A/H1N1 and A/H3N2 strains in the 2002-2003 vaccine respectively, it was expected that the vaccine would still have had good cross-reactivity against the new subtype (27). The 2002-2003 influenza season was indeed a mild one in Quebec (28) and the lead times were obtained were average.

Although a year-by-year analysis may not always be helpful in deciding which subgroup may be most likely to provide the strongest and earliest signals for influenza surveillance each year, this approach may nonetheless be a beneficial and perhaps crucial complement to an overall analysis, as the yearly variation observed in this study and other studies suggests (29–31). Analyzing yearly variations may

prevent the mistake of making inappropriate generalizations or reveal patterns underlying different years demonstrating common traits.

Only a few other studies have taken a year-by-year approach to comparing the timing of different age and setting specific subsets of ILI visits. Although none of these studies compared specific age-groups and visit settings simultaneously as we did, they also found a variation in lead time across different years (29–31), but only a few analyzed community setting data (30) or used methodology that controls for autocorrelation (29, 31).

Limitations of this study include the inability to distinguish the impact of respiratory syncytial virus (RSV), another major viral respiratory pathogen clinically similar to influenza and with a particularly high incidence among young children (32) and the elderly (33); and the use of counts of ILI visits rather than episodes of care (which are difficult to define). We also acknowledge that our epidemic period definition is unverified, but there has been no consistent definition and a variety of approaches have been used (29, 30, 34). We also tried other slight variations in defining the epidemic period with little impact on the results. Finally, with only five influenza seasons' worth of data, it is difficult to make generalizations for specific influenza subtypes.

In conclusion, our year-by-year analysis of ILI visits has emphasized that annual variation in lead times, possibly caused by the frequent changes in the subtypes and strains of influenza viruses circulating each year, makes it difficult to pinpoint particular subsets that would always be the earliest indicators of an influenza epidemic every year. It suggests that ILI syndromic surveillance should not focus on any single subgroup but a combination of several optimally-performing subgroups. With a data set covering more influenza seasons, future research could further explore the effect of different influenza subtypes on the timeliness

of ILI visits for influenza surveillance. Future research could also examine the contribution of RSV, and other factors that are known to vary from year to year, including environmental factors. A better understanding of these relationships in the context of annual variation would help improve the accuracy of infectious disease surveillance and forecasting systems as well as the planning of public health interventions.

5.6 Acknowledgements

We thank Jean Gratton at the Direction de Santé Publique de Montréal for providing the hospitalizations data, Rodica Gilca at the Institut National de Santé Publique du Québec for the viral isolates data, and members of the PPWC meetings at our research group for their comments on the manuscripts. This research was supported by a scholarship from the McGill Centre for Bioinformatics.

5.7 Biographical Sketch

Emily Chan is currently a graduate research assistant at the Surveillance Lab led by David Buckeridge of the McGill Clinical and Health Informatics Research Group, and recently submitted her MSc thesis in the Department of Epidemiology at McGill University, Montreal, Canada. Her research interests include infectious disease, syndromic surveillance, and health care data analysis.

5.8 References

[1] Henning KJ. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004 Sep 24;53 Suppl:5–11.

[2] Lombardo JS, Burkom H, Pavlin J. ESSENCE II and the framework for evaluating syndromic surveillance systems. MMWR Morb Mortal Wkly Rep. 2004 Sep;53 Suppl:159–165.

[3] Boussard E, Flahault A, Vibert JF, Valleron AJ. Sentiweb: French communicable disease surveillance on the World Wide Web. BMJ. 1996 Nov;313(7069):1381–2; discussion 1382–4.

[4] Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis. 2004 May;10(5):858–64.

[5] Chan EH. Evaluating the use of physician billing data for age and setting specific influenza surveillance. McGill University; 2008.

[6] Barker WH. Excess pneumonia and influenza associated hospitalization during influenza epidemics in the United States, 1970-78. Am J Public Health. 1986 Jul;76(7):761–765.

[7] McBean AM, Hebert PL. New estimates of influenza-related pneumonia and influenza hospitalizations among the elderly. Int J Infect Dis. 2004 Jul;8(4):227–235.

[8] Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. PLoS Med. 2007 Aug

7;4(8):e247.

[9] Gill PW, Murphy AM, Cunningham AL. Influenza A(H1N1): a widening spectrum? Med J Aust. 1991 Sep;155(6):362–367.

[10] Statistiques Annuelles. Québec, Québec: Régie de l'Assurrance-Maladie du Québec (RAMQ);1995.

[11] Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007 Sep 5;298(9):993–1001.

[12] SM.13 Nombre de participants, nombre de services médicaux et leur coût par participant selon le sexe, le groupe d'âge et le type de service, rémunération à l'acte, médecine et chirurgie. Search Tool for Statistical Information (STSI). [Online Tables]. Régie de l'assurance maladie du Québec (RAMQ);[Accessed: Oct. 23, 2008]. Available from: http://www.ramq.gouv.qc.ca/fr/statistiques/index.shtml.

[13] Box G, Jenkins G. Time series analysis: Forecasting and control. San Francisco: Holden-Day; 1970.

[14] Helfenstein U. Box-Jenkins modelling in medical research. Stat Methods Med Res. 1996 Mar;5(1):3–22.

[15] Zeger SL, Irizarry R, Peng RD. On time series analysis of public health and biomedical data. Annu Rev Public Health. 2006;27:57–79.

[16] Box GEP, Newbold P. Some Comments on a Paper of Coen, Gomme and Kendall. J R Stat Soc [Ser A]. 1971;134(2):229–240.

[17] Andersson E, Bock D, Frisn M. Modeling influenza incidence for the purpose of on-line monitoring. Stat Methods Med Res. 2008 Aug;17(4):421–438.

[18] Hay AJ, Gregory V, Douglas AR, Lin YP. The evolution of human influenza viruses. Philos Trans R Soc Lond B Biol Sci. 2001 Dec;356(1416):1861–1870.

[19] Fleming DM. The contribution of influenza to combined acute respiratory infections, hospital admissions, and deaths in winter. Commun Dis Public Health. 2000 Mar;3(1):32–8.

[20] Simonsen L, Clarke MJ, Williamson GD, Stroup DF, Arden NH, Schonberger LB. The impact of influenza epidemics on mortality: introducing a severity index. Am J Public Health. 1997 Dec;87(12):1944–50.

[21] Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. J Infect Dis. 2000 Mar;181(3):831–7.

[22] Influenza in Canada: 2000-2001 season. Can Commun Dis Rep. 2002 Feb 1;28(3):17–28.

[23] Fox JP, Hall CE, Cooney MK, Foy HM. Influenzavirus infections in Seattle families, 1975-1979. I. Study design, methods and the occurrence of infections by time and age. Am J Epidemiol. 1982 Aug;116(2):212–27.

[24] Monto AS, Kioumehr F. The Tecumseh Study of Respiratory Illness. IX. Occurence of influenza in the community, 1966–1971. Am J Epidemiol. 1975 Dec;102(6):553–563.

[25] Shaw MW, Xu X, Li Y, Normand S, Ueki RT, Kunimoto GY, et al. Reappearance and Global Spread of Variants of Influenza B/Victoria/2/87 Lineage Viruses in the 2000-2001 and 2001-2002 Seasons. Virology. 2002;303(1):1–8.

[26] Macey JF, Tam TW, Li Y, Winchester B, Zabchuk P. Influenza in Canada: 2001-2002 season. Can Commun Dis Rep. 2003 Mar 15;29(6):45–59.

[27] Centers for Disease Control and Prevention (CDC). Update: Influenza activity–United States and worldwide, 2001-02 season, and composition of the 2002-03 influenza vaccine. MMWR Morb Mortal Wkly Rep. 2002

Jun;51(23):503–506.

[28] Influenza in Canada: 2003-2004 season. Can Commun Dis Rep. 2005 Jan;31(1):1–18.

[29] Lemay R, Mawudeku A, Shi Y, Ruben M, Achonu C. Syndromic surveillance for influenzalike illness. Biosecur Bioterror. 2008 Jun;6(2):161–170.

[30] Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, specificity and predictive values of health service based indicators for the surveillance of influenza A epidemics. Int J Epidemiol. 1994 Aug;23(4):849–55.

[31] 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.

[32] Schanzer DL, Langley JM, Tam TW. Hospitalization attributable to influenza and other viral respiratory illnesses in Canadian children. Pediatr Infect Dis J. 2006 Sep;25(9):795–800.

[33] Falsey AR, Walsh EE. Respiratory syncytial virus infection in elderly adults. Drugs Aging. 2005;22(7):577–587.

