Using big data for improving two surveillance systems: influenza surveillance using Google flu-related search query data and probationers absconding surveillance using chronological case notes data

Jialiang Liu

Department of Biostatistics and Epidemiology

University of North Texas Health Science Center

Dissertation Committee:

Sumihiro Suzuki, PhD (Major advisor)

Eun-Young Mun, PhD

Rajesh Nandy, PhD

Contents

Chapter 1 Introduction

1.1. Introduction: Influenza Surveillance

Each year, the incidence of seasonal influenza (flu) and its financial costs are substantial in the Unite State (U.S.) every year. Flu causes major economic burden due to hospitalization and absenteeism form work and school. The Centers for Disease Control and Prevention (CDC) indicates that approximately 25 million people in the U.S. were infected with flu during the 2015-2016 flu season, leading to 11 million flu-related medical visits, and 12,000 flu-associated deaths.[1] Timeliness in detecting the onset of flu season is a critical component of flu surveillance to delay the spread of the disease and mitigate its adverse consequences.[2] The earlier we can detect a flu season onset, the more time we have to plan and implement proactive prevention strategies against the spread of the disease.[2] Preparation for preventing and controlling diseases, including seasonal flu, is not a quick and simple process.[3] The time required for the planning and implementation of a detailed and comprehensive plan for managing seasonal flu takes anywhere from a few days to even weeks.[3] The current gold standard of flu surveillance as practiced by the CDC includes reporting an onset of flu season whenever flu activity levels exceed a predetermined epidemic threshold.[4,5] However, as the flu activity is estimated based on clinical data, there is always a delay of up to three weeks between the occurrence of flu season onset and dissemination of this information.[5] That is, using the current gold standard of flu surveillance, we are only able to detect the onset after flu season has already begun. Thus, there is an urgent need for improving and strengthening the seasonal flu surveillance system to provide timely information of flu season onset for guiding public health decisions that seek to prevent and control the disease. To this end, the first goal of this dissertation was to test an innovative strategy that applies a statistical detection algorithm to the near real-time seasonal flu activity data to predict the onset of flu season weeks prior to flu season beginning.

The crucial first component of our innovative strategy for improving flu surveillance is the availability of real-time or at least, near real-time flu activity data. Studies suggest that internet-based information, such as volume of online search queries on flu-related topics may serve as novel, convenient, and cost-effective data sources for providing near real-time information of flu activity to be in complement, or even in lieu of, traditional flu information.[6-9] The Pew Internet and American Life Project reported that 80 percent of American internet users indicated that they accessed health-related information through internet searches.[10] Specifically, Google search query volume of specific flu-related search terms has attracted the most research interest.[11-14] Recently, Yang and colleagues developed a statistical model called AutoRegression with General Online (ARGO) data model which can be used to accurately estimate flu activity using Google search query data.[15,16] The aggregated Google search query data are publicly available on a near real-time basis, the resulting ARGO flu activity estimates are far superior and pragmatic for flu season onset detection compared to the current gold standard of flu activity reporting.

The second critical component of our strategy for improving flu surveillance is applying change point detection (CPD) methods to real-time flu activity data obtained from the ARGO model. Change point detection is a class of statistical methods in sequential analysis applied to time series data, e.g., stock market data, to determine a point in time when the distribution of the series is different before and after.[17,18] CPD methods are classified as “offline” or “online”.[17] The online CPD methods search for change points concurrently as data become available. Therefore, online CPD methods are ideal for early detection of flu season onset due to their ability of identifying change points in real-time. In this dissertation, we are proposing to apply a Bayesian online change point detection (BOCPD) method to the real-time data. However, because the BOCPD algorithm requires the specification of necessary conditions for assumptions, there is a need to determine the best way to specify these conditions to make the BOCPD algorithm more robust and practical when applying to for flu surveillance. Another barrier of direct application of BOCPD method for flu surveillance is it lacks a systematic way to determine informative change points that may signal the onset of flu season. Thus**,** we are proposing to modify the BOCPD method to expand its application to flu surveillance. We hypothesize that by applying the modified BOCPD method to the ARGO real-time estimated flu activity data, we will be able to create a surveillance system that will allow for the prediction of an imminent onset of flu season with enough lead time for public health officials to take appropriate actions to prevent and control the spread of the disease. Our goal to create a more practical flu surveillance system was accomplished through the following Aim.

Aim 1: To apply the modified Bayesian online change point detection (BOCPD) algorithm to real-time flu activity data obtained from the ARGO model to create a new surveillance system that will provide early detection of the onset of flu season. The number of weeks prior to the actual flu season beginning was determined. A systematic way to satisfy the necessary conditions for assumptions of the BOCPD algorithm was developed. A systematic approach to determine informative change points that may signal the onset of flu season was established.

1.2. Introduction: Probationers Absconding Surveillance

Probation is a court-order period of correctional supervision in the community, generally the most widely used alternative sanction to incarceration for qualifying offenders.[19] Probationers can maintain their normal lives in the community if they abide by certain conditions of probation.[19] If a probationer fails to comply with all required conditions, the court may revoke probation and require the probationer to serve a prison sentence.[20] Despite the opportunity for avoiding incarceration, there is a significant segment of offenders who are sentenced to probation fail to complete probation by absconding from supervision.[19] According to the Annual Probation Survey, approximately 10% of probationers abscond from supervision each year.[21] Based on the report from the U.S. Department of Justice, 16% of offenders sentenced to probation were re-arrested for committing new crimes during their period of supervision. It is reasonable to assume that the re-arrest rate would be even higher if we include the probationers who abscond from probation supervision. However, due to the limited financial resources and the increasing population of probationers, little effort has been made toward locating and examining these probation absconders.[19] The current knowledge about absconders may be insufficient to prevent the occurrence of probation absconding. There is a scarcity of research on probation absconders. Limited studies have examined risk factors associated with probation absconding and have found associations with demographic characteristics, substance use, offense types, and offender risk scores.[22]

However, in most probation systems, much of the detailed information about each probationer are written in text form as *chronological case notes*, which are part of the standard record keeping procedures. These case notes are electronically recorded by probation officers in an ad hoc fashion and describe their interactions with the probationers.[23] The case notes contain details about their meeting and notes on probationer's behaviors, compliance with conditions, and violations while in custody or under supervision management.[23,24] These records may be a rich source for important yet previously untapped information that could be used to discover novel and meaningful knowledge that can be used to prevent probation absconding. Nonetheless, due to the free-text nature of these case notes, using case notes as source of data is challenging. A possible solution is to apply natural language processing (NLP) techniques to the chronologic case notes data. NLP is a field of study that deals with the comprehension and analysis of human-produced texts.[25-27] Utilizing NLP techniques allows the researchers to extract meaningful information that can be used to generate critical knowledge buried in the text documents.[28] Thus, we hypothesize that by applying NLP to chronological case notes we will be able to discover knowledge that may help to reduce the number of probationers who abscond. Our goal to discover novel and meaningful knowledge about probationers who abscond supervision was accomplished through the following Aim.

