Using administrative medical claims data to supplement state disease registry systems for reporting zoonotic infections

Stephen G Jones,1 Steven Coulter,2 William Conner3

ABSTRACT
Objective To determine what, if any, opportunity exists in using administrative medical claims data for supplemental reporting to the state infectious disease registry system.

Materials and methods Cases of five tick-borne (Lyme disease (LD), babesiosis, ehrlichiosis, Rocky Mountain spotted fever (RMSF), tularemia) and two mosquito-borne diseases (West Nile virus, La Crosse viral encephalitis) reported to the Tennessee Department of Health during 2000—2009 were selected for study. Similarly, medically diagnosed cases from a Tennessee-based managed care organization (MCO) claims data warehouse were extracted for the same time period. MCO and Tennessee Department of Health incidence rates were compared using a complete randomized block design within a general linear mixed model to measure potential supplemental reporting opportunity.

Results MCO LD incidence was 7.7 times higher (p<0.001) than that reported to the state, possibly indicating significant under-reporting (~196 unreported cases per year). MCO data also suggest about 33 cases of RMSF go unreported each year in Tennessee (p<0.001). Three cases of babesiosis were discovered using claims data, a significant finding as this disease was only recently confirmed in Tennessee.

Discussion Data sharing between MCOs and health departments for vaccine information already exists (eg, the Vaccine Safety Datalink Rapid Cycle Analysis project). There may be a significant opportunity in Tennessee to supplement the current passive infectious disease reporting system with administrative claims data, particularly for LD and RMSF.

Conclusions There are limitations with administrative claims data, but health plans may help bridge data gaps and support the federal administration’s vision of combining public and private data into one source.

BACKGROUND AND SIGNIFICANCE
Notifiable diseases are infectious diseases which require regular, frequent, and timely reporting of individual diagnosed cases to aid in prevention and control (eg, Lyme disease (LD), giardiasis, salmonella).1 2 Public health officials from state health departments collaborate annually with the Centers for Disease Control and Prevention (CDC) to determine which diseases should be listed. Reporting of disease cases by healthcare providers and laboratories is currently mandated only at the state level and therefore can vary from state to state.2 3 In 2009, over 7000 cases of notifiable communicable diseases were reported to the Tennessee Department of Health (TDH) Communicable and Environmental Disease Services.4 Although state regulations or contractual obligations may require the reporting of certain diseases, traditional passive reporting by the diagnosing clinician can be burdensome, incomplete, and delayed.5 Thus, there is under-reporting of diseases as not all diagnosed or suspected cases are reported by healthcare providers.5–10 and reporting can vary by physician specialty.11 Many healthcare providers may not understand the importance of public health surveillance, and generally how, when, why, and what to report.5 9

With the recent passing of the Health Information Technology for Economic and Clinical Health (HITECH) Act and the Patient Protection and Affordable Care Act, there is a great opportunity to modernize healthcare delivery and improve the quality of care received in the USA. Through the HITECH Act, the Assistant Secretary for Planning and Evaluation for the Department of Health and Human Services is building a multi-payer, multi-claim database to support patient-centered research using public and private payer claims data. Combining public and private data into one source may alleviate certain problems surrounding existing disparate data sources and provide the ability to study less common conditions.12 California is already planning for the integration of public and private data through their State Health Information Exchange Cooperative Agreement Program. A major initiative for them is to ‘transmit and integrate data across multiple internal and external data sources’ in order to ‘prepare for and respond to emergencies, diseases, outbreaks, epidemics, and emerging threats.’13