[34] Menec VH, Black C, MacWilliam L, Aoki FY. The impact of influenza-associated respiratory illnesses on hospitalizations, physician visits, emergency room visits, and mortality. Can J Public Health. 2003;94(1):59–63.

[35] Li Y. 1998-1999 influenza season: Canadian laboratory diagnoses and strain characterization. Can Commun Dis Rep. 1999 Nov 1;25(21):177–81.

[36] Li Y. 1999-2000 influenza season: Canadian laboratory diagnoses and strain characterization. Can Commun Dis Rep. 2000 Nov 15;26(22):185–9.

[37] Centers for Disease Control and Prevention (CDC). Update: influenza activity–United States and worldwide, 1999-2000 season, and composition of the 2000-01 influenza vaccine. MMWR Morb Mortal Wkly Rep. 2000

May;49(17):375–381.

[38] Statement on influenza vaccination for the 2000-2001 season. An Advisory Committee Statement (ASC). National Advisory Committee on Immunization (NACI). Can Commun Dis Rep. 2000 Jun 1;26:1–16.

[39] Marsden-Haug N, Foster VB, Gould PL, Elbert E, Wang H, Pavlin JA. Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. Emerg Infect Dis. 2007 Feb;13(2):207–16.

Table 5–1: Epidemic periods (defined using viral data) and strains circulating in Quebec (bolded if predominant) according to Canadian Communicable Disease Reports for each full influenza season during 1998-2003 (22, 26, 28, 35, 36).

Season	Epidemic Period (Duration in weeks)	Strains Circulating in Quebec	Subtype	Vaccine Mismatch?	Last Seen in Quebec*
1998-1999	Dec. 9-Apr. 28 (21)	**A/Sydney/5/97-like**	**H3N2**	no	p.s.
		B/Beijing/184/93-like	B	no	p.s.
1999-2000	Oct. 6-May 3 (31)	**A/Sydney/5/97-like**	**H3N2**	no	p.s.
		A/New Caledonia/20/99-like	H1N1	yes†	new
		B/Beijing/184/93-like	B	no	p.s.
2000-2001	Dec. 20-Apr. 25 (19)	**A/New Caledonia/20/99-like**	**H1N1**	no	p.s.
		B/Yamanashi/166/98-like‡	**B**	no	p.s.
2001-2002	Nov. 14-May 8 (26)	**A/Panama/2007/99-like**	**H3N2**	no	new
		B/Hong Kong/330/01-like§	B	yes	1988-1989
2002-2003¶	Nov. 13-May 14 (27)	**A/H1N2**	**H1N2**	no	new
		A/Panama/2007/99-like	H3N2	no	p.s.
		A/New Caledonia/20/99-like	H1N1	no	2000-2001
		B/Hong Kong/330/01-like	**B**	no	p.s.

* p.s. = previous season

† The A/New Caledonia/20/99 strain is a new antigenic variant that evolved from A/Beijing/262/95-like viruses, which was a component of the vaccine but it induced lower titers of antibodies to A/New Caledonia/20/99-like strains (37).

‡ The B/Yamanashi/166/98 strain is a B/Beijing/184/93-like virus (38).

§ The B/Hong Kong/330/01 strain is a B/Victoria/02/87 lineage virus, one of the two lineages that influenza B viruses have evolved into, but which has been absent from North America for a decade (26).

¶ For the 2002-2003 season, provincial statistics were not available and therefore, data pertaining to all of Canada were used for this season.

Table 5–2: A sex and age-group distribution comparison between the population of study patients* who made at least one visit with an influenza-like illness (ILI) diagnosis, and the total population of all study patients* (all visits) for each influenza season (epidemic weeks only) during 1998-2003.

	Proportion of Patients				
	Influenza Season				
	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003
Influenza-Like Illness Study Patients					
Number of patients	348921	432540	305706	334756	297200
Sex					
Female	0.58	0.58	0.58	0.58	0.58
Male	0.42	0.42	0.42	0.42	0.42
Total	1.00	1.00	1.00	1.00	1.00
Age-Group					
<2 yr	0.08	0.08	0.06	0.07	0.06
2-4 yr	0.10	0.10	0.11	0.11	0.11
5-12 yr	0.11	0.11	0.14	0.14	0.16
13-17 yr	0.05	0.04	0.05	0.04	0.05
18-39 yr	0.25	0.26	0.26	0.24	0.24
40-64 yr	0.25	0.26	0.25	0.25	0.25
≥65 yr	0.15	0.14	0.14	0.14	0.14
Total	1.00	1.00	1.00	1.00	1.00
All Study Patients					
Number of patients	1617413	1874619	1567539	1742367	1723604
Sex					
Female	0.57	0.57	0.57	0.57	0.57
Male	0.43	0.43	0.43	0.43	0.43
Total	1.00	1.00	1.00	1.00	1.00
Age-Group					
<2 yr	0.04	0.04	0.03	0.03	0.02
2-4 yr	0.04	0.04	0.05	0.04	0.04
5-12 yr	0.07	0.08	0.08	0.08	0.08
13-17 yr	0.05	0.05	0.04	0.04	0.04
18-39 yr	0.30	0.30	0.29	0.29	0.29
40-64 yr	0.31	0.31	0.33	0.33	0.34
≥65 yr	0.19	0.18	0.19	0.18	0.18
Total	1.00	1.00	1.00	1.00	1.00

* Each of these study patients had at least 1 billing claim from a community-based care setting (i.e. private offices, private clinics and local community health and social services centres) or a hospital emergency department in Quebec between 1998 and 2003. People with missing demographic data (<0.2%) were omitted.

Table 5–3: The cross-correlation function (CCF) was computed between various age-group and setting-specific subsets of influenza-like illness (ILI) visits, and a common reference time series of pneumonia and influenza (P&I) hospitalizations. This table shows the subsets that demonstrated the greatest "lead times" based on (1) the peak correlation and (2) the earliest significant correlation (not necessarily peak) for each influenza season during 1998-2003. Among these subsets, the one that demonstrated the greatest correlation for each column and season is bolded.

Subset		Peak Correlation		Earliest Significant Correlation	
Setting	Age-Group	Lag (weeks)	Correlation*	Lag (weeks)	Correlation*
1998-1999					
ED	**2-4 years**	2	0.63	**3**	**0.57**
ED	**5-12 years**	**3**	**0.47**	3	0.47
1999-2000					
Community	**< 2 years**	1	0.65	**2**	**0.43**
Community	**13-17 years**	**2**	**0.38**	2	0.38
ED	**18-39 years**	0	0.54	**2**	**0.45**
2000-2001					
Community	2-4 years	2	0.59	2	0.59
Community	**5-12 years**	**2**	**0.75**	**2**	**0.75**
ED	2-4 years	0	0.66	2	0.58
ED	5-12 years	0	0.58	2	0.55
2001-2002					
Community	**<2 years**	**3**	**0.50**	**3**	**0.50**
Community	**2-4 years**	**2,3**†	**0.50**	**3**	**0.50**
Community	13-17 years	3	0.40	3	0.40
2002-2003					
Community	5-12 years	2	0.41	2	0.41
Community	**13-17 years**	**2**	**0.72**	**2**	**0.72**
ED	5-12 years	2	0.45	2	0.45
ED	13-17 years	2	0.57	2	0.57

ED = emergency department; Community = community-based care setting

* All correlations shown were significant ($\alpha = 0.05$)

† Correlation at a lags of 2 and 3 weeks were the same.