Aim 2: To apply natural language processing (NLP) techniques to the chronological case notes of adult probationers in Tarrant County, Texas to discover hidden risk factors related to probationers absconding from supervision. To our knowledge, no study has ever applied NLP techniques to the case note data to explore content associated with probation absconders and completers. Factors that discriminate between probation absconders and completers were identified.

Chapter 2 Literature Review

2.1. Overview

Surveillance is a critical foundation of effective public health practice. The key objective of public health surveillance is to provide valid information to guide appropriate public health actions in a timely manner to contribute to the health of a population.[29] Reasons for conducting public health surveillance include the need to monitor trends and patterns of disease activity (e.g., influenza, dengue, and specific cancers), track health-related behaviors (e.g., illicit drug use, violence, and injuries), establish public health priorities, and facilitate prevention program planning and management.[2,30,31] However, many current surveillance methods and procedures in public health have a variety of deficiencies, such as out-of-date and possible vulnerable technology, lack of real-time surveillance systems that impact the ability to effectively detect and control threats in public health.[32,33] With the advancement of technology and availability of various sources of data, there should be opportunities for substantial improvement in these areas. This dissertation focused on two important types of surveillance: influenza surveillance and probationers absconding surveillance.

As data become more and more ubiquitous every day, different kinds of data are constantly generated from a plurality of sources, including web-based searches, social media, electronic health records, occupational injury records from Occupational Safety and Health Administration (OSHA) system, records from probation system, and many more.[34-37] Such data that have large volumes, velocity, and/or variety are commonly referred to as big data.[38] These data can take on many forms: structured or unstructured, quantitative or qualitative, collected near real-time or retrospectively, etc.[38] Big data analytics exhibits abundant potential to support a wide range of public health functions, such as tracking infectious diseases outbreaks, monitoring health-related behavior, and improving planning and delivery of public health interventions.[38-40] We believed that taking advantage of big data and associated technologies will benefit current public health surveillance systems. This dissertation consists of two topics with application of big data, one is to create an innovative flu surveillance system using Google flu-related search queries data, the other is to create a new surveillance system of probationers absconding from supervision using data generated from chronological case notes written by probation officers.

2.2. Literature Review: Influenza Surveillance

Influenza (flu) is an acute viral disease that mainly impacts the respiratory system.[41] The World Health Organization (WHO) estimates that worldwide 5-15% of the population is affected by influenza each year, with between three and five million cases of severe illness.[41,42] Flu has become one of the leading causes of death, where up to 56,000 are killed each year.[42] Health and economic burden of seasonal flu are substantial. The Centers for Disease Control and Prevention (CDC) indicated that approximately 25 million people in the United States (U.S.) were infected with influenza during the 2015-2016 flu season, leading to 11 million flu-related medical visits, 12,000 flu-associated deaths, and over $10 billion in medical costs, lost productivity, and lost wages.[43] Although seasonal flu happens every year, the timing of onset is different from season to season. Timeliness in detecting the onset of flu season plays a critical role in flu surveillance to delay the spread and control the impact of the disease. Early knowledge about the onset of flu season allows health officials to properly prepare prevention strategies, such as reinforcing flu prevention messages to the general public to increase vaccination rates. Early uptake of flu vaccinations is critical in preventing the spread of the flu because flu vaccines are not fully effective until about 2 weeks after the shot.[44] In addition, early detection can help health administrators make optimal staffing and medical resourcing decisions in preparation for potential surges of patient visits to hospital facilities. Preparation for preventing and controlling diseases, including seasonal flu, is not a quick and simple process. The time required for the planning and implementation of a detailed and comprehensive strategy for managing a flu season takes anywhere from a few days to even weeks. The earlier the detection of the potential onset of flu season, the more time there is to implement proactive prevention strategies against the spread of the disease. Presently, the gold standard of flu surveillance in U.S. is the one practiced by the CDC. CDC is responsible for flu surveillance using the following key indicators from laboratories and clinics: percent of patient visits for influenza-like illness (ILI), percent of respiratory specimens testing positive for flu viruses, rate of influenza-associated hospitalizations, and percent of deaths resulting from pneumonia or influenza. CDC uses this information to determine the onset of flu season and the types of flu viruses that are circulating, detect non-seasonal flu viruses (e.g., H1N1), and measure the impact of flu on hospitalization and deaths. CDC reports the onset of flu season whenever flu activity levels exceed a predetermined epidemic threshold[4]. Among the many indicators the CDC collects, the onset of flu season is determined based on clinical data. However, due to the time needed to process clinical information and report it to the CDC, there is always a delay between the actual time of flu season onset and public dissemination of this information. That is, using the current gold standard of flu surveillance, there is no way to monitor flu activity in real-time fashion, and an onset can only be detected after flu season has already begun. Furthermore, the current flu surveillance system only captures information from patients who are seen by health care providers, and hence, it cannot capture information from patients who have the flu but do not go to healthcare providers, which most likely leads to an underestimation of flu activity. Thus, there is an urgent need for improving and strengthening the seasonal flu surveillance system to provide timely flu season onset information for guiding public health decisions that seek to prevent and control the disease. In particular, there needs to a surveillance system, that can monitor flu activity in real or near real time and can utilize this real-time information to detect the onset of flu season prior to flu season beginning.

To this end, the crucial first step of developing a more practical flu surveillance system is the availability of real-time, or at least near real-time flu activity data. In recent decades, the rapid expansion of the internet has dramatically changed how people search for information, especially about health-related information. An increasing amount of health-related information has become available on Web sites[9]. The Pew Internet and American Life Project reported that 80 percent of American internet users indicated that they accessed health-related information using internet search engines (e.g., Google, Yahoo, etc.).[10] Thus, the frequency of online search queries may provide information regarding disease activity, e.g., seasonal flu activity. Because internet-based information can be obtained in real-time, studies have suggested that volume of queries from online search engines on flu-related topics may serve as a novel, convenient, and cost-effective way to track flu activity.[6-9] Unfortunately, there is no existing surveillance system that utilizes online search query data to monitor trend of flu activity in U.S. In previous versions, Google launched the Google Flu Trends (GFT) service, which attempted to track flu activity by monitoring and analyzing health care seeking behavior in the form of queries to its online search engine. However, the GFT algorithm for flu activity estimation was found lacking in reliability and accuracy.[6,45] As a result, this service was shut down in August of 2015.[46] Recently, Yang and colleagues developed a statistical model called AutoRegression with General Online (ARGO) data model which can be used to accurately estimate real-time flu activity using Google search queries data.[15,16] As the arrogated results of the Google search query data are publicly available on a near real-time basis, the resulting flu activity estimates are pragmatic for flu season onset detection compared to the gold standard of flu activity reporting.

