Introduction

Even though health care provider reporting of diseases to public health authorities is common, often there is under-reporting by providers, including for notifiable diseases; frequently, under-reporting occurs by wide margins. Two causal factors for this under-reporting by providers have been that: (1) disclosing data may violate their patients’ privacy, and (2) disclosed data may be used to evaluate their performance. A reluctance to disclose information due to privacy concerns exists despite the U.S. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule permitting disclosures of personal health information (PHI) for public health purposes without patient authorization. On the other hand, such patient privacy concerns are somewhat justified: there have been documented breaches of patient information from public health data custodians.

A common way to address this privacy issue is to de-identify patient data before it is disclosed to public health. Worries about performance evaluation can be addressed by ensuring provider anonymity when data is collected, as well as by attesting that collected data will be held securely, not shared with third parties, and only used for pre-defined purposes (i.e., anonymity and confidentiality assurances).

In this paper we provide empirical and theoretical examinations illustrating practical challenges with current approaches to de-identification in the context, as an example, of syndromic surveillance, ensuring provider anonymity, and providing explicit confidentiality assurances. We then propose approaches for addressing these challenges.

The De-identification of Syndromic Surveillance Data

Routine monitoring of specific indicator variables can provide a valuable early warning of a disease outbreak. Such indicator variables come from pre-diagnostic data sources, such as OTC drug purchases, school absenteeism data, emergency department data, tele-health lines, and electronic medical record (EMR) data. The emergency department data source is the most common.

Emergency department syndromic surveillance systems are designed to be a leading indicator of time- and geography-dependent abnormal occurrences (i.e., frequency and total numbers) of health events. They can be used to alert responders early that an outbreak may be in progress, and can also be used to track the progress of an outbreak after it has been detected. In a recent syndromic surveillance project, individual level patient data was being collected from the emergency department (ED) at the Children’s Hospital of Eastern Ontario (CHEO) and sent to the municipal public health department. For meaningful analysis, it is important to know when a patient presents at the hospital and where they live. Given that one of the primary purposes of a syndromic surveillance system is to detect outbreaks early, knowing the exact date of
presentation is important. Syndromic surveillance data is transmitted to public health every few hours; therefore, it is easy to know the exact date of a patient visit to the ED.

To be able to determine if outbreaks are specific to a particular geography, postal code information is necessary. For example, detailed location information is needed to determine if a food-borne illness outbreak is localized to people living in a particular neighborhood, which may suggest the food source or a contamination site.

In our particular example, the fields requested by public health from the CHEO ED were the following: date of presentation, postal code, date of birth, and gender, as well as syndrome (e.g., gastrointestinal, constitutional, respiratory, rash, hemorrhagic, botulinic, or neurological) and severity codes. The latter two fields cannot be easily used to re-identify individual patients; therefore, the focus of the analysis was the first four fields. A de-identification analysis on 108,344 ED visits between June 2007 and June 2009 was performed to address patient privacy concerns by the hospital. For this analysis we used a risk threshold of 0.2, which is common in practice and has strong precedents. This means that 0.2 is the maximum probability of an individual record being re-identified on these four variables.

**Data De-identification**

To de-identify this data we used generalization and suppression. Generalization refers to reducing the precision of the fields in the data, and suppression refers to removing specific values on these four variables for some records. There is a trade-off between these two data transformations. We can retain more precision at the expense of higher suppression, for example. Dates are typically generalized by reducing the precision to month and year, year, and then year ranges. Postal codes are typically generalized by cropping the last set of characters.

Interpretable ways to assess data quality due to de-identification is to look at how much precision was lost through generalization and the percentage of records that have some suppression applied to them. The de-identification of this data using an optimization algorithm resulted in a number of alternative solutions summarized in Table 1. All of these solutions would retain the full date of presentation.

Solution 1 is the original data set. As can be seen, all of the records in the data set would need to have some suppression applied to them if the data is disclosed without any generalization. Solutions 4 and 5 would reduce the age of patients to 5- and 10-year intervals. Given that the data pertains to children with a limited age range, such generalizations would reduce variation, and hence data utility, considerably. Solutions 3 to 5 would retain only the first character of the postal code, which represents a very large geographic area that would not be useful for outbreak cluster detection.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of Possible De-Identification Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Solution</strong></td>
<td><strong>Precision of Variables</strong></td>
</tr>
<tr>
<td>Date of Presentation</td>
<td>Postal Code</td>
</tr>
<tr>
<td>1</td>
<td>day/month/year</td>
</tr>
<tr>
<td>2</td>
<td>day/month/year</td>
</tr>
<tr>
<td>3</td>
<td>day/month/year</td>
</tr>
<tr>
<td>4</td>
<td>day/month/year</td>
</tr>
<tr>
<td>5</td>
<td>day/month/year</td>
</tr>
</tbody>
</table>
The compromise solution is number 2, but it would also result in all of the records having some suppression applied to them. As can be seen from these results, it is difficult to retain both detailed information about when a visit occurred and location information about where the patient lives. Traditional de-identification methods will not be sufficient for some public health purposes, such as syndromic surveillance.

A Possible Solution
In our example, there was little freedom to transform the date of presentation; the best we can do is generalize the patient’s date of birth to year of birth (or age). However, instead of cropping postal codes, we can apply better methods to de-identify geospatial information. It has been demonstrated that other simple approaches such as adding noise to location information\(^10\) can be reverse engineered.\(^11\) A recent linear programming approach to choose optimal random perturbations to location information has been proposed.\(^12\) However, this will not scale to large data sets. An alternative approach that public health professionals should consider is clustering adjacent areas.\(^13\) Clustering has the potential to retain small areal units and to better ensure that privacy can be protected. Whether clustering will be sufficient to allow the disclosure of detailed date of presentation information for syndromic surveillance remains an empirical question.