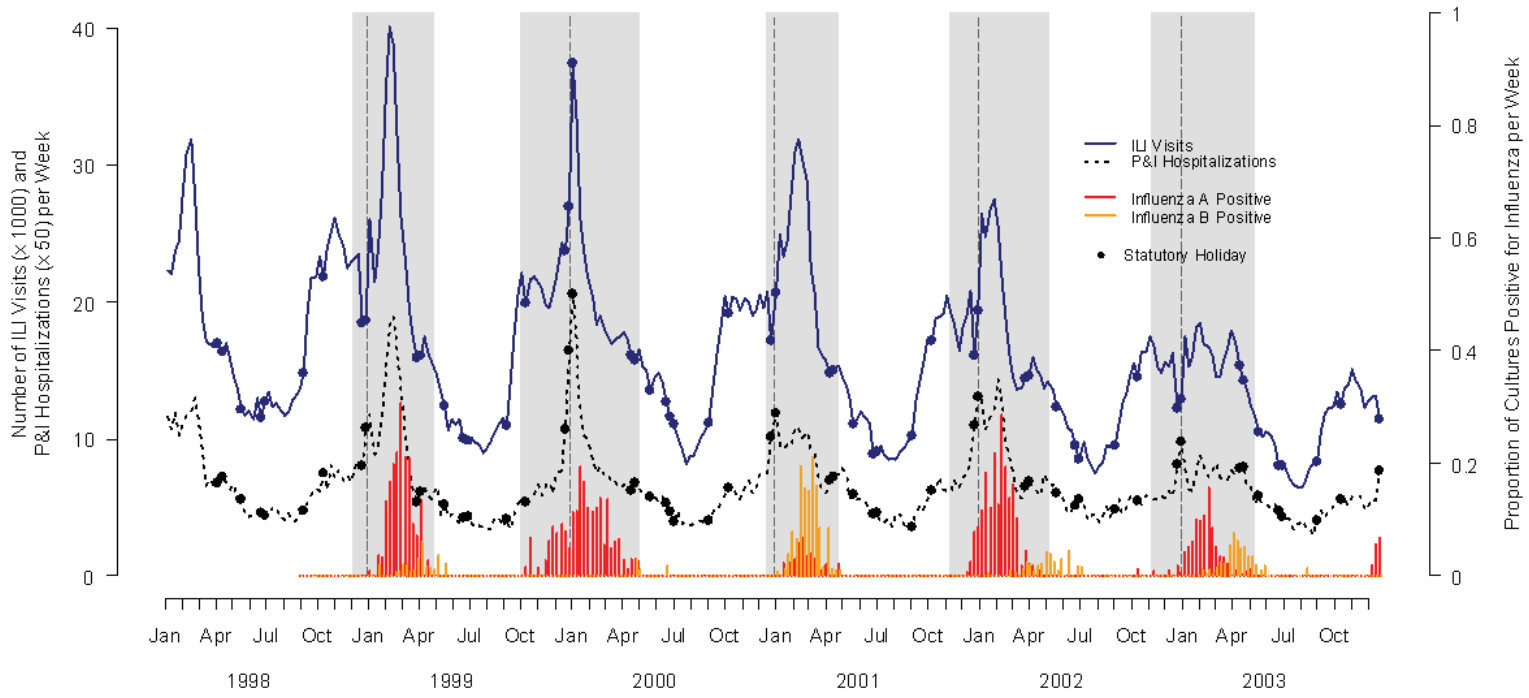

Figure 5–1: Weekly total counts of influenza-like illness (ILI) visits to community-based care settings and hospital emergency departments, and pneumonia and influenza (P&I) hospitalizations, and weekly proportions of viral cultures positive for influenza A and B, 1998-2003. Shaded regions indicate epidemic periods. Viral data prior to week 35 of 1998 were not available.

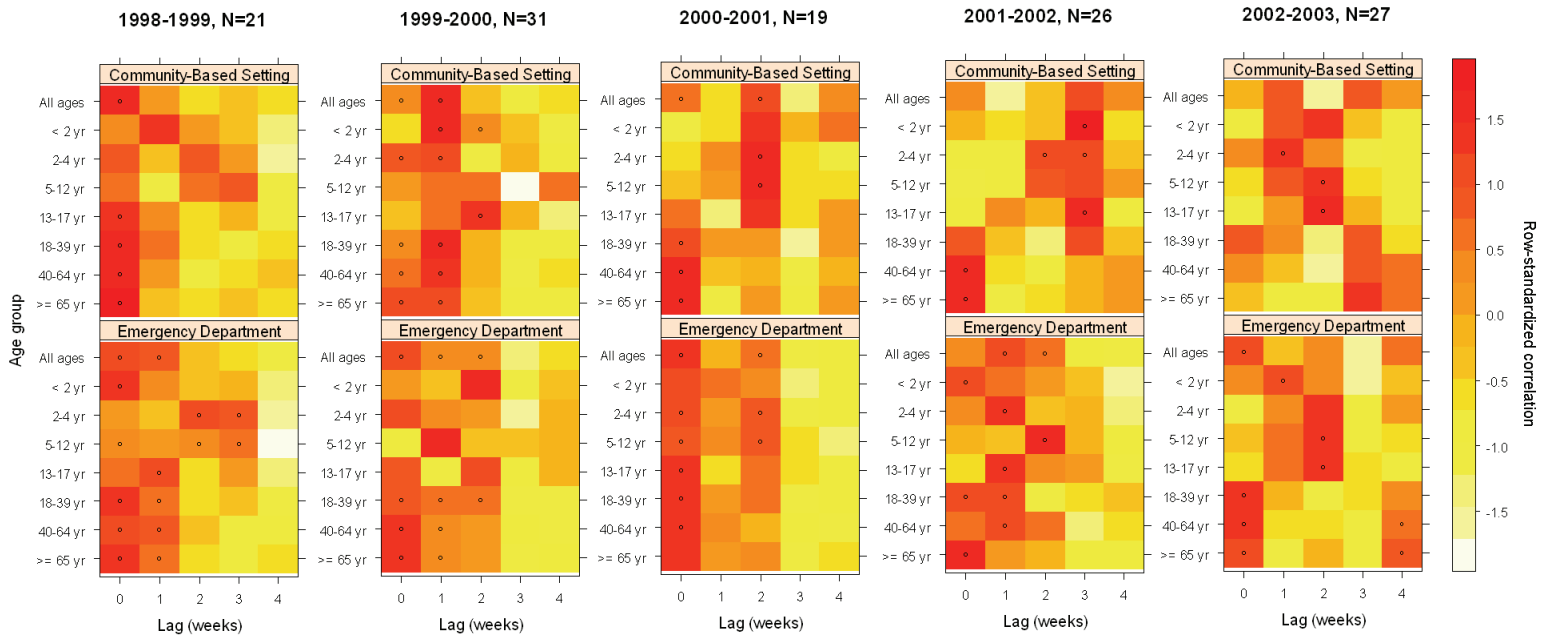

Figure 5–2: A heat-map representation of the cross-correlation functions (CCFs) between each influenza-like illness (ILI) visits time series subset by age-group and type of visit setting, against a common reference time series of pneumonia and influenza (P&I) hospitalizations during epidemic periods for each influenza season during 1998-2003. Correlations are represented on a colour gradient after having been standardized to each subset. Dots indicate correlations that were significant ($\alpha = 0.05$). N = number of weeks included for each influenza season

5.10 Appendix

Table 5–4: The set of ICD-9 codes used in our ILI syndrome categorization, based on ILI groupings validated by Marsden-Haug et al. (39).

ICD-9 Code	Description
079.9	Unspecified viral and chlamydial infections
382.9	Unspecified otitis media
460	Acute nasopharyngitis [common cold]
461.9	Acute sinusitis, unspecified
465.8	Acute upper respiratory infections of other multiple sites
465.9	Acute upper respiratory infections of an unspecified site
466.0	Acute bronchitis
486	Pneumonia, organism unspecified
487.0	Influenza with pneumonia
487.1	Influenza with other respiratory manifestations
487.8	Influenza with other manifestations
490	Bronchitis, not specified as acute or chronic
780.6	Fever
786.2	Cough

Table 5–5: Holiday weeks (during which statutory holidays fell*) were represented by one of two holiday indicator variables in our autoregressive integrated moving average (ARIMA) models.

Statutory Holiday	Date	Indicator Variable
New Year's Day	January 1	Winter holiday
Good Friday	Friday before Easter Sunday (varies)	Other holiday
Easter Monday	Monday after Easter Sunday (varies)	Other holiday
Victoria Day	Monday preceding May 25	Other holiday
St-Jean-Baptiste Day	June 24	Other holiday
Canada Day	July 1	Other holiday
Labour Day	First Monday of September	Other holiday
Thanksgiving Day	Second Monday in October	Other holiday
Christmas Day	December 25	Winter holiday
Boxing Day	December 26	Winter holiday

* If the holiday fell on a Saturday, both that week and the following week were treated as holiday weeks.

Table 5–6: The cross-correlation function (CCF) between various subsets of influenza-like illness (ILI) visits to community-based care settings and hospital emergency departments (ED), and a common reference time series of pneumonia and influenza (P&I) hospitalizations for each influenza season during 1998-2003. Both the peak correlation and the earliest significant correlation (not necessarily peak) for each subset are bolded.