Once real-time flu activity data are available, the second critical component of our innovative strategy for improving flu surveillance is to apply a statistical method that can detect the onset of flu season prior to flu season beginning when applied to the real-time data. In each season, flu activity starts to elevate gradually leading to the onset of flu season. Specifically, CDC identifies the onset of flu season as the first week when the percentage of patients seeking medical attention with ILI symptoms is at or above a predetermined epidemic baseline for two consecutive weeks.[4] As flu activity tends to grow gradually, the initial uptick in %ILI data should be an indicator for the eventual onset of flu season. Change point detection (CPD) is a statistical method that can be used to identify such initial uptick in the data. CPD is a class of statistical methods in sequential analysis applied to time series data, e.g., CDC's weekly ILI data, to determine a point in time when the distribution of the series is different before and after, i.e. point where the distribution changes.[17,18] A data point separates two different distributions of the time series is known as a change point. We posit that in general, CDC's identified date of flu season onset is usually not a change point. Rather, the change point most likely occurs prior, when there is an initial uptick in flu activity. Then the change point identified would provide early detection for imminent flu season onset. CPD methods are classified as either "offline" or "online".[17] Offline CPD methods identify the change point retrospectively, after the entire sequence of data is observed, which makes meaningless for flu surveillance, because the flu season onset can only be detected after the fact. In contrast, online CPD methods search for change points concurrently as data become available. Therefore, online CPD methods are ideal for detection of flu season onset due to their capability of identifying change points in a real-time, or at least near real-time fashion.

In this dissertation, we applied a Bayesian online change point detection (BOCPD) method to real-time flu activity data obtained from the ARGO model to create a flu surveillance system that can provide early detection of the imminent flu season onset. The BOCPD algorithm uses posterior probabilities to determine change points, where the posterior probabilities are updated at each time point given newly observed data. In particular, posterior probabilities are calculated for *run length,* which is the number of observations since the most recent change point (actual change point and not a detected change point). Because a change point can be detected at any time, run length is a random variable. At each time point, posterior probabilities are calculated for all possible run lengths, where the support is defined to be all positive integers less than or equal to the current sample size. This method makes two assumptions to calculate the posterior probabilities for run lengths. First, the distribution of the data and its parameters must be specified a priori. Second, because only one new datum is observed at each time point, given the run length at each time point, the run length at the next time point can either grows by 1 if no change point occurs or resets to 0 if a change point occurs at this time. The probability of these two events must also be specified a priori. These two assumptions play a crucial role in the process and they are applied at every time point no matter how long the process continues. Thus, the practical application of the BOCPD algorithm to detect flu season onset would require determining the best way to specify the conditions of these assumptions. Because the historical flu activity data collected by the CDC's flu surveillance system are available, we incorporated these data to develop a systematic way to satisfy the necessary conditions for the aforementioned assumptions. Another barrier of direct application of BOCPD method for flu surveillance is there is no systematic way to determine informative change points that may signal the onset of flu season. During the change point detection process, there may be multiple change points detected over time, some of which may be false detections or uninformative change points. There is need to develop a systematic way to determine change points that can provide early signal of flu season onset.

2.2. Literature Review: Probationers Absconding Surveillance

Based on the degrees of severity and control, the criminal justice system encompasses a variety of punishments for criminal behaviors.[20,47] Depending on the types and risk level of crimes that the offender has committed, probation can be used as an alternative to incarceration.[20,47] Probation is a court-ordered period of correctional supervision in the community, generally as an alternative to incarceration for qualifying offenders.[20] As the most widely used alternative sanction to incarceration, probation provides an opportunity for offenders to avoid prisons and the criminal environment contained therein.[19] Probationers can maintain their normal lives in the community if they abide by certain conditions of probation and report regularly to an appointed probation officer.[20] General conditions of probation may include participating in rehabilitation programs, submitting to drug and alcohol tests and maintaining employment.[20] If a probationer fails to comply with all required conditions, the court may revoke probation and require the probationer to serve a prison sentence.[20,21] Despite the opportunity for avoiding prison, there is a significant segment of offenders sentenced to probation who fail to complete probation by absconding from supervision.[21] By the end of 2016, about 4.5 million adults were under some form of criminal justice supervision in the community in the U.S.[21] Approximately 10% of probationers abscond from supervision each year.[21] According to the report from the U.S. Department of Justice, 16% of offenders sentenced to probation were re-arrested from committing a new crime during their period of supervision.[21] It is reasonable to assume that the re-arrest rate would be even higher if the probationers who abscond from probation supervision are included. Crime is a serious public health issue. Criminal activities can carry heavy health consequences for victims, including physical injury and mental health difficulties.[48] Absconders pose potential threats to community safety, as they may be reengaged in criminal activities, that are going undetected.[19] locating absconders is a critical component to prevent criminal activities and improve community safety.

However, locating absconders is made difficult by limited financial resources and increasing population of probationers.[19] Moreover, there is scarcity of research on absconders. The current knowledge about absconders may be insufficient to prevent the occurrence of probation absconding. Without adequately addressing probation absconders, a relatively risky population, public safety risks will remain high. Limited studies have examined risk factors associated with probation absconding and have found associations with demographic characteristics, offense types, and offender risk scores.[19,22] However, much of the detailed information about each probationer are written in text form as chronological case notes, which are part of the standard record keeping procedures for most probation systems.[49] Probation officers record case notes every time they meet with probationers. These case notes are electronically recorded and describe the interactions between the probation officers and the probationers during their meetings.[23] The interactions can be from formally scheduled meetings or from unscheduled meetings when probationers show up unannounced.[23,24] In addition, case notes also contain details about probationer's behaviors, compliance with conditions, and violations while in custody or under supervision management.[23,24] These case notes may be a rich source of previously untapped information that could be used to discover novel and meaningful knowledge about offenders who abscond from probation supervision. Key words and phrases in the case notes may be linked to the knowledge that are critical to probation officers to have a better understanding of factors associated with probation absconding and strengthen probation system. However, using case notes as data is challenging. The first challenge is the free-text nature of this type of data. The case notes are rather unstructured because the contents documented in case notes are based on conversations between the officer and probationer. Conventional statistical methods cannot be directly applied to data in text form. In addition, there is a large amount (in the hundreds of thousands) of case notes corresponding to the growing probationer population. Trying to make sense of them in an ad hoc fashion is generally not feasible. Therefore, there is a need to apply more complex statistical methods that can incorporate words, phrases, and their meaning to determine additional risk factors for probation absconding. To this end, applying natural language processing techniques may be a possible solution to analyzing the chronological case notes more efficiently and systematically.