Health insurance plans (ie, managed care organizations, MCOs) could play a major role in the reporting of infectious diseases.8 14 Kaiser Permanente is one such entity already engaged in collaborative disease reporting with the CDC through their Vaccine Safety Datalink (VSD) Rapid Cycle Analysis (RCA) project. Their goal is to monitor potential adverse events (side effects) following vaccinations in near real time. Section 164.512(b) of the federal Health Information Portability Accountability Act (HIPAA) establishes safeguards allowing health plans to report such data. It states that all healthcare providers, health plans, and clearinghouses are permitted to disclose protected health information without individual authorization for the reporting of disease and vital events, and the conducting of public health surveillance, investigations, and interventions.15
Medical encounter data are recorded within the healthcare system every time a patient visits their doctor or hospital for a medical service, fills a prescription medicine, or seeks a consultation from a clinician. Considering that more than 253 million Americans have health insurance and will most likely utilize that service when needed, administrative data captured in medical encounters could serve as a useful data supplement to the current disease reporting system. Zoonotic diseases are infectious diseases that can be transmitted from or through animals to humans, and arthropods often act as vectors for transmission. Zoonotic diseases are of significant concern regarding public health and account for approximately 75% of recently emerging infectious diseases; approximately 60% of all human pathogens originate from animals. The aim of our study was to determine what, if any, opportunity may exist for supplementing the state infectious disease registry system with administrative medical claims data. We measure opportunity by comparing disease incidence rates across two data sources for five tick-borne (LD, babesiosis, ehrlichiosis, Rocky Mountain spotted fever (RMSF), tularemia) and two mosquito-borne (West Nile virus (WNV), La Crosse viral encephalitis) diseases known to occur in the southeastern USA. Any measurable differences between the two systems will indicate an opportunity for potential data sharing efforts.

MATERIALS AND METHODS

Case definition

Disease occurrence data within the proposed study area of Tennessee were collected from two separate data sources and tested for statistical differences. The first data source was administrative medical claims data obtained from BlueCross BlueShield of Tennessee, a large southeastern MCO located in Tennessee. The second data source was an extract provided by the TDH Center for Environmental and Communicable Diseases. A specific and separate case definition was needed for each data source due to the origin of the data.

Case definition—MCO

The participating MCO insures approximately 50% of the entire state’s population. For the purposes of this study, cases are defined as all medical claims filed to the MCO having a primary or secondary arthropod-borne disease diagnosis code of interest with at least three separate corroborating events, using the member’s first recorded occurrence (see below). Claims were extracted for the January 1, 2000—December 31, 2009 time period using the exact ICD-9 diagnosis code for disease identification (eg, 066.4: West Nile virus), not the more generic root level diagnosis available to providers (066: Other arthropod-borne viral diseases). In this manner, the minimum (ie, lower limit of) disease incidence according to MCO data is estimated. Tick-borne diseases include:

- Babesiosis (ICD-9 code: 088.32)
- Borreliosis—LD (ICD-9 code: 083.81)
- Ehrlichiosis—human monocytic ehrlichiosis (HME) (ICD-9 code: 082.41)
- Rickettsiosis—RMSF (ICD-9 code: 082.0)
- Tularemia (ICD-9 code: 021)
- Mosquito-borne diseases include:
  - La Crosse viral encephalitis (ICD-9 code: 062.5)
  - West Nile virus (WNV) (ICD-9 code: 066.4)

These diseases were selected because they represent a growing public health concern, are known to occur in the southeastern USA, and provide a mix of relatively common cases (eg, borreliosis, rickettsiosis) and very rare cases (eg, babesiosis, tularemia) for study. Using the MCO database, any patient receiving medical services for one of the selected diseases prior to the start of the study period (January 1, 2000) or after the study period (December 31, 2009) was removed from the analysis. To best ensure the diagnosis for the patient was in fact their first diagnosis, only the first recorded diagnosis date for each patient was retained. Any subsequent claims for the patient were removed and not considered in this study. To improve data validity, only claims with at least three instances of corroborating evidence were retained; here, corroborations occurred if the same diagnosis code in question was mentioned in separate line items on the same claim and/or on any follow-up claims. (A ‘line item’ is a term used to describe the structure of data storage on a medical claim. For every clinical procedure performed on a patient within a given claim, each procedure and an accompanying medical diagnosis will be recorded in separate line items.) For example, if a claim had only two line items with a primary diagnosis code of LD, this claim was not considered for study. This analysis utilized the exact diagnosis code for disease identification and allowed the minimum (ie, lower limit of) incidence according to MCO data to be estimated. Institutional Review Board approval was not needed for this study because only administrative claims data were examined.