1998-1999

Subset		Correlation					ARIMA
Type of Setting	Age-Group	Lag: 0 weeks	Lag: 1 weeks	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	Model (p,d,q)†
Community-based setting	All ages	**0.63***	0.26	0.10	0.17	0.11	(2,1,2)
	<2 years	0.22	**0.42**	0.17	0.07	−0.16	(2,1,4)
	2-4 years	0.37	0.17	**0.40**	0.28	−0.04	(0,1,2)
	5-12 years	0.35	0.03	0.31	**0.39**	−0.01	(2,1,2)
	13-17 years	**0.46***	0.29	0.17	0.25	0.07	(2,1,3)
	18-39 years	**0.64***	0.35	−0.01	−0.02	0.01	(2,1,5)
	40-64 years	**0.73***	0.29	−0.04	0.02	0.06	(3,1,2)
	≥65 years	**0.67***	−0.02	−0.12	−0.03	−0.09	(2,1,2)
Emergency department	All ages	**0.66***	**0.58***	0.23	0.18	−0.01	(4,1,4)
	<2 years	**0.69***	0.35	0.10	0.14	−0.30	(3,1,3)
	2-4 years	0.42	0.30	**0.63***	**0.57***	0.03	(0,1,4)
	5-12 years	0.44*	0.38	0.44*	**0.47***	−0.01	(3,1,4)
	13-17 years	0.37	**0.50***	0.14	0.30	−0.05	(3,1,3)
	18-39 years	**0.69***	**0.45***	0.15	0.25	0.06	(2,1,3)
	40-64 years	**0.67***	**0.60***	0.17	0.00	−0.01	(7,1,0)
	≥65 years	**0.87***	**0.56***	0.04	−0.01	0.02	(1,1,0)

* Significant ($\alpha = 0.05$)

† All ARIMA models also included a first-order seasonal autoregressive term and two holiday variables

1999-2000

Subset		Correlation					ARIMA
Type of Setting	Age-Group	Lag: 0 weeks	Lag: 1 weeks	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	Model (p,d,q)†
Community-based setting	All ages	0.39*	**0.63***	0.24	0.14	0.18	(2,1,2)
	<2 years	0.26	**0.65***	**0.43***	0.29	0.15	(2,1,4)
	2-4 years	0.37*	**0.42***	0.02	0.16	0.04	(0,1,2)
	5-12 years	0.12	**0.18**	**0.18**	−0.12	**0.19**	(2,1,2)
	13-17 years	0.20	0.30	**0.38***	0.22	0.09	(2,1,3)
	18-39 years	0.40*	**0.73***	0.25	0.07	0.06	(2,1,5)
	40-64 years	0.48*	**0.68***	0.27	0.01	0.15	(3,1,2)
	≥65 years	0.54*	**0.59***	0.12	0.00	−0.05	(2,1,2)
Emergency department	All ages	**0.62***	0.41*	**0.38***	−0.15	0.06	(4,1,4)
	<2 years	0.14	0.10	**0.27**	0.03	0.11	(3,1,3)
	2-4 years	0.32‡	0.17	0.12	−0.22	0.09	(0,1,4)
	5-12 years	−0.05	**0.20**	0.01	0.02	0.04	(3,1,4)
	13-17 years	0.24	−0.25	**0.35**	−0.19	0.03	(3,1,3)
	18-39 years	**0.54***	0.43*	**0.45***	−0.01	0.03	(2,1,3)
	40-64 years	**0.78***	**0.38***	0.35	−0.09	−0.10	(7,1,0)
	≥65 years	**0.78***	**0.42***	0.27	−0.09	−0.05	(1,1,0)

* Significant ($\alpha = 0.05$)

† All ARIMA models also included a first-order seasonal autoregressive term and two holiday variables

‡ A significant peak correlation was actually found at a lag of -1 week.

2000-2001

Subset		Correlation					ARIMA
Type of Setting	Age-Group	Lag: 0 weeks	Lag: 1 weeks	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	Model (p,d,q)†
Community-based setting	All ages	0.46*	0.27	**0.54***	0.17	0.43	(2,1,2)
	<2 years	0.05	0.10	**0.35**	0.17	0.24	(2,1,4)
	2-4 years	0.26	0.43	**0.59***	0.28	0.25	(0,1,2)
	5-12 years	0.24	0.37	**0.75***	0.15	0.15	(2,1,2)
	13-17 years	0.27	0.08	**0.36**	0.15	0.22	(2,1,3)
	18-39 years	**0.46***	0.24	0.26	−0.07	0.26	(2,1,5)
	40-64 years	**0.50***	0.12	0.17	0.00	0.19	(3,1,2)
	≥65 years	**0.51***	−0.08	0.18	−0.02	0.20	(2,1,2)
Emergency department	All ages	**0.70***	0.17	**0.47***	−0.25	−0.15	(4,1,4)
	<2 years	**0.34**	0.21	0.18	−0.26	−0.16	(3,1,3)
	2-4 years	**0.66***	0.35	**0.58***	−0.14	−0.06	(0,1,4)
	5-12 years	**0.58***	0.32	**0.55***	−0.00	−0.38	(3,1,4)
	13-17 years	**0.58***	−0.10	0.32	−0.13	−0.18	(3,1,3)
	18-39 years	**0.59***	0.16	0.31	−0.30	−0.14	(2,1,3)
	40-64 years	**0.68***	0.30	0.10	−0.24	−0.31	(7,1,0)
	≥65 years	**0.39**	0.14	0.15	−0.18	−0.08	(1,1,0)

* Significant ($\alpha = 0.05$)

† All ARIMA models also included a first-order seasonal autoregressive term and two holiday variables

2001-2002

Subset		Correlation					ARIMA
Type of Setting	Age-Group	Lag: 0 weeks	Lag: 1 weeks	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	Model (p,d,q)†
Community-based setting	All ages	0.25	0.01	0.18	**0.38**	0.27	(2,1,2)
	<2 years	0.15	0.04	0.12	**0.50***	0.04	(2,1,4)
	2-4 years	0.01	−0.01	**0.50***	**0.50***	0.14	(0,1,2)
	5-12 years	−0.04	0.02	0.25	**0.31**	0.15	(2,1,2)
	13-17 years	−0.04	0.19	0.11	**0.40***	−0.01	(2,1,3)
	18-39 years	0.32	0.23	0.15	**0.36**	0.22	(2,1,5)
	40-64 years	**0.47***	0.11	0.06	0.19	0.23	(3,1,2)
	≥65 years	**0.52***	−0.09	0.02	0.11	0.15	(2,1,2)
Emergency department	All ages	0.36	**0.50***	**0.39***	0.05	0.01	(4,1,4)
	<2 years	**0.42***	0.33	0.22	0.12	−0.19	(3,1,3)
	2-4 years	0.38	**0.49***	0.27	0.32	0.17	(0,1,4)
	5-12 years	0.14	0.11	**0.54***	0.14	−0.01	(3,1,4)
	13-17 years	−0.08	**0.46***	0.21	0.14	−0.21	(3,1,3)
	18-39 years	**0.54***	**0.53***	−0.00	0.13	0.21	(2,1,3)
	40-64 years	0.32	**0.46***	0.32	−0.19	0.03	(7,1,0)
	≥65 years	**0.65***	0.19	0.15	−0.06	−0.06	(1,1,0)

* Significant ($\alpha = 0.05$)

† All ARIMA models also included a first-order seasonal autoregressive term and two holiday variables

2002-2003

Subset		Correlation					ARIMA
Type of Setting	Age-Group	Lag: 0 weeks	Lag: 1 weeks	Lag: 2 weeks	Lag: 3 weeks	Lag: 4 weeks	Model (p,d,q)[†]
Community-based setting	All ages	0.11	**0.22**	−0.04	**0.22**	0.13	(2,1,2)
	<2 years	−0.18	0.25	**0.35**	−0.07	−0.22	(2,1,4)
	2-4 years	0.21	**0.39**[*]	0.20	−0.03	0.01	(0,1,2)
	5-12 years	0.06	0.32	**0.41**[*]	0.02	−0.01	(2,1,2)
	13-17 years	−0.36	0.12	**0.72**[*]	0.01	−0.29	(2,1,3)
	18-39 years	0.23	0.14	−0.27	**0.26**	−0.09	(2,1,5)
	40-64 years	0.18[‡]	−0.05	−0.39	0.27	0.24	(3,1,2)
	≥65 years	−0.06[‡]	−0.12	−0.10	0.10	0.03	(2,1,2)
Emergency department	All ages	**0.58**[*]	0.01	0.27	−0.41	0.34	(4,1,4)
	<2 years	0.19	**0.39**[*]	0.17	−0.35	−0.00	(3,1,3)
	2-4 years	−0.03	0.20	**0.37**	0.01	0.18	(0,1,4)
	5-12 years	0.10	0.29	**0.45**[*]	0.01	0.08	(3,1,4)
	13-17 years	−0.10	0.33	**0.57**[*]	−0.05	−0.15	(3,1,3)
	18-39 years	**0.56**[*]	−0.01	−0.30	−0.13	0.23	(2,1,3)
	40-64 years	**0.71**[*]	−0.23	−0.20	−0.24	**0.39**[*]	(7,1,0)
	≥65 years	**0.56**[*]	−0.33	0.02	0.27	**0.47**[*]	(1,1,0)

[*] Significant ($\alpha = 0.05$)

[†] All ARIMA models also included a first-order seasonal autoregressive term and two holiday variables

[‡] Peak correlation was actually found at a lag of -1 week; it was significant for the 40-64 year old subsets but insignificant for the ≥65 year old subsets.