Natural language processing (NLP) is a field of study that deals with the comprehension and analysis of human-produced texts.[25-27] By utilizing NLP techniques, we may be able to extract meaningful information in terms of key words and/or phrases that can be used to discover critical knowledge buried in the text documents.[28] NLP techniques have been successfully applied in many domains, such as surveillance in occupational injury and illness, improving medical diagnosis, etc.[28,50] To our knowledge, no study has applied NLP techniques to chronological case notes to explore contents associated with probation absconders and completers. By applying NLP to case notes of probationers, we may be able to shed light on previously untapped risk factors for probation absconders and generate useful knowledge that may help to reduce the number of probationers who abscond. The potential major contributions of this topic included identifying previously unknow commonalities, i.e., key words and phrases, in the case note of absconders and completers as well as contributing to a new surveillance system that uses case notes systematically to prevent probation absconding.

Chapter 3 Methods

3.1. Overview of study design

The goal of this study was to create innovative strategies for improving surveillance for two public health issues: influenza and probationers absconding. For Aim 1, the approach for improving seasonal flu surveillance was to apply a modified Bayesian online change point detection (BOCPD) algorithm to real-time flu activity data. We used data that represent the percentages of patients seeking medical attention with influenza-like illness (ILI) symptoms, which were estimated by applying the ARGO model to Google flu-related search query volume data. For Aim 2, our strategy for improving probation absconding surveillance was to apply natural language processing (NLP) techniques to data generated from chronological case notes to identify commonalities in contents associated with probation absconders.

Aim 1: To apply the modified Bayesian online change point detection (BOCPD) algorithm to real-time flu activity data obtained from the ARGO model to create a new surveillance system that will provide early detection of the onset of flu season.

3.2. Methods of Aim 1

Our strategy for improving seasonal flu surveillance involved two components: (1) using a data source that can monitor flu activity real-time; (2) using a statistical method that can detect the onset of flu season prior to flu season beginning.

3.2.1. Estimation of real-time flu activity by applying the ARGO model to the Google flu-related search query volume data and the historical CDC %ILI data

3.2.1.1. Data sources

The ARGO model used to estimate real-time flu activity required data from two sources: (a) Google flu-related search query volume data and (b) the historical CDC %ILI data.

a. Selection of Google flu-related search terms

The flu-related search terms, e.g., “flu duration,” “symptoms of flu,” “flu versus cold,” “flu care,” etc. were identified from Google Correlate (www.google.com/trends/correlate), which is a Google tool that provides the top 100 most highly correlated search terms given a user specified time series data, e.g., CDC %ILI data.[51] Previous studies indicated that internet search behavior was different during the 2009 influenza virus A (H1N1) pandemic. Internet users tended to search for different flu-correlated search terms to seek information after the H1H1 outbreak. Thus, to capture the most commonly used flu-related Google search terms, as done in Yang et al.,[15] two sets of top 100 highly flu-correlated search terms were used to estimate flu activity before and after H1N1 period.[52] The first set prior to the H1N1 pandemic was identified using CDC %ILI data from January 2004 to December 2006 and were used to estimate flu activity from January 2007 through May 2010. The second set was identified using CDC %ILI data from January 2004 to May 2010 (including the H1N1 season) to estimate real-time flu activity from May 2010 through August 2015. In these two sets, 52% of search terms overlapped.

b. Google flu-related search query volume data

Search volume data of all identified Google flu-related search terms were obtained from Google Correlate (www.google.com/trends/correlate) and Google Trends (www.google.com/trends). The identified flu-related search terms for a given estimation time period were fixed, but the search volume of each term may vary from week to week. The search volume of each term was standardized to a standard normal distribution (i.e., mean of 0 and standard deviation of 1) using the search volume of all queries submitted for that week. Google Correlate data were available on a weekly basis across time, but data were only available from January 2004 to March 2015 from Google Correlate. Search volume data for the time period after March 2015 were obtained from Google Trends which publishes relative weekly search volume of search terms specified by the user. These search frequencies were divided by the total number of online searches done on Google over the same time period. This number was normalized to an integer values from 0 to 100, where 100 corresponds to the maximum weekly search since January 2004. To make search volume data obtained from Google Correlate compatible with those obtained from Google Trends, the Google Correlate data were linearly transformed to the same scale of 0 to 100.

c. CDC percentage of ILI (%ILI) data

The weighted version of the historical CDC percentage (%) of ILI data (available at www.cdc.gov/grasp/fluview/fluportaldashboard. html) were used in the ARGO model. The CDC % ILI data were collected through the Outpatient Influenza-like Illness Surveillance Network (ILINet).[5] The CDC ILI Surveillance consists of a network of health care providers who record the weekly total number of patients seen and the number of those patients with ILI and report these information to the CDC on a weekly basis.[5] An ILI is defined as a body temperature of 100° Fahrenheit or greater and a cough or a sore throat in the absence of a known cause other than influenza.[5]

3.2.1.2. ARGO formulation

The ARGO model with the following form was used to retrospectively estimate real-time %ILI from 2007 week 21 to 2015 week 20,

$$y_T = \mu_y + \sum_{j=1}^{N} \alpha_j Z_{T-j} + \sum_{i=1}^{K} \beta_i X_{i,T} + \epsilon_T, \quad \epsilon_T \sim \mathcal{N}(0, \sigma^2),$$

where $\boldsymbol{\alpha}$ = (α_1,…,α_N) are the autoregressive coefficients defining the importance of lagged values; $\boldsymbol{\beta}$=(β_1,...,β_K) are the coefficients indicating the strength for the i^{th} search term; μ_y is the intercept; and ϵ_T is the normally distributed error term which represents the randomness in the time series. Because we do not have the actual %ILI at time *T*, we assumed that the %ILI of current time has a regressive relationship with the previous CDC %ILI data. The components used to estimate this were *N* logit-transformed historical CDC %ILI observations $Z_{T-1}, Z_{T-2}, \ldots, Z_{T-N}$ at time *T*, and $X_{i,T}$, the log-transformed relative search volume of the i^{th} identified Google flu-related search term at time *T*. As most flu seasons tend to have an annual trend, we used *N* = 52 (weeks) to capture the within-year seasonality in the %ILI visits. At any given time *T*, ARGO only uses CDC data available up to time *T* - *j* to estimate the flu activity at time *T*, whereas the CDC will have the same information at time *T* + *k* (i.e., *k* weeks later than ARGO) where *j*, *k* are some positive integers. Without loss of generality (WLOG), we assumed that *k = 1* for the application of the ARGO model, i.e., a one-week delay in availability of CDC %ILI data. This assumes the best-case scenario; in reality, the lag could be as long as 4 weeks.