Performance Measurement
Provider concern about performance measurement can manifest itself at the institutional level as well as the individual practitioner level. For example, Long Term Care Homes (LTCH) in Ontario do not have to report Antimicrobial Resistant Organism colonizations and infections among their residents to the public or to public health. Given that almost 60 percent LTCHs are privately owned\(^14\) and that home funding consists of a mixture of government support and monthly resident fees, this may also increase resistance to sharing ARO rates. Similarly, individual practitioners may not want their screening rates for particular diseases and conditions to be known if it will affect their fees or require them to gain additional educational credits. To alleviate these concerns, public health can collect data anonymously and provide strong confidentiality assurances to the providers (that the data will be handled responsibly and not shared with others). Recent protocols have been developed that would guarantee both.\(^15\) However, ensuring anonymity and providing strong confidentiality assurances may also have negative consequences on the quality of data collected. Empirical results in behavioral economics illustrate why.

Anonymity Concerns
Previous studies demonstrated that while individuals who remained anonymous in a task were sometimes able to report more socially undesirable behaviors, they were also less accurate reporters of their behavior.\(^16\) Further, Nina Mazar et al.\(^17\) demonstrate that when given the opportunity to behave dishonestly (i.e., by shredding test papers before reporting a score), study participants are more likely to take advantage of the situation and inflate their test scores. This work suggests that when reporting is anonymous, individuals might be particularly vulnerable to misreporting their behavior — ironically, when they have no chance of being rewarded for more socially appropriate behavior.\(^18\) Finally, in the domain of online shopping,\(^19\) when shoppers have individual identities (versus remaining anonymous), they have greater levels of trust, and possibly, accountability of actions. As such, one consequence of guaranteed anonymity is the possibility of reduced accountability of responses.

The option that seems most promising for reasonable response rates — and most accurate data from providers — is to maintain anonymity and have good practices in place to protect confidentiality, but not present this information in an overt fashion. For example, one could make available a detailed privacy policy or disclosure notice on a website, without showcasing this information to participants. If policies remain strong, but less focal, then response rates and accuracy of responses are likely to remain unaffected, without compromising the protection of respondents.
Confidentiality Assurances
When collecting sensitive information, intuition suggests that additional confidentiality notices may increase response rates and disclosure, and individuals indeed report that they would be more willing to disclose sensitive information if data gathering and use is regulated. One systematic review demonstrated that confidentiality assurances offered a small positive effect for increased response rates. However, more recent research suggests that informing individuals of the strength of security protecting their data actually counter-intuitively and ironically encourages them to withhold more information than when those securities are not mentioned. For example, John, Acquisti, and Loewenstein demonstrate that in circumstances when confidentiality is guaranteed, individuals are more inclined to withhold sensitive information, such as whether they have cheated on a spouse. However, when privacy and data protection practices are not mentioned, suggesting that no policy for security of data is in place, individuals are more inclined to give information freely. Similarly, individuals given assurances of confidentiality are less willing to complete a questionnaire than those receiving no assurance, suggesting that response rates might be influenced by the presence of an explicit privacy policy.

Individuals may thus see a privacy policy as a warning instead of seeing it as a notice of safety, which may induce skepticism and careful assessment of the potentially unsafe environment. As a consequence, individuals “clam up” and fail to take advantage of a safe outlet for information sharing.

A Possible Way Forward
Hypothetical outcomes given different levels of anonymity and confidentiality assurances are provided in Table 2. The option that seems most promising for reasonable response rates — and most accurate data from providers — is to maintain anonymity and have good practices in place to protect confidentiality, but not present this information in an overt fashion. For example, one could make available a detailed privacy policy or disclosure notice on a website, without explicitly showcasing this information to participants. If policies remain strong, but less focal, then response rates and accuracy of responses are likely to remain unaffected, without compromising the protection of respondents.

One additional variable for consideration is baseline levels of trust and disclosure. Without any information with regard to anonymity or confidentiality, a provider might base disclosure on their trust in the organization collecting the data. As such, when trust in an organization is low (for example, if the public health unit has a poor reputation regarding privacy, or public health in general has a poor reputation due to well publicized data breaches), disclosure in general will be low; in such cases, assurances of confidentiality may encourage greater disclosure. In this scenario of low trust, we might predict a reversal of Table 2: response rates and accuracy may be highest when anonymity and confidentiality assurances are high, whereas response rates and accuracy may be lowest when anonymity and confidentiality assurances are low or not mentioned. This possibility highlights the important role of trust in data disclosures to public health, and how it can mediate the impact of anonymity and confidentiality assurances.

Conclusions
In this paper we examined two contemporary problems faced by public health professionals in collecting data from health care providers. Previous work has indicated that providers want to ensure that their patients cannot be re-identified in data they disclose, and that providers have concerns that their performance may be evaluated from the patient data they disclose. New methods for de-identifying geographic information will allow useful de-identified data to be disclosed to public health. Ensuring provider anony-
ity and having strong data handling practices, but not communicating that explicitly and in detail to providers when data is collected, will likely ensure a higher response rate with more accurate data. However, this must be accompanied by steps to grow trust between the providers and public health.

Note
The analysis on the emergency department data set was approved by the research ethics board of the Children's Hospital of Eastern Ontario Research Institute.

References
5. See El Emam, supra note 2.
8. See El Emam, (2009), id.
9. Id.
15. See El Emam, supra note 2.