CHAPTER 6
Summary and Conclusion

We found that among physician billing data for ILI diagnoses, although there is some degree of year-to-year variation in optimal subsets and lead times each year, community setting visits by school-aged children of 5 to 12 and 13 to 17 years tended to be the most strongly correlated with P&I hospitalizations at the earliest lags (often at least a 2 week lead time) and therefore, they may be the earliest sentinels of influenza infection.

Our findings targeting pediatric age-groups are consistent with previous research. Like Sebastian et al. (2008) and Olson et al. (2007), our results point toward school-aged children, although others have found pre-school aged children to be the earliest indicators of an influenza epidemic (Brownstein et al., 2005; Lemay et al., 2008). These findings are compatible with the well established role of school-aged children as the primary vectors in the transmission of influenza in the community. This role likely reflects an interplay between their innate ability to shed the virus earlier and for a longer time compared to adults (Carrat et al., 2008; Frank et al., 1981; Hall et al., 1979), their extensive social contact patterns (Glass and Glass, 2008; Mikolajczyk et al., 2008; Mossong et al., 2008; Wallinga et al., 2006), as well as the behavioural patterns of concerned parents who tend to bring their ill children to a doctor earlier than adults would seek care themselves (Brownstein et al., 2005). Several studies, including one randomized controlled study, have found that when day care or school children are vaccinated, the incidence of ILI among their contacts, and in the general population is reduced (Hurwitz

et al., 2000; King et al., 2006; Piedra et al., 2005; Reichert et al., 2001). In fact, the evidence that has accumulated for the role of school-aged children in influenza epidemics has led the Advisory Committee on Immunization Practices (ACIP) of the Centers for Disease Control and Prevention (CDC) in the United States to recently add all children aged 5 to 18 years to their recommended target groups for annual influenza vaccination (Fiore et al., 2008). In previous seasons, the ACIP targeted only children 6 months to 4 years among children who are healthy. Canada has not yet adopted this expanded vaccination policy 200 (2008).

We additionally demonstrated that community health care settings have the potential to generate an earlier signal than the ED. A potential explanation for this finding may be that mild initial symptoms often do not require ED care (Heffernan et al., 2004) and therefore patients in the initial stages of illness may tend to seek primary care at community-based settings first (Lazarus et al., 2001). Although the potential of automated syndromic surveillance in primary care has previously been alluded to by a few (Lazarus et al., 2001; Smith et al., 2007), there has been little research evaluating its implementation in primary care (Lazarus et al., 2002; Marsden-Haug et al., 2007; Miller et al., 2004; Sloane et al., 2006; Smith et al., 2007; van den Wijngaard et al., 2008; Yang et al., 2008), and even fewer studies doing so using rigorous time series methodology (Yang et al., 2008). Rather, the ED setting (or no setting stratification) has predominated most studies evaluating automated syndromic surveillance (Lemay et al., 2008; Lober et al., 2003; Olson et al., 2007; Sebastian et al., 2008; Zheng et al., 2007). Our visit setting findings are novel since to our knowledge, no researcher has ever compared the timeliness of community-based and ED setting data for the syndromic surveillance of ILI using data drawn from a single source population to avoid potential confounders, while also simultaneously assessing the effect of patient age.

Only a few studies have conducted a year-by-year analysis for assessing the timeliness of syndromic data (Lemay et al., 2008; Quenel et al., 1994; Zheng et al., 2007), although none have compared different age and setting specific subsets simultaneously. Like in these other studies, our year-by-year analysis demonstrated annual variation, in both the optimal subsets by age and setting and in the lead times, which may be a consequence of the circulation of different strains each season. The timing of different subsets may depend on an interaction between age and influenza subtype factors. In particular, we have highlighted the impact of a re-emerged strain on the earlier presenting behaviour of children to community settings for ILI visits.

Methodologically, this study adds to a growing body of literature that has demonstrated the utility of time series methods to control for autocorrelation in the study of infectious disease. Ignoring autocorrelation while correlating two time series can lead to artificially inflated estimates of the correlation coefficient (Box and Newbold, 1971; Bowie and Prothero, 1981; Helfenstein, 1996). However, only a few studies (Brownstein et al., 2005; Lemay et al., 2008; Olson et al., 2007; Yang et al., 2008; Zheng et al., 2007) have similarly applied time series methods for the purpose of evaluating the timeliness of data sources for syndromic surveillance as we have and to our knowledge, none have used such methods to compare timeliness of subsets simultaneously restricted by age-group and visit setting.

In using heat maps to represent the cross-correlation function (CCF), it became clear to us that there are often other significant correlations surrounding the peak correlation. We believe the heat-maps facilitate a broader understanding of the patterns in the often complex relationship between two time series being correlated, and underscore the benefit of an *overall* examination rather than

focusing on just the single peak correlation in the CCF as has typically been done in the past.

With a data set covering more influenza seasons, future directions could include a further examination of the effect of different influenza subtypes on the timeliness of ILI visits for influenza surveillance, an examination of the contribution of respiratory syncytial virus (RSV, another major respiratory viral pathogen, especially among children), and other factors that are known to vary from year to year such as environmental effects.

In conclusion, our results provide compelling support for the implementation of syndromic surveillance in primary care, with a specific focus on school-aged children, as a beneficial complement to existing surveillance systems. However, annual variations in lead-lag relationships may also make it difficult to pinpoint one subset that would consistently be the earliest indicator of an influenza season each year. Along this note, the implication may be that surveillance for influenza probably should not focus on any single particular group but a combination of several potentially early indicators across different data sources. A better understanding of the lead-lag relationships between potential syndromic data sources and viral circulation in the context of annual variation would help improve the accuracy of infectious disease surveillance and forecasting systems as well as the planning of public health interventions such as vaccination and school closure policies.

CHAPTER 7
Bibliography

SM.13 Nombre de participants, nombre de services médicaux et leur coût par participant selon le sexe, le groupe d'âge et le type de service, rémunération à l'acte, médecine et chirurgie. Search Tool for Statistical Information (STSI). [Online Tables]. Régie de l'assurance maladie du Québec (RAMQ). [Accessed: Oct. 23, 2008]. Available from: http://www.ramq.gouv.qc.ca/fr/statistiques/index.shtml.

Statistiques Annuelles. Québec, Québec: Régie de l'Assurrance-Maladie du Québec (RAMQ). 1995.

(2000). Statement on influenza vaccination for the 2000-2001 season. An Advisory Committee Statement (ASC). National Advisory Committee on Immunization (NACI). *Can Commun Dis Rep*, 26:1–16.

(2002). Influenza in Canada: 2000-2001 season. *Can Commun Dis Rep*, 28(3):17–28.

(2005). Influenza in Canada: 2003-2004 season. *Can Commun Dis Rep*, 31(1):1–18.

(2008). Statement on influenza vaccination for the 2008-2009 season. An Advisory Committee Statement (ASC). National Advisory Committee on Immunization (NACI). *Can Commun Dis Rep*, 34:1–46.

Andersson, E., Bock, D., and Frisn, M. (2008). Modeling influenza incidence for the purpose of on-line monitoring. *Stat Methods Med Res*, 17(4):421–438.

Ashford, D. A., Kaiser, R. M., Bales, M. E., Shutt, K., Patrawalla, A., McShan, A., Tappero, J. W., Perkins, B. A., and Dannenberg, A. L. (2003). Planning against biological terrorism: lessons from outbreak investigations. *Emerg Infect Dis*, 9(5):515–519.

Barker, W. H. (1986). Excess pneumonia and influenza associated hospitalization during influenza epidemics in the united states, 1970-78. *Am J Public Health*, 76(7):761–765.

Bartlett, M. S. (1935). Some aspects of the time-correlation problem in regard to tests of significance. *J R Stat Soc*, 98(3):536–543.

Bazarian, J. J., Veazie, P., Mookerjee, S., and Lerner, E. B. (2006). Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes. *Acad Emerg Med*, 13(1):31–8.

Beitel, A. J., Olson, K. L., Reis, B. Y., and Mandl, K. D. (2004). Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. *Pediatr Emerg Care*, 20(6):355–60.

Birkhead, G., Chorba, T. L., Root, S., Klaucke, D. N., and Gibbs, N. J. (1991). Timeliness of national reporting of communicable diseases: the experience of the national electronic telecommunications system for surveillance. *Am J Public Health*, 81(10):1313–1315.