For any given time *t* = *T*, CDC ILI data and Google search volume data from the previous two years (104 weeks, i.e., *t* = *T* – 1, *T* – 2,…,*T* – 104) were used as the training data to estimate the parameters $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mu}_y$ of the ARGO model. Because this window moves forward with time, all parameters of the ARGO model were dynamically updated every week. The parameters of the ARGO model were estimated using a least absolute shrinkage and selection operator (LASSO) method to minimize the sum of squared residuals plus the sum of absolute values of the coefficients,

$$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mu}_y = arg\ min \sum_{t=T-1}^{T-104} \left(y_t - \mu_y - \sum_{j}^{N} \alpha_j Z_{t-j} - \sum_{i}^{K} \beta_i X_{i,t} \right)^2 + \lambda_\alpha \sum_{j=1}^{N} |\, \alpha_j| \\ + \lambda_\beta \sum_{i=1}^{K} |\, \beta_i|$$

where λ_α and λ_β are regularization hyperparameters. Theoretically, some kind of cross-validation method should be used to estimate all 2 hyperparameters in the ARGO model. However, because the ARGO model only uses a 2-year rolling window as the training data, any cross-validation result would be highly variable. Therefore, we set $\lambda_\alpha = \lambda_\beta$ to obtain more stable estimates. The final estimated ARGO model has the form,

$$\hat{y}_T = \hat{\mu}_y + \sum_{j=1}^{N} \hat{\alpha}_j Z_{T-j} + \sum_{i=1}^{K} \hat{\beta}_j X_{i,T}.$$

3.2.2. Detection informative change points that may signal the imminent onset of flu season by applying the Bayesian online change point detection (BOCPD) method to ARGO estimated %ILI data.

The second step of our strategy of flu surveillance was to apply a statistical change point detection method to the ARGO estimated %ILI at each time *t* to identify an informative change point that signal the imminent onset of flu season prior to the actual flu season beginning.

3.2.2.1. Rationale

Change point detection is a method applied to time series data to determine a point in time when the distribution of the series changes. Each flu season, flu activity starts to elevate gradually leading to the onset of flu season. CDC identifies the onset of flu season as the first week when the percentage of patients seeking medical attention with ILI symptoms is at or above a predetermined epidemic baseline for two consecutive weeks[4]. Figure 1 shows the CDC %ILI data at the national level from 2012 to 2016. The red lines represent the epidemic baselines for each season, which is the mean percentage of patient visits of ILI for the previous three seasons. Periods above the red lines are considered to be flu seasons.

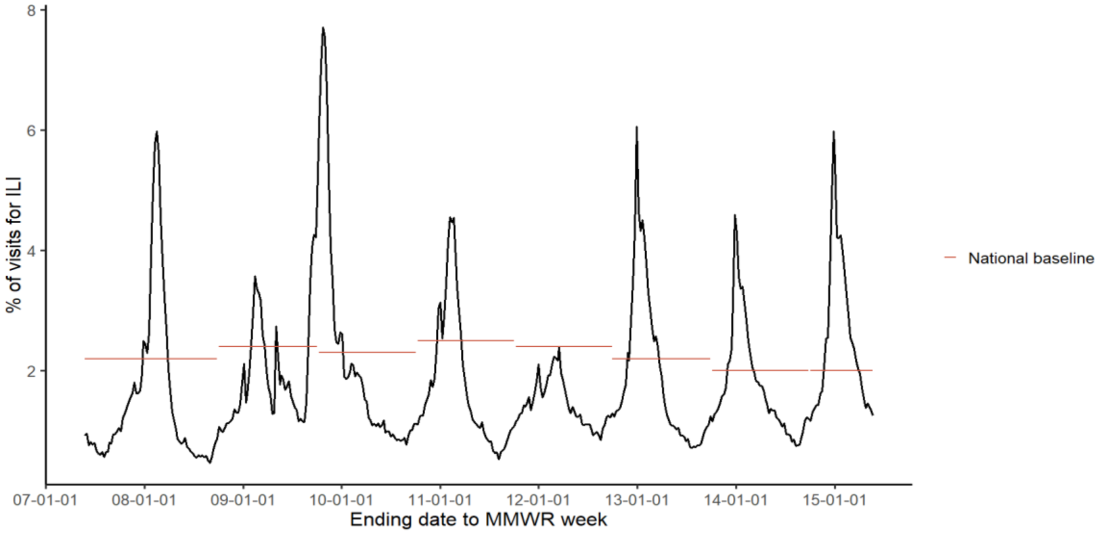

Figure 1. CDC-reported percentage of visits with ILI symptoms, 2007 – 2015 seasons.

We hypothesized that the CDC identified flu season onset (solid orange line in Figure 2) is not actually a change point. Rather, the actual change point occurs prior when there is an initial uptick in %ILI data (dotted green line in Figure 2). Thus, the identified change point would provide early detection for imminent onset of flu season.