Case definition—TDH

The second data source was an extract provided by the TDH, Center for Environmental and Communicable Diseases detailing all notifiable ‘confirmed’ or ‘probable’ disease cases reported to the State of Tennessee during the study period according to the CDC guidelines. For example, the CDC case definitions for LD are as follows:

- Confirmed case—A case of erythema migrans (‘bulls-eye’ rash) with known exposure, or a case of erythema migrans with laboratory evidence of infection and without a known exposure, or a case with at least one late manifestation that has laboratory evidence of infection.
- Probable case—Any other case of physician-diagnosed LD that has laboratory evidence of infection.

This resulted in a comparison of medically diagnosed cases to CDC confirmed or probable disease cases. For example, a medical diagnosis for LD should be based on an individual’s history of possible exposure to ticks that carry LD, the presence of typical signs and symptoms, and the results of blood tests. Data from different sources can be compared because specific ICD-9 diagnosis codes and not generic root-level codes were used to extract medical claims. Thus we assumed if a clinician codes at this detail, they have evidence to support the diagnosis beyond suspicion. The pathogens of interest are ‘[not] likely to be diagnosed clinically without confirmatory testing, - either [providers] make the right diagnosis with lab confirmation or they never make the diagnosis at all’ (Laura Cooley, MD, infectious disease specialist, personal communication, November 8, 2010). Furthermore, this very difference and ambiguity is the main focus of the study, which is to compare the difference between state reported incidence rates and medically diagnosed case rates to determine if claims data could support the current surveillance system. Because TDH serves as the compiler of all data sources to the state level, these data represent a theoretically complete set of reported cases for the state under the assumption that no under-reporting exists.

Statistical analyses

To estimate if and to what extent under-reporting of notifiable diseases exists, a randomized control block design was employed within a generalized linear mixed model (GLMM) approach to
compare TDH and MCO case counts. For the purposes of this study, the degree to which under-reporting is present is used to measure the amount of ‘opportunity’ that may exist in using administrative claims data to supplement the state registry system. GLMM models are particularly useful in estimating trends in disease rates where the response variable is not necessarily normally distributed.\textsuperscript{19} \textsuperscript{20} Input values into the models included a yearly (n=10) county (n=95) case total, which produced 950 observations for each data source. Separate models were built for each disease, and the response variable of interest was disease counts assumed to be Poisson distributed with a log-transformed population count as an exposure offset. Disease counts were expected to vary by county (ie, spatial heterogeneity) due to varying population denominators, socio-economic factors, and geographic and habitat characteristics.\textsuperscript{21–24} Therefore, county was used as a blocking factor to remove the expected county-to-county variability when TDH values were compared with MCO values. Space (county) was considered a random effect, while time (year), data source (MCO vs TDH), and a time\times data source interaction were considered fixed effects. Fixed effects were examined for statistical significance using the F test with a \( \alpha \) level of 0.05. Variability in case counts across counties was tested using covariance tests.

**RESULTS**

Approximately 58 million medical claims were filed to the MCO during the 2000–2009 study period. Of these, 3535 claims contained a primary or secondary diagnosis for one of the seven described arthropod-borne diseases. After removal of invalid claims (patients without at least three separate ICD-9 entries for the disease, patients with claim dates starting or ending outside of the time period, duplicate patient entries, and patients having non-unique disease coding issues such as codes for RMSF and LD on the same claim), 1654 unique case-persons were distributed across the seven diseases of interest and remained for study. The majority of disease cases were LD (n=903; 55%), followed by RMSF (n=661; 40%). The remaining five diseases made up the residual 5% of disease cases. Three cases of babesiosis were found within the MCO claims data, specifically in Davidson, Lincoln, and Washington Counties.

Results from the GLMM suggest only LD and RMSF values are statistically different, as none of the other models converged (table 1). The average yearly numbers of medically diagnosed cases of LD from MCO data were 7.7 times higher than those reported to the state (\( F=83.44; \ p<0.0001 \)). LD rates varied significantly over the 10-year study period (\( F=2.08; \ p=0.0283 \)) and there was a significant temporal interaction with year\times data source (\( F=2.84; \ p=0.0026 \)). Based on the residual pseudo-likelihood, a test of covariance suggests there is significant spatial variation in LD cases across the state (\( \chi^2=54.8; \ p<0.0001 \)). The average yearly numbers of medically diagnosed cases of RMSF from MCO data were 1.24 times higher than those reported to the state (\( F=14.45; \ p=0.0001 \)). RMSF disease rates significantly varied over the 10-year study period (\( F=14.52; \ p<0.0001 \)) and there was a significant temporal interaction with year\times data source (\( F=10.14; \ p<0.0001 \)). There was also significant spatial variation in RMSF cases across the state (\( \chi^2=1135.01; \ p<0.0001 \)).