Boussard, E., Flahault, A., Vibert, J. F., and Valleron, A. J. (1996). Sentiweb: French communicable disease surveillance on the World Wide Web. *BMJ*, 313(7069):1381–2; discussion 1382–4.

Bowie, C. and Prothero, D. (1981). Finding causes of seasonal diseases using time series analysis. *Int J Epidemiol*, 10(1):87–92.

Box, G. and Jenkins, G. (1970). *Time series analysis: Forecasting and control*. Holden-Day, San Francisco.

Box, G. E. P. and Newbold, P. (1971). Some comments on a paper of Coen, Gomme and Kendall. *J R Stat Soc [Ser A]*, 134(2):229–240.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York, 2nd edition.

Brownstein, J. S., Kleinman, K. P., and Mandl, K. D. (2005). Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. *Am J Epidemiol*, 162(7):686–93.

Brownstein, J. S. and Mandl, K. D. (2008). Pediatric population size is associated with local timing and rate of influenza and other acute respiratory infections among adults. *Ann Emerg Med*.

Cadieux, G. and Tamblyn, R. (2008). Accuracy of physician billing claims for identifying acute respiratory infections in primary care. *Health Serv Res*, 43(6):2223–2238.

Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., and Valleron, A.-J. (2008). Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol*, 167(7):775–785.

Cauchemez, S., Valleron, A. J., Boelle, P. Y., Flahault, A., and Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750–4.

Centers for Disease Control and Prevention (CDC) (2000). Update: influenza activity–united states and worldwide, 1999-2000 season, and composition of the 2000-01 influenza vaccine. *MMWR Morb Mortal Wkly Rep*, 49(17):375–381.

Centers for Disease Control and Prevention (CDC) (2002). Update: Influenza activity–united states and worldwide, 2001-02 season, and composition of the 2002-03 influenza vaccine. *MMWR Morb Mortal Wkly Rep*, 51(23):503–506.

Chan, E. H. (2008). Evaluating the use of physician billing data for age and setting specific influenza surveillance. Master's thesis, McGill University.

Dato, V., Wagner, M. M., and Fapohunda, A. (2004). How outbreaks of infectious disease are detected: a review of surveillance systems and outbreaks. *Public Health Rep*, 119(5):464–471.

Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press, Oxford.

Esposito, S., Marchisio, P., Cavagna, R., Gironi, S., Bosis, S., Lambertini, L., Droghetti, R., and Principi, N. (2003). Effectiveness of influenza vaccination of children with recurrent respiratory tract infections in reducing respiratory-related morbidity within the households. *Vaccine*, 21(23):3162–3168.

Falsey, A. R., Hennessey, P. A., Formica, M. A., Cox, C., and Walsh, E. E. (2005). Respiratory syncytial virus infection in elderly and high-risk adults. *N Engl J Med*, 352(17):1749–1759.

Falsey, A. R. and Walsh, E. E. (2005). Respiratory syncytial virus infection in elderly adults. *Drugs Aging*, 22(7):577–587.

Fiore, A. E., Shay, D. K., Broder, K., Iskander, J. K., Uyeki, T. M., Mootrey, G., Bresee, J. S., and Cox, N. S. (2008). Prevention and control of influenza: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2008. *MMWR Recomm Rep*, 57(RR-7):1–60.

Fleischauer, A. T., Silk, B. J., Schumacher, M., Komatsu, K., Santana, S., Vaz, V., Wolfe, M., Hutwagner, L., Cono, J., Berkelman, R., and Treadwell, T. (2004). The validity of chief complaint and discharge diagnosis in emergency department-based syndromic surveillance. *Acad Emerg Med*, 11(12):1262–7.

Fleming, D. M. (2000). The contribution of influenza to combined acute respiratory infections, hospital admissions, and deaths in winter. *Commun Dis Public Health*, 3(1):32–8.

Fleming, D. M. and Cross, K. W. (1993). Respiratory syncytial virus or influenza? *Lancet*, 342(8886-8887):1507–1510.

Fleming, D. M., Pannell, R. S., Elliot, A. J., and Cross, K. W. (2005). Respiratory illness associated with influenza and respiratory syncytial virus infection. *Arch*

Dis Child, 90(7):741–746.

Fox, J. P., Hall, C. E., Cooney, M. K., and Foy, H. M. (1982). Influenzavirus infections in Seattle families, 1975-1979. I. Study design, methods and the occurrence of infections by time and age. *Am J Epidemiol*, 116(2):212–27.

Frank, A. L., Taber, L. H., Wells, C. R., Wells, J. M., Glezen, W. P., and Paredes, A. (1981). Patterns of shedding of myxoviruses and paramyxoviruses in children. *J Infect Dis*, 144(5):433–441.

Gill, P. W., Murphy, A. M., and Cunningham, A. L. (1991). Influenza a(h1n1): a widening spectrum? *Med J Aust*, 155(6):362–367.

Glass, L. M. and Glass, R. J. (2008). Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health*, 8:61.

Glezen, W. P. (1996). Emerging infections: Pandemic influenza. *Epidemiol Rev*, 18(1):64–76.

Golomb, M. R., Garg, B. P., Saha, C., and Williams, L. S. (2006). Accuracy and yield of ICD-9 codes for identifying children with ischemic stroke. *Neurology*, 67(11):2053–5.

Hall, C. B., Douglas, R. G., Geiman, J. M., and Meagher, M. P. (1979). Viral shedding patterns of children with influenza B infection. *J Infect Dis*, 140(4):610–613.

Hay, A. J., Gregory, V., Douglas, A. R., and Lin, Y. P. (2001). The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci*, 356(1416):1861–1870.

Heffernan, R., Mostashari, F., Das, D., Karpati, A., Kulldorff, M., and Weiss, D. (2004). Syndromic surveillance in public health practice, New York City. *Emerg Infect Dis*, 10(5):858–64.

Helfenstein, U. (1996). Box-Jenkins modelling in medical research. *Stat Methods Med Res*, 5(1):3–22.

Henning, K. J. (2004). What is syndromic surveillance? *MMWR Morb Mortal Wkly Rep*, 53 Suppl:5–11.

Hurwitz, E. S., Haber, M., Chang, A., Shope, T., Teo, S., Ginsberg, M., Waecker, N., and Cox, N. J. (2000). Effectiveness of influenza vaccination of day care children in reducing influenza-related morbidity among household contacts. *JAMA*, 284(13):1677–82.

Irvin, C. B., Nouhan, P. P., and Rice, K. (2003). Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. *Ann Emerg Med*, 41(4):447–52.

Ivanov, O., Gesteland, P. H., Hogan, W., Mundorff, M. B., and Wagner, M. M. (2003). Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. *AMIA Annu Symp Proc*, pages 318–322.

Jajosky, R. A. and Groseclose, S. L. (2004). Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health*, 4:29.

Jansen, A. G. S. C., Sanders, E. A. M., Wallinga, J., Groen, E. J., van Loon, A. M., Hoes, A. W., and Hak, E. (2008). Rate-difference method proved satisfactory in estimating the influenza burden in primary care visits. *J Clin Epidemiol*, 61(8):803–812.

John, R. S., Finlay, B., and Blair, C. (2001). Bioterrorism in Canada: An economic assessment of prevention and postattack response. *Can J Infect Dis*, 12(5):275–284.

King, J. C., Stoddard, J. J., Gaglani, M. J., Moore, K. A., Magder, L., McClure, E., Rubin, J. D., Englund, J. A., and Neuzil, K. (2006). Effectiveness of school-based

influenza vaccination. *N Engl J Med*, 355(24):2523–2532.

Koplan, J. (2001). CDC's strategic plan for bioterrorism preparedness and response. *Public Health Rep*, 116 Suppl 2:9–16.

Lazarus, R., Kleinman, K., Dashevsky, I., Adams, C., Kludt, P., DeMaria, A., J., and Platt, R. (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerg Infect Dis*, 8(8):753–60.

Lazarus, R., Kleinman, K. P., Dashevsky, I., DeMaria, A., and Platt, R. (2001). Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health*, 1:9.

Lee, M.-S., Walker, R. E., and Mendelman, P. M. (2005). Medical burden of respiratory syncytial virus and parainfluenza virus type 3 infection among US children. Implications for design of vaccine trials. *Hum Vaccin*, 1(1):6–11.

Lemay, R., Mawudeku, A., Shi, Y., Ruben, M., and Achonu, C. (2008). Syndromic surveillance for influenzalike illness. *Biosecur Bioterror*, 6(2):161–170.

Li, Y. (1999). 1998-1999 influenza season: Canadian laboratory diagnoses and strain characterization. *Can Commun Dis Rep*, 25(21):177–81.