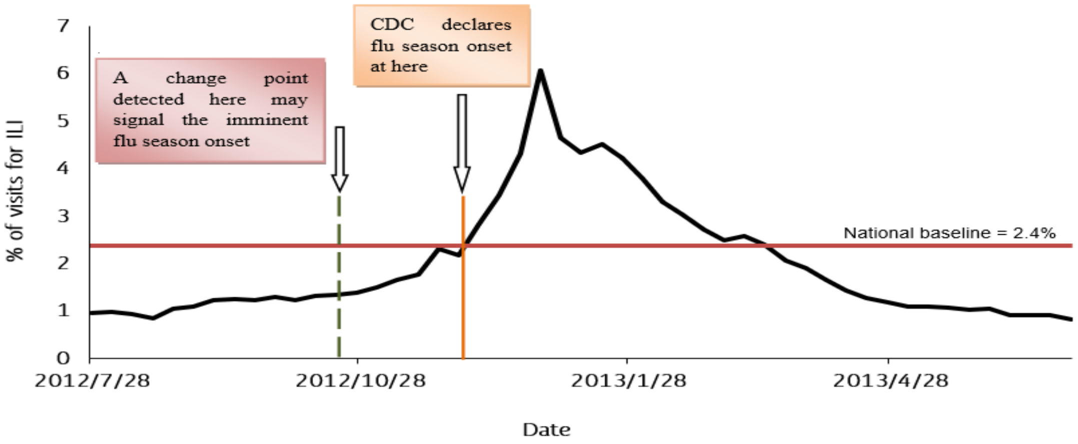

Figure 2. Rational for using change point detection for flu season onset.

3.2.2.2. Bayesian online change point detection

We applied the Bayesian online change point detection (BOCPD) method to the ARGO estimated weekly %ILI data to detect change points that may signal the onset of flu season. BOCPD is an online change point detection method that can detect change points concurrently as data become available. The BOCPD algorithm uses posterior probabilities to determine change points[53]. It applies a Bayesian framework at each time point to update the posterior probabilities given newly observed data. The BOCPD algorithm calculates the posterior probabilities for the *run length (r)* at each time point. A run length is defined as the number of observations since the most recent change point. Because a change point can be detected at any time, run length is a random variable. At time *t*, posterior probabilities, i.e., $P(r_t|\boldsymbol{y}_{1:t})$, are calculated for all possible run lengths, r_t = 0, 1, …, t – 1, where $\boldsymbol{y}_{1:t}$ is vector of observed values from the beginning to time *t*. WLOG, we assumed that a change point occurred at *t* = 1. Therefore, for example, r_t = *t* - 1 indicated that no change point had occurred during the last *t* observations, while r_t = 0 indicates that a change point just occurred at time *t*. At each *t*, the goal is to recursively estimate the posterior distribution of all possible run lengths given all the observed data up to *t*, so that change points can be determined from these distributions. The posterior distribution is calculated as follows.

Let $y_t \in \mathbb{R}$ be an observation from a time series data at each time point *t*. Define a partition of the sequence $y_1, y_2, \ldots, y_t$ as any subset of observations between two consecutive change points. Observations in each partition are assumed to be independent and identically distributed (*i.i.d.*) from a distribution $P\left(y_t|\boldsymbol{\theta}_t^{(r)}\right)$ and independent from observations in the other partitions. The notation $\boldsymbol{\theta}_t^{(r)}$ represents the parameters associated with a subset of observations given run length r_t in each time *t*.

The posterior distribution of the run length is found by normalizing the joint probability $P(r_t, \boldsymbol{y}_{1:t})$,

$$P(r_t|\boldsymbol{y}_{1:t}) = \frac{P(r_t,\, \boldsymbol{y}_{1:t})}{P(\boldsymbol{y}_{1:t})}.$$

The joint probability $P(r_t, \boldsymbol{y}_{1:t})$ is updated at every *t* by using a recursive message passing scheme,

$$\begin{aligned} P(r_t, \boldsymbol{y}_{1:t}) &= \sum_{r_{t-1}=1}^{t-1} P(r_t, r_{t-1}, \boldsymbol{y}_{1:t}) \\ &= \sum_{r_{t-1}=1}^{t-1} P(r_t, \boldsymbol{y}_t | r_{t-1}, \boldsymbol{y}_{1:t-1})\, P(r_{t-1}, \boldsymbol{y}_{1:t-1}) \\ &= \sum_{r_{t-1}=1}^{t-1} P(y_t | r_t, r_{t-1}, \boldsymbol{y}_{1:t-1}) P(r_t | r_{t-1}, \boldsymbol{y}_{1:t-1}) P(r_{t-1}, \boldsymbol{y}_{1:t-1}) \\ &= \sum_{r_{t-1}=0}^{t-1} \underbrace{P(r_t | r_{t-1})}_{\text{hazard}}\, \underbrace{P\left(y_t | r_{t-1}, \boldsymbol{y}_{t-1}^{(r_{t-1})}\right)}_{\substack{\text{predictive} \\ \text{probability}}} \underbrace{P(r_{t-1}, \boldsymbol{y}_{1:t-1})}_{\substack{\text{run length probability} \\ \text{from previous time step}}} \end{aligned}$$

Figure 3 illustrates the recursive updates of the posterior probability of all possible run length at a given time point *t* in BOCPD algorithm. In this figure, the path of calculation under the given hazard, where, in red, all possible routes to get a run length of 1 at time 4 (i.e., r_4 = 1) are shown. The BOCPD algorithm incorporates all these possible paths, estimating the probability of reaching that point under every path.

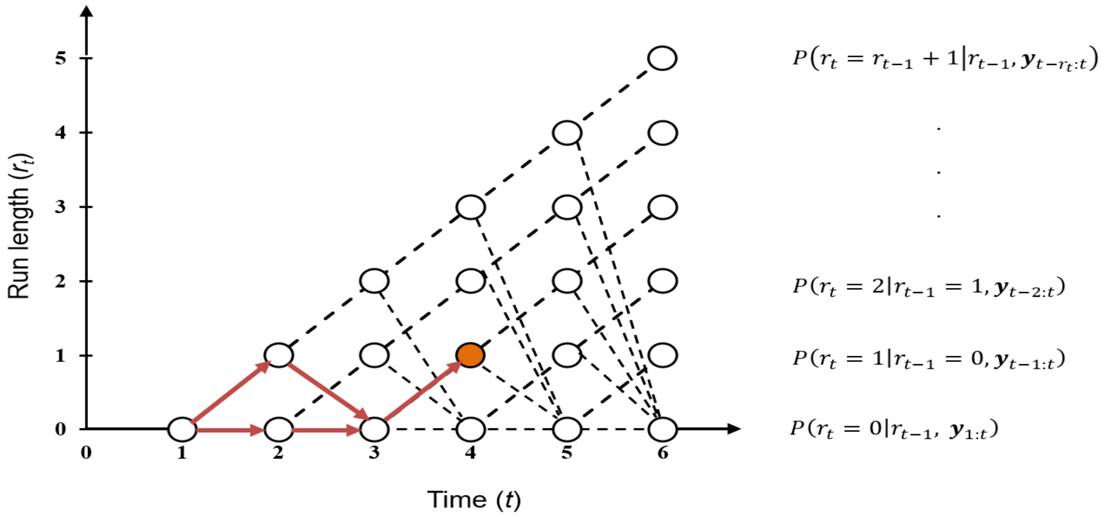
$P(r_t = r_{t-1} + 1 | r_{t-1}, \boldsymbol{y}_{t-r_t:t})$
$P(r_t = 2 | r_{t-1} = 1, \boldsymbol{y}_{t-2:t})$
$P(r_t = 1 | r_{t-1} = 0, \boldsymbol{y}_{t-1:t})$
$P(r_t = 0 | r_{t-1}, \boldsymbol{y}_{1:t})$
Run length (r_t)
5
4
3
2
1
0
0
1
2
3
4
5
6
Time (t)

Figure 3. The run length illustrated for *t* = 1,…,6. The orange shaded node gives *r₄* = 1, where the red path signifies the forward paths possible to reach that run length under the model assumptions.