Gender distributions and the median ages of cases from MCO and TDH data were similar for LD and RMSF (tables 2 and 3). (Babesiosis was excluded as no TDH cases were recorded.)

Temporal trending indicates the aforementioned per 100 000 population rate differences varied from year to year (table 1), and MCO rates were not consistently higher throughout the study period (figure 1). LD rates from MCO data were consistently higher than TDH rates. TDH data indicated no evidence of babesiosis, but MCO data showed three separate cases in the years 2004, 2005, and 2009. RMSF rates from TDH data were lower than MCO rates for 2000–2005 and 2008, but were higher for 2006, 2007 and 2009 (figure 1).

**DISCUSSION**

Administrative medical claims data are an important resource for the research and surveillance of chronic diseases.\textsuperscript{25} Our results suggest there may be an opportunity to supplement the current disease reporting system with administrative medical claims data. We make no attempts to quantify the number of cases shared between the data because our intent is to compare differences in disease rates at a population level, not an individual level. Thus at a minimum, the number of potential cases not reported to the state is the difference between the MCO rate and the TDH rate; therefore assuming a 100% overlap between MCO and TDH data, we identify the smallest possible benefit of utilizing MCO data. Although our personal interpretation, this work was recently presented to senior TDH and the TDH rate; therefore assuming a 100% overlap between MCO and TDH data, we identify the smallest possible benefit of utilizing MCO data. Although our personal interpretation, this work was recently presented to senior TDH

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Gender distribution across disease cases and data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCO</strong></td>
<td><strong>TDH</strong></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td><strong>Female</strong></td>
</tr>
<tr>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>LD</td>
<td>372 (44.1%)</td>
</tr>
<tr>
<td>RMSF</td>
<td>348 (52.6%)</td>
</tr>
<tr>
<td>HME</td>
<td>7 (50.0%)</td>
</tr>
<tr>
<td>Tularemia</td>
<td>26 (61.9%)</td>
</tr>
<tr>
<td>LACV</td>
<td>9 (90.0%)</td>
</tr>
<tr>
<td>WNV</td>
<td>10 (47.8%)</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>772 (46.8%)</td>
</tr>
</tbody>
</table>

*Six TDH cases had unknown gender.

HME, human monocytic ehrlichiosis; LACV, La Crosse viral encephalitis; LD, Lyme disease; MCO, managed care organization; TDH, Tennessee Department of Health; RMSF, Rocky Mountain spotted fever; WNV, West Nile virus.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Median age of recorded cases across disease categories and data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCO</strong></td>
<td><strong>TDH</strong></td>
</tr>
<tr>
<td><strong>Rocky Mountain spotted fever</strong></td>
<td>36</td>
</tr>
<tr>
<td><strong>Human monocytic ehrlichiosis</strong></td>
<td>36</td>
</tr>
<tr>
<td><strong>Tularemia</strong></td>
<td>47</td>
</tr>
<tr>
<td><strong>La Crosse viral encephalitis</strong></td>
<td>8</td>
</tr>
<tr>
<td><strong>West Nile virus</strong></td>
<td>61</td>
</tr>
</tbody>
</table>

MCO, managed care organization; TDH, Tennessee Department of Health.
Figure 1  Temporal comparison of Lyme disease (top) and Rocky Mountain spotted fever (bottom) incidence rates using managed care organization (MCO) medical claims data and Tennessee Department of Health (TDH) reported data.
epidemiologists/physicians who showed great enthusiasm and a desire to implement this type of data sharing. They recognize the importance of data enhancements and we are currently in discussions on how to best proceed. Kaiser Permanente and the CDC have laid the groundwork to engage in data sharing projects, and with great success having produced more than 75 scientific articles since the inception of its VSD program in 1990.26 We envision a similar collaboration, providing data elements to the TDH such as ICD-9 diagnosis and CPT4-procedural codes and patient identifiers.