Li, Y. (2000). 1999-2000 influenza season: Canadian laboratory diagnoses and strain characterization. *Can Commun Dis Rep*, 26(22):185–9.

Lober, W. B., Trigg, L. J., Karras, B. T., Bliss, D., Ciliberti, J., Stewart, L., and Duchin, J. S. (2003). Syndromic surveillance using automated collection of computerized discharge diagnoses. *J Urban Health*, 80(2 Suppl 1):i97–106.

Lombardo, J. S., Burkom, H., and Pavlin, J. (2004). ESSENCE II and the framework for evaluating syndromic surveillance systems. *MMWR Morb Mortal Wkly Rep*, 53 Suppl:159–165.

Macey, J. F., Tam, T. W., Li, Y., Winchester, B., and Zabchuk, P. (2003). Influenza in Canada: 2001-2002 season. *Can Commun Dis Rep*, 29(6):45–59.

Marsden-Haug, N., Foster, V. B., Gould, P. L., Elbert, E., Wang, H., and Pavlin, J. A. (2007). Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. *Emerg Infect Dis*, 13(2):207–16.

Mathur, U., Bentley, D. W., and Hall, C. B. (1980). Concurrent respiratory syncytial virus and influenza A infections in the institutionalized elderly and chronically ill. *Ann Intern Med*, 93(1):49–52.

McBean, A. M. and Hebert, P. L. (2004). New estimates of influenza-related pneumonia and influenza hospitalizations among the elderly. *Int J Infect Dis*, 8(4):227–235.

Menec, V. H., Black, C., MacWilliam, L., and Aoki, F. Y. (2003). The impact of influenza-associated respiratory illnesses on hospitalizations, physician visits, emergency room visits, and mortality. *Can J Public Health*, 94(1):59–63.

Mikolajczyk, R. T., Akmatov, M. K., Rastin, S., and Kretzschmar, M. (2008). Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiol Infect*, 136(6):813–822.

Miller, B., Kassenborg, H., Dunsmuir, W., Griffith, J., Hadidi, M., Nordin, J. D., and Danila, R. (2004). Syndromic surveillance for influenzalike illness in ambulatory care network. *Emerg Infect Dis*, 10(10):1806–11.

Molinari, N. A., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., and Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, 25(27):5086–96.

Monto, A. S. and Kioumehr, F. (1975). The Tecumseh Study of Respiratory Illness. IX. Occurence of influenza in the community, 1966–1971. *Am J Epidemiol*, 102(6):553–563.

Morse, S. S. (2007). Global infectious disease surveillance and health intelligence. *Health Aff (Millwood)*, 26(4):1069–1077.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5(3):e74.

Muscatello, D. J., Churches, T., Kaldor, J., Zheng, W., Chiu, C., Correll, P., and Jorm, L. (2005). An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. *BMC Public Health*, 5:141.

Neuzil, K. M., Mellen, B. G., Wright, P. F., Mitchel, E. F., J., and Griffin, M. R. (2000). The effect of influenza on hospitalizations, outpatient visits, and courses of antibiotics in children. *N Engl J Med*, 342(4):225–31.

Olson, D. R., Heffernan, R. T., Paladini, M., Konty, K., Weiss, D., and Mostashari, F. (2007). Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. *PLoS Med*, 4(8):e247.

Perrotta, D. M., Decker, M., and Glezen, W. P. (1985). Acute respiratory disease hospitalizations as a measure of impact of epidemic influenza. *Am J Epidemiol*, 122(3):468–476.

Piedra, P. A., Gaglani, M. J., Kozinetz, C. A., Herschler, G., Riggs, M., Griffith, M., Fewlass, C., Watts, M., Hessel, C., Cordova, J., and Glezen, W. P. (2005). Herd immunity in adults against influenza-related illnesses with use of the

trivalent-live attenuated influenza vaccine (CAIV-T) in children. *Vaccine*, 23(13):1540–1548.

Poehling, K. A., Edwards, K. M., Weinberg, G. A., Szilagyi, P., Staat, M. A., Iwane, M. K., Bridges, C. B., Grijalva, C. G., Zhu, Y., Bernstein, D. I., Herrera, G., Erdman, D., Hall, C. B., Seither, R., Griffin, M. R., and the New Vaccine Surveillance, N. (2006). The underrecognized burden of influenza in young children. *N Engl J Med*, 355(1):31–40.

Quenel, P., Dab, W., Hannoun, C., and Cohen, J. M. (1994). Sensitivity, specificity and predictive values of health service based indicators for the surveillance of influenza A epidemics. *Int J Epidemiol*, 23(4):849–55.

Reichert, T. A., Sugaya, N., Fedson, D. S., Glezen, W. P., Simonsen, L., and Tashiro, M. (2001). The Japanese experience with vaccinating schoolchildren against influenza. *N Engl J Med*, 344(12):889–96.

Reis, B. Y., Pagano, M., and Mandl, K. D. (2003). Using temporal context to improve biosurveillance. *Proc Natl Acad Sci U S A*, 100(4):1961–5.

Schanzer, D. L., Langley, J. M., and Tam, T. W. (2006). Hospitalization attributable to influenza and other viral respiratory illnesses in Canadian children. *Pediatr Infect Dis J*, 25(9):795–800.

Schneeweiss, S., Robicsek, A., Scranton, R., Zuckerman, D., and Solomon, D. H. (2007). Veteran's affairs hospital discharge databases coded serious bacterial infections accurately. *J Clin Epidemiol*, 60(4):397–409.

Sebastian, R., Skowronski, D. M., Chong, M., Dhaliwal, J., and Brownstein, J. S. (2008). Age-related trends in the timeliness and prediction of medical visits, hospitalizations and deaths due to pneumonia and influenza, British Columbia, Canada, 1998-2004. *Vaccine*, 26(10):1397–403.

Shaw, M. W., Xu, X., Li, Y., Normand, S., Ueki, R. T., Kunimoto, G. Y., Hall, H., Klimov, A., Cox, N. J., and Subbarao, K. (2002). Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000-2001 and 2001-2002 seasons. *Virology*, 303(1):1–8.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and its Applications: With R Examples*. Springer.

Simonsen, L., Clarke, M. J., Williamson, G. D., Stroup, D. F., Arden, N. H., and Schonberger, L. B. (1997). The impact of influenza epidemics on mortality: introducing a severity index. *Am J Public Health*, 87(12):1944–50.

Simonsen, L., Fukuda, K., Schonberger, L. B., and Cox, N. J. (2000). The impact of influenza epidemics on hospitalizations. *J Infect Dis*, 181(3):831–7.

Sloane, P. D., MacFarquhar, J. K., Sickbert-Bennett, E., Mitchell, C. M., Akers, R., Weber, D. J., and Howard, K. (2006). Syndromic surveillance for emerging infections in office practice using billing data. *Ann Fam Med*, 4(4):351–8.

Smith, G., Hippisley-Cox, J., Harcourt, S., Heaps, M., Painter, M., Porter, A., and Pringle, M. (2007). Developing a national primary care-based early warning system for health protection–a surveillance tool for the future? Analysis of routinely collected data. *J Public Health (Oxf)*, 29(1):75–82.

Suyama, J., Sztajnkrycer, M., Lindsell, C., Otten, E. J., Daniels, J. M., and Kressel, A. B. (2003). Surveillance of infectious disease occurrences in the community: an analysis of symptom presentation in the emergency department. *Acad Emerg Med*, 10(7):753–63.

Tamblyn, R., Abrahamowicz, M., Dauphinee, D., Wenghofer, E., Jacques, A., Klass, D., Smee, S., Blackmore, D., Winslade, N., Girard, N., Du Berger, R., Bartman, I., Buckeridge, D. L., and Hanley, J. A. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory

authorities. *JAMA*, 298(9):993–1001.

Upshur, R. E., Knight, K., and Goel, V. (1999). Time-series analysis of the relation between influenza virus and hospital admissions of the elderly in Ontario, Canada, for pneumonia, chronic lung disease, and congestive heart failure. *Am J Epidemiol*, 149(1):85–92.

van den Wijngaard, C., van Asten, L., van Pelt, W., Nagelkerke, N. J. D., Verheij, R., de Neeling, A. J., Dekkers, A., van der Sande, M. A. B., van Vliet, H., and Koopmans, M. P. G. (2008). Validation of syndromic surveillance for respiratory pathogen activity. *Emerg Infect Dis*, 14(6):917–925.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am J Epidemiol*, 164(10):936–944.

Wilchesky, M., Tamblyn, R. M., and Huang, A. (2004). Validation of diagnostic codes within medical services claims. *J Clin Epidemiol*, 57(2):131–41.