WLOG, we assumed that a change point just occurred before the first datum, i.e., at *t* =1, therefore the probability for the initial run length is $P\ (r_1 = 0) = 1$. As shown above, the posterior probability is calculated through the following components: (1) the *hazard function H(y)* which represents the conditional prior probability of a change point occurring given the run length r_t at time *t*, (2) the predictive probability of a newly observed datum belonging to each run length, and (3) the run length probability from the previous time step *t* - 1. At time *t*, the current run length r_t can take one of two values: 0 if a change point occurs at this time or $r_{t-1} + 1$ if no change point occurs. The probability of these two events must be specified a priori as the hazard function *H*,

$$P\left(r_t \middle| r_{t-1}\right) = \begin{cases} H(r_{t-1}+1) & if\ r_t = 0 \\ 1 - H(r_{t-1}+1) & if\ r_t = r_{t-1}+1 \\ 0 & otherwise \end{cases} .$$

Here,

$$H(y) = \frac{P_{gap}(g = y)}{\sum_{t=y}^{\infty} P_{gap}(g = y)},$$

and $P_{gap}(g)$ is a prior distribution for the interval between change points. As with Adams and MacKay,[53] we assumed that the length of the interval between change points followed a geometric distribution with fixed time scale parameter λ_{gap}. Thus, the hazard function was constant at the prior hazard rate 1/ λ_{gap}. Although by convention, regular flu season occurs during the fall and winter months every year, we used a very conservative window of time for flu season to occur (i.e., from week 40 to week 20 of the following year). Therefore, we set λ_{gap} to be 20, representing our prior belief of the time interval between change points, and hazard rate *H* = 1/20. Algorithm 1 shows the mathematical steps of the BOCPD algorithm.

Algorithm 1 BOCPD algorithm

```
1: Initialize:
        $P(r_1 = 0) = 1; P(y_1, r_1) = P\left(y_1 \middle| \boldsymbol{\theta}_1^{(0)}\right)$
        $\boldsymbol{\theta}_1^{(0)} = \boldsymbol{\theta}_{prior}$
2: Update sufficient statistics:
        $\boldsymbol{\theta}_2^{(0)} = \boldsymbol{\theta}_{prior};$
        $\boldsymbol{\theta}_2^{(1)} = \boldsymbol{\theta}_1^{(0)} + U(y_1)$
3: for $t = 2, 3, \ldots$ do
```

```
4: Calculate predictive probabilities:
    $\pi_t^{(r)} = P(y_t|r_t = r, \boldsymbol{y}_{t-1}^{r_{t-1}}) = P\left(y_t|\boldsymbol{\theta}_t^{(r)}\right)$
5: Calculate growth probabilities:
    $P(r_t = r_{t-1} + 1, \boldsymbol{y}_{1:t}) = P(r_{t-1}, \boldsymbol{y}_{1:t-1})\, \pi_t^{(r_t)}(1 - H)$
6: Calculate change point probabilities:
    $P(r_t = 0, \boldsymbol{y}_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, \boldsymbol{y}_{1:t-1})\, \pi_t^{(0)} H$
7: Calculate marginal probabilities:
    $P(\boldsymbol{y}_{1:t}) = \sum_{r=0}^{t-1} P(r_t = r, \boldsymbol{y}_{1:t})$
8: Normalize the run length probabilities:
    $P(r_t = r|\boldsymbol{y}_{1:t}) = \frac{P(r_t = r, \boldsymbol{y}_{1:t})}{P(\boldsymbol{y}_{1:t})}, r = 0, \ldots, t - 1$
9: Update sufficient statistics:
    $\theta_{t+1}^{(0)} = \theta_{prior};$
    $\theta_{t+1}^{(r_t+1)} = \theta_t^{(0)} + U(y_t)$
end for
```

The predictive probability $P(y_t|r_t = r, \boldsymbol{y}_{t-1}^{r_{t-1}}) = P\left(y_t|\boldsymbol{\theta}_t^{(r)}\right)$ of a new datum y_t given the run length r_t at time *t* was calculated based on the data distribution and its parameters. We assumed that the ARGO estimated %ILI data followed a normal distribution with unknown mean and unknown precision (inverse of variance). Because the normal distribution is an exponential family distribution with conjugate priors, the predictive probability associated with a particular current run length can be characterized by a finite number of sufficient statistics, which in turn can be calculated incrementally as data arrives. If the prior density is specified as

$$\mathrm{p}(\theta) \propto g(\theta)^{\eta} e^{\phi(\theta)^T \nu},$$

then the posterior density is

$$p(\theta) \propto g(\theta)^{\eta+n} e^{\phi(\theta)^T(\nu+t(y))}.$$

This is of the same distributional form as the prior and shows that the prior density is a conjugate[54]. For data that follow a normal distribution with unknown mean μ and unknow precision τ, the conjugate prior is the normal-gamma distribution[55], a combination of a normal prior on μ and a gamma prior on τ,

$$NG\ (\mu, \tau|\mu_0, k_0, \alpha_0, \beta_0) = \mathcal{N}(\mu, \tau|\mu_0, (k_0\tau)^{-1})Gamma(\tau|\alpha_0, \beta_0),$$

where $\theta = \{\mu_0, k_0, \alpha_0, \beta_0\}$ are the model hyperparameters. Using a conjugate prior, the posterior predictive distribution at each r_t will always have a known, closed parametric form of the same family as the prior distribution, giving the simplified computations of all desired predictive probabilities $P\left(y_t|\boldsymbol{\theta}_t^{(r)}\right)$ without involving inefficient integrations.[54] In particular, the parameters of the posterior predictive distribution $NG\left(\mu, \tau\middle|\mu_{r_t}, \kappa_{r_t}, \alpha_{r_t}, \beta_{r_t}\right)$ given r_t have the form,

$$\mu_{r_t} = \frac{\kappa_0\mu_0 + \sum_{i=t-r_t}^{t} y_i}{\kappa_0 + r_t}$$

$$\kappa_{r_t} = \kappa_0 + r_t$$

$$\alpha_{r_t} = \alpha_0 + r_t/2$$

$$\beta_{r_t} = \beta_0 + \frac{\kappa_0 r_t\left(\left(\frac{\sum_{i=t-r_t}^{t} y_i}{r_t}\right) - \mu_0\right)^2}{2(\kappa_0 + r_t)}.$$

Note that the application can be extended to non-exponential families, provided that the posterior distributions can be computed numerically. In our application, y_t will be the ARGO estimated %ILI data.