The authors understand the challenges that data sharing presents and the expected opposition by some to using claims data for case identification. We reduce the potential for claims error by including only exact diagnosis coding as opposed to root level coding, requiring at least three separate corroborations and retaining only the first recorded diagnosis. Admittedly, we are limited by the inability to relate, via a patient identifier, a medically diagnosed case to a CDC defined ‘confirmed’ or ‘probable’ case. However, it must be noted that, similar to MCO diagnosed cases, cases reported to the state are not necessarily laboratory confirmed either. According to the CDC guidelines, laboratory confirmation of LD is not necessary if the patient manifests with erythema migrans and the classic symptoms of the disease. Further, the sensitivity of ELISA testing on culture manifests with erythema migrans and the classic symptoms of the disease. Additionally in 2010, the TDH acknowledged this report of possible infection as accurate.30 Claims data suggest that one of the three cases also occurred in 2009, so it is possible, although not definite, that the Mosites et al finding38 and the MCO claims findings are the same. Mosites et al38 are now attempting to identify animal reservoir hosts and tick vectors. Data from the MCO could aid in this effort, and suggest at least two other cases occurred prior to this discovery.

As previously mentioned, a limitation in this particular study is the known tendency of some physicians to over-diagnose LD31 32 particularly in non-endemic areas. While this could be an anomaly peculiar to LD, it also emphasizes the need for coordinated information sharing between health plans and the state health department. Additionally, if health plans contributed to the current reporting system using medical claims filings, the state could potentially investigate these claims for confirmation. These investigations could alleviate some of the concerns surrounding over- and mis-diagnosis of certain zoonotic diseases. Other limitations of this study include the inability to empirically filter out claims in error. Even though the data were filtered to include only cases with at least three line items, claims coding errors are still possible. However, administrative data have proven to be very reliable as a source of disease identification when compared to medical chart reviews.40 41 We are further limited by the inability to relate, via a patient identifier, a medically diagnosed case based on claims data to a CDC defined ‘confirmed’ or ‘probable’ case. A patient could be coded with LD in the MCO claim system without necessarily having a laboratory confirmed case without laboratory confirmation. Additionally, the Mosites et al finding38 are now attempting to identify animal reservoir hosts and tick vectors. Data from the MCO could aid in this effort, and suggest at least two other cases occurred prior to this discovery.

CONCLUSIONS

Data sharing to improve disease surveillance is not without challenges. However, significant opportunity may exist in Tennessee to supplement the current passive reporting system with administrative medical claims data, particularly for LD and RMSF and other rare infectious diseases. There are known limitations with using administrative claims data, but health plans may help bridge data gaps as well as support the federal
administration’s vision of combining public and private data into one source. Additionally, the benefit to the public health system should take precedence over any potential data discrepancies at the individual level. State and local public health officials rely on healthcare providers, laboratories, and other public health personnel to report the occurrence of notifiable diseases to state and local health departments.53 Missing from this system are health plans and the data they could provide to state and national surveillance efforts. Without such data, trends may not be accurately monitored, unusual occurrences of diseases might not be detected, and the effectiveness of intervention activities cannot be easily evaluated.

Acknowledgments We wish to thank BlueCross BlueShield of Tennessee (employer of authors SJ and SC) and the Tennessee Department of Health for their willingness to engage in this project, especially Dr Inga Himelright, Chief Medical Officer, for her expertise in infectious diseases and surveillance, and Dr Tim Jones, Tennessee State Epidemiologist, for his fervor to implement this system. We thank Dr James Rieck, Department of Mathematical Sciences, Clemson University, for his statistical expertise. We also recognize Dr Martin Kuldoff, Department of Population Medicine Harvard Medical School and Harvard Pilgrim Health Care Institute, for his support related to space-time analysis for this and subsequent related work.

Contributors SJ conceived of and designed the study and analyzed the data. SJ, SC, and WE contributed to the writing and review of the paper. Funding This is Technical Contribution number 6000 of the Clemson University Experiment and is based upon work supported by NIFA/USDA, under project number SC-1700424. This work was funded in-kind through BlueCross BlueShield of Tennessee and Clemson University.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES