Journal of Biomedical Informatics 53 (2015) 15–26

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A method for detecting and characterizing outbreaks of infectious disease from clinical reports

Gregory F. Cooper *, Ricardo Villamarin, Fu-Chiang (Rich) Tsui, Nicholas Millett, Jeremy U. Espino, Michael M. Wagner

Real-time Outbreak and Disease Surveillance (RODS) Laboratory, Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Boulevard, Pittsburgh, PA 15206-3701, USA

ARTICLE INFO

Article history:
Received 5 April 2014
Accepted 22 August 2014
Available online 30 August 2014

Keywords:
Infectious disease
Outbreak detection
Outbreak characterization
Clinical reports
Bayesian modeling

ABSTRACT

Outbreaks of infectious disease can pose a significant threat to human health. Thus, detecting and characterizing outbreaks quickly and accurately remains an important problem. This paper describes a Bayesian framework that links clinical diagnosis of individuals in a population to epidemiological modeling of disease outbreaks in the population. Computer-based diagnosis of individuals who seek healthcare is used to guide the search for epidemiological models of population disease that explain the pattern of diagnoses well. We applied this framework to develop a system that detects influenza outbreaks from emergency department (ED) reports. The system diagnoses influenza in individuals probabilistically from evidence in ED reports that are extracted using natural language processing. These diagnoses guide the search for epidemiological models of influenza that explain the pattern of diagnoses well. Those epidemiological models with a high posterior probability determine the most likely outbreaks of specific diseases; the models are also used to characterize properties of an outbreak, such as its expected peak day and estimated size. We evaluated the method using both simulated data and data from a real influenza outbreak. The results provide support that the approach can detect and characterize outbreaks early and well enough to be valuable. We describe several extensions to the approach that appear promising.

1. Introduction

There remains a significant need for computational methods that can rapidly and accurately detect and characterize new outbreaks of disease. In a cover letter for the July 2012 "National Strategy for Biosurveillance" report, President Obama wrote: *As we saw during the H1N1 influenza pandemic of 2009, decision makers—from the president to local officials—need accurate and timely information in order to develop the effective responses that save lives* [1]. The report itself calls for "situational awareness that informs decision making" and innovative methods to "forecast that which we cannot yet prove so that timely decisions can be made to save lives and reduce impact." The report echoes a call made by Ferguson in 2006 in *Nature* for similar forecasting capabilities [2].

The current paper describes a Bayesian method for detecting and characterizing infectious disease outbreaks. The method is part of an overall framework for probabilistic disease surveillance that we have developed [3], which seeks to improve situational awareness and forecasting of the future course of epidemics. As depicted in Fig. 1, the framework supports disease surveillance end-to-end, from patient data to outbreak detection and characterization. Moreover, since detection and characterization are probabilistic, they can serve as input to a decision-theoretic decision-support system that aids public-health decision making about disease-control interventions, as we describe in [3].

In the approach, a case detection system (CDS) obtains patient data (evidence) from electronic medical records (EMRs) [4]. The patient data include symptoms and signs extracted by a natural language processing (NLP) system from text reports. CDS uses data about the patient and probabilistic diagnostic knowledge in the form of Bayesian networks [5] to infer a probability distribution over the diseases that a patient may have. For a given patient-case j, the result of this inference is expressed as likelihoods of the patient's data E_j, both with and without an outbreak disease dx. In a recently reported study, CDS achieved an area under the ROC curve of 0.75 (95% CI: 0.69 to 0.82) in identifying influenza cases from findings in ED reports [6].

A second component of the system, which is the focus of this paper, is the outbreak detection and characterization system (ODS). ODS receives from CDS the likelihoods of monitored

* Corresponding author. Postal address: The Offices at Baum, Suite 524, 5607 Baum Boulevard, Pittsburgh, PA 15206-3701, USA.
E-mail address: gfc@pitt.edu (G.F. Cooper).

http://dx.doi.org/10.1016/j.jbi.2014.08.011
1532-0464/

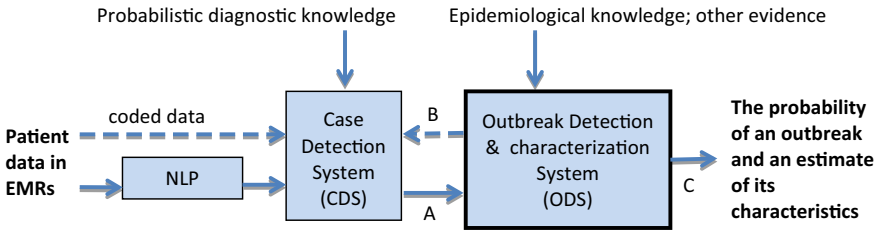

Fig. 1. Schematic of the probabilistic disease surveillance system. CDS transmits to ODS the likelihoods of each patient's findings, given the diseases being monitored (see arc A). ODS computes the probabilities of the epidemic models that were found during its model search. From these models, ODS can compute the probability of an outbreak, as well as estimate outbreak characteristics, such the outbreak size. For each of the monitored diseases, ODS also computes the prior probability that the next patient has that disease; it passes this information to CDS to use in deriving the posterior probability distribution over the diseases for that patient (see arc B). Thus, in an iterative, back-and-forth fashion, diagnostic information on past patients supports outbreak detection, and outbreak detection supports diagnosis of the next patient. This paper focuses on ODS and arcs A and C in the figure.

diseases for all patients over time. ODS searches a space of possible epidemic models that fit the likelihoods well, and it computes the probability of each model, denoted as $P(epidemic\ model_i \mid data_{all})$. The distribution over these epidemic models can be used to detect, characterize, and predict the future course of disease outbreaks. The output of ODS may be used to inform decisions about disease control interventions.

For each day, ODS also computes a prior probability that a patient seen on that day will have disease *dx*. To do so, ODS uses its estimate of (1) the extent of *dx* in the population, and (2) the fraction of people in the population with *dx* who will seek medical care. These ODS-derived patient priors can be used by CDS to compute the posterior probability that patient *j* has disease *dx*, that is, $P(dx \mid data_j)$. The probability that a patient has a disease can inform clinical decisions about treatment and testing for that patient, public health case finding, and public health disease reporting.

We previously described the overall disease surveillance system architecture shown in Fig. 1, including a high-level description of ODS [3]. The purpose of the current paper is to provide a detailed mathematical description of the current ODS model and inference methodology, as well as an initial evaluation of it using data from a real outbreak and from simulated disease outbreaks. The paper focuses on epidemiologic applications of ODS, which includes all the information flowing from left to right that are shown with solid arrows in Fig. 1.

2. Background

Outbreak detection and characterization (OD&C) is a process that detects the existence of an outbreak and estimates the number of cases and other characteristics, which can guide the application of control measures to prevent additional cases [7]. In this section, we review representative prior work on OD&C algorithms, and we describe the novel characteristics of our approach.

Non-Bayesian OD&C algorithms can be classified as temporal [8–15], spatial [16–22], or spatio-temporal [23]. Almost all of these approaches follow a frequentist paradigm and share a key limitation: they only compute a *p value* (or something related to it) of a monitored signal; given the signal, they do not derive the posterior probability that there is an outbreak of disease *dx*, which is what decision makers typically need. It is also difficult for frequentist approaches to incorporate many types of prior epidemiological knowledge about disease outbreaks.

Bayesian algorithms have been developed for outbreak detection [24–39]. These algorithms can derive the posterior probabilities of disease outbreaks, which are needed in setting alerting thresholds and performing decision analyses to inform public-health decision-making. Bayesian algorithms have also been developed to perform some types of outbreak characterization [31,38,40,41]. However, all of these algorithms have a major limitation: the evidence they receive as input is constrained to be counts, such as the daily number of patients presenting to outpatient clinics with symptoms of cough and fever. Although such counts are informative about outbreaks, they cannot feasibly express many rich sources of information, such as that found in a patient's emergency department (ED) report, which includes a mix of history, symptoms, signs, and lab information.

In the current paper, we describe a more flexible and general approach that models probabilistically the available evidence using data likelihoods, such as the probability of the findings in a patient's ED report conditioned on the patient having influenza (or alternatively some other disease). This approach can use counts as evidence, but it is not limited to doing so. It leverages the intrinsic synergy between individual patient diagnosis and population OD&C. In particular, in this approach OD&C is derived based on probabilistic patient diagnostic assessments, expressed as likelihoods. In general, the more informative is available patient evidence about the diseases being monitored, the more informative are the resulting probabilities of those diseases. For example, evidence that a patient has a fever, cough, and several other symptoms consistent with influenza will generally increase the probability of influenza in that patient, relative to having evidence regarding only one symptom, such as cough. The higher those probabilities (if well calibrated), the more informed the OD&C method will be about which patients have the outbreak disease, which in turn supports the detection and characterization of the outbreak in the population. In general, it is desirable to be able to incorporate whatever evidence happens to be available for each individual patient (including symptoms, signs, and laboratory tests) as early as possible in order to support outbreak detection and characterization. The method described in this paper provides such flexibility and generality.

In addition, the diagnosis of a newly arriving patient is influenced by prior probabilities that are derived from probabilistic inference over current OD&C models. To our knowledge, no prior research (either Bayesian or non-Bayesian) has (1) used a rich set of clinical information in EMR records as evidence in performing disease outbreak detection and characterization, nor (2) taken an integrated approach to patient diagnosis and population OD&C. While the power of this synergy is intuitive, the contribution of this paper is in describing a concrete approach for how to realize it computationally. In addition, we evaluate the approach.

Beyond being able to use a variety of evidence, the approach we propose can be applied with many different types of disease outbreak models. In the current paper we investigate the use of SEIR (Susceptible, Exposed, Infectious, and Recovered) compartmental models that use difference equations to capture the dynamics of contagious disease outbreaks, which is a highly relevant and important class of outbreak diseases in public health [42]. *SEIR models* have been extensively developed and applied to model