Yang, L., Wong, C. M., Lau, E. H. Y., Chan, K. P., Ou, C. Q., and Peiris, J. S. M. (2008). Synchrony of clinical and laboratory surveillance for influenza in Hong Kong. *PLoS ONE*, 3(1):e1399.

Zambon, M. C., Stockton, J. D., Clewley, J. P., and Fleming, D. M. (2001). Contribution of influenza and respiratory syncytial virus to community cases of influenza-like illness: an observational study. *Lancet*, 358(9291):1410–1416.

Zeger, S. L., Irizarry, R., and Peng, R. D. (2006). On time series analysis of public health and biomedical data. *Annu Rev Public Health*, 27:57–79.

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

APPENDIX A
Appendix A

Table A–1: Lead times obtained by studies comparing data from the ED, community setting, and mixed ambulatory settings to a P&I mortality standard.

Ages	Author and Year	Lead Time (weeks)	Specific Setting, Age-Group
ED			
all ages	Brownstein et al., 2005*	1.5	general ED
adults	Brownstein et al., 2005*	1.7	general ED, 18-39 yr
		−1.9	general ED, 40-64 yr
		0.1	general ED, > 64 yr
		1.7	adult ED, all adults
		1.0	adult ED, 18-39 yr
		0.6	adult ED, 40-64 yr
		1.6	adult ED, > 64 yr
		0.4	community ED, 18-39 yr
		3.0	community ED, 40-64 yr
		1.9	community ED, > 64 yr
children	Brownstein et al., 2005*	2.7	general ED, <3 yr
		3.3	general ED, 3-4 yr
		2.1	general ED, 5-10 yr
		4.6	general ED, 11-17 yr
		5.4	pediatric ED, all children
		5.0	pediatric ED, <3 yr
		7.1	pediatric ED, 3-4 yr
		5.7	pediatric ED, 5-10 yr
		4.0	pediatric ED, 11-17 yr
		3.0	community ED, <3 yr
		3.7	community ED, 3-4 yr
		1.7	community ED, 5-10 yr
		3.6	community ED, 11-17 yr
Community Settings			
all ages	Brownstein et al., 2005*	2.0	community ED
		4.1	ambulatory care
	Miller et al., 2004	1-2	ambulatory care
adults	Brownstein et al., 2005*	5.1	ambulatory care, 18-39 yr
		4.3	ambulatory care, 40-64 yr
		4.7	ambulatory care, > 64 yr
children	Brownstein et al., 2005*	5	ambulatory care, <3 yr
		5.3	ambulatory care, 3-4 yr
		3.6	ambulatory care, 5-10 yr
		3.6	ambulatory care, 11-17 yr
Mixed Ambulatory Care Settings			
all ages			
adults	Sebastian et al., 2008*	−0.3	20-49 yr
		−0.3	50-64 yr
		−0.8	≥65 yr
children	Sebastian et al., 2005*	−0.5	<6 mo
		−0.3	6-23 mo
		0	2-4 yr
		0.3	5-9 yr
		0.3	10-19 yr

* Original unit of measurement in the study was days

Table A–2: Lead times obtained by studies comparing data from the ED, community setting, and mixed ambulatory settings to a P&I hospitalizations standard.

Ages	Author and Year	Lead Time (weeks)	Specific Setting, Age-Group
ED			
all ages			
adults			
children	Ivanov et al., 2003*	1.1	<5 yr
Community Settings			
all ages	Lazarus et al., 2002	2	
	van den Wijngaard et al., 2008	−1	
adults			
children			
Mixed Ambulatory Care Settings			
all ages			
adults	Sebastian et al., 2008*	0.3	20-49 yr
		0.3	50-64 yr
		−0.3	≥65 yr
children	Sebastian et al., 2008*	0	<6 mo
		0.3	6-23 mo
		0.5	2-4 yr
		0.8	5-9 yr
		0.8	10-19 yr

* Original unit of measurement in the study was days

Table A–3: Lead times obtained by studies comparing data from the ED, community setting, and mixed ambulatory settings to an influenza virological gold standard.

Ages	Author and Year	Lead Time (weeks)	Specific Setting, Age-Group
ED			
all ages	Zheng et al., 2007*	0.4-2.6 (mean 1.1)	
adults	Olson et al., 2007	0	18-39 yr
		0	40-64 yr
		-1	≥ 65 yr
children	Olson et al., 2007	0	<2 yr
		1	2-4 yr
		1	5-12 yr
		1	13-17 yr
	Lemay et al., 2008	2-4 (mean 2.8)	<5 yr
Community Settings			
all ages	van den Wijngaard et al., 2008	1	influenza A
		2	influenza B
	Marsden-Haug et al., 2007	0	variety of ICD-9 codes for ILI
	Yang et al., 2008	4	GPs
		2	GOPCs
adults			
children			
Mixed Ambulatory Care Settings			
all ages			
adults			
children			

* Original unit of measurement in the study was days

Table A–4: Set of Régie de l'assurance maladie du Québec (RAMQ) establishment codes used in our ambulatory care categorization.

Code	Setting	Setting Type
000	private office without municipality number	community
6XX	private office with municipality number	community
512	private medical and/or dental clinic (with anaesthesia privilege)	community
54X	private medical clinic considered a general practice establishment in the context of particular medical activities	community
8X5	C.L.S.C.* : service point	community
9X2	C.L.S.C.*	community
0X7	emergency department	hospital ED

* Local community service centres providing health and social services in Quebec, Canada

APPENDIX B
Appendix B

McGill

Faculty of Medicine
3655 Promenade Sir William Osler
Montreal, QC H3G 1Y6

Faculté de médecine
3655, Promenade Sir William Osler
Montréal, QC, H3G 1Y6

Fax/Télécopieur: (514) 398-3595

April 28, 2008

Dr. David Buckeridge
Epidemiology and Biostatistics
1140 Pine Avenue West
Montreal, Quebec
H3A 1A3

Dear Dr. Buckeridge,

Thank you for submitting the research proposal entitled "Optimizing Fee-For-Service Physician Billing Data for Syndromic Surveillance of Influenza in Ambulatory Care Settings"

As this study involves no more than minimal risk and in accordance with Article 1.6 of the Canadian Tri-Council Policy Statement of Ethical Conduct for Research Involving Humans and U.S. Title 45 CFR 46, Section 110 (b), paragraph (1), we are pleased to inform you that approval for the study (April 2007) was provided via an expedited review by the Chair on April 28, 2008 valid until **April 2009**. The study proposal will be presented for corroborative approval at the next meeting of the Committee and a certification document will be issued to you at that time.

A review of all research involving human subjects is required on an annual basis in accord with the date of initial approval. The annual review should be submitted at least one month before **April 2009**. Should any modification to the study occur over the next twelve months, please advise IRB appropriately.

Yours sincerely,

Serge Gauthier, MD
Chair
Institutional Review Board

cc: A04-E13-08B

McGill

Faculty of Medicine
3655 Promenade Sir William Osler
Montreal, QC H3G 1Y6

Faculté de médecine
3655, Promenade Sir William Osler
Montréal, QC, H3G 1Y6

Fax/Télécopieur: (514) 398-3595

CERTIFICATION OF ETHICAL ACCEPTABILITY FOR RESEARCH INVOLVING HUMAN SUBJECTS

The Faculty of Medicine Institutional Review Board (IRB) is a registered University IRB working under the published guidelines of the Tri-Council Policy Statement, in compliance with the Plan d'action ministériel en éthique de la recherche et en intégrité scientifique, (MSSS, 1998) and the Food and Drugs Act (17 June 2001); and acts in accordance with the U.S. Code of Federal Regulations that govern research on human subjects. The IRB working procedures are consistent with internationally accepted principles of good clinical practice.

At a full Board meeting on April 28, 2008, the Faculty of Medicine Institutional Review Board, consisting of:

PAUL BRASSARD, MD	SERGE GAUTHIER, MD
VINCENT GRACCO, PHD	KATHERINE GRAY-DONALD, PHD
MARLYNNE GURSKY, BN, M.ED.	MARIGOLD HYDE, B.SC.
ROBERT L. MUNRO, BCL	SALLY TINGLEY, B.COM

Examined the research project **A04-E13-08B** entitled *Optimizing Fee-for-Service Physician Billing Data for Syndromic Surveillance of Influenza in Ambulatory Care Settings*

As proposed by: Dr. David Buckeridge (Applicant) to _____ (Granting Agency, if any)

And consider the experimental procedures to be acceptable on ethical grounds for research involving human subjects.

April 28, 2008
Date

Chair, IRB

Dean of Faculty

Institutional Review Board Assurance Number: FWA 00004545