Specifying the prior values of the hyperparameters of the data distribution will influence the results in the process. The BOCPD algorithm requires that the initial values of hyperparameters of the data distributions are specified a priori. These initial values are used to update the parameters in the recursive scheme shown above (Algorithm 1). However, unlike many numerical algorithms that replace the initial values with newly obtained values each run, the BOCPD algorithm uses the same initial values for the entirety of the process. Thus, correctly specifying these initial values and/or modifying the algorithm to allow for the reconciliation of these values is crucial for a pragmatic application of the algorithm. As flu activity is seasonal, and the historical flu activity data collected by the CDC are always publicly available, we incorporated these data to determine the initial values systematically.

Direct application of BOCPD to the ARGO data will result in a never-ending process where more and more change points are detected as time progresses. As such, the number of possible run lengths (i.e., the support of the run length) keeps growing although many run length values will have a very small posterior probability in order to account for the possibility that no change point has occurred since the beginning, resulting in a computationally inefficient algorithm. However, seasonal flu is a recurrent infectious disease, and hence, it is improbable that no change point occurs within a given season year. Thus, by allowing the BOCPD algorithm to restart each year and re-estimate the initial values of the hyperparameters based on data from previous flu seasons, it may result in a more efficient and accurate detection process. Therefore, we assessed the performance of the following four strategies: (1) without restarting the BOCPD algorithm and using an uninformative prior for the data distribution; (2) restarting the BOCPD algorithm on week 21 (around May 20th) every year with an uninformative prior for the data distribution; (3) without restarting the BOCPD algorithm and using historical CDC %ILI data to estimate prior values; (4) restarting the BOCPD algorithm and re-estimating prior values using historical CDC %ILI data on week 21 every year. For strategies (1) and (2), we set the prior values of hyperparameters of prior predictive distribution as $\mu_0 = 0$, $k_0 = 0.001$, $\alpha_0 = 1$, and $\beta_0 = 0.00001$. The choice of these values was arbitrary, representing the situation where the prior of hyperparameters were set up without a systematic principle. For strategy (3), the hyperparameters of the data distribution were estimated only once at the beginning of the study period (i.e., 2007 week 21), and the same initial values were used until the end of the study period (i.e., 2015 week 20). For strategy (4), we restarted the BOCPD algorithm in week 21 and re-estimated the initial values of hyperparameters of the data distribution each time using historical CDC %ILI data from all past regular flu seasons available to up each year. For this strategy, different initial values of hyperparameters were used to detect change points each year. For example, all historical CDC %ILI data up to week 20 of 2007 were used to estimate the hyperparameters for the detection process during the 2007 – 2008 season. However, because H1N1 outbreak that occurred in 2009 was not like other flu seasons, the trends of flu activity for this season was very different. Therefore, the CDC %ILI data from 2009 week 21 to 2010 week 20 were not used to estimate hyperparameters for subsequent seasons. For example, during the 2010 – 2011 season, only historical CDC %ILI data up to week 20 of 2009 were used to estimate the hyperparameters. We used the properties of the BOCPD algorithm to estimate the initial values of hyperparameters with the historical CDC %ILI data by computing the derivatives of the log marginal likelihood of the predictive distribution given r_t with respect to the hyperparameters $\boldsymbol{\theta}_t$={μ, k, α, β} of the posterior predictive distribution,[56]

$$log\ P\ (\boldsymbol{Z}_{1:T}|\boldsymbol{\theta}_t) = log \sum_{t=1}^{T} P(Z_t|\boldsymbol{Z}_{1:t-1}, \boldsymbol{\theta}_t),$$

where $\boldsymbol{Z}_{1:T}$ represent all historical CDC ILI data available up to each year (Algorithm 2). Then the partial derivatives of hyperparameters $\frac{\partial P(\boldsymbol{Z}_{1:t})}{\partial \theta}$ were plugged into a conjugate gradient optimizer to find the optimal values of hyperparameters. These values were used as the prior in the BOCPD algorithm.

Algorithm 2 Estimation of prior values of hyperparameter using property of BOCPD algorithm

```
1: Initialize:
        $P(r_1 = 0) = 1; P(Z_1, r_1) = P\left(Z_1 \middle| \boldsymbol{\theta}_1^{(0)}\right)$
        $\boldsymbol{\theta}_1^{(0)} = \boldsymbol{\theta}_{prior}$= {$\mu_0 = 0, k_0 = 0.1, \alpha_0 = 0.1, \beta_0 = 0.1$}
        $\frac{\partial P(Z_1, r_1)}{\partial \boldsymbol{\theta}} = 0$
2: Update sufficient statistics:
        $\boldsymbol{\theta}_2^{(0)} = \boldsymbol{\theta}_{prior};$
        $\boldsymbol{\theta}_2^{(1)} = \boldsymbol{\theta}_1^{(0)} + U(Z_1)$
3: for $t$ = 2, 3, … do
4: Calculate predictive probabilities:
    $\pi_t^{(r)} = P\left(x_t|r_t = r, \boldsymbol{Z}_{t-1}^{r_{t-1}}\right) = P\left(\boldsymbol{Z}_t|\boldsymbol{\theta}_t^{(r)}\right)$
5: Calculate growth probabilities:
    $P(r_t = r_{t-1} + 1, \boldsymbol{Z}_{1:t}) = P(r_{t-1}, \boldsymbol{Z}_{1:t-1})\, \pi_t^{(r_t)}(1 - H)$
6: Calculate partial derivatives of growth probabilities w.r.t $\boldsymbol{\theta}$:
    $\frac{\partial P(r_t = r_{t-1} + 1, \boldsymbol{Z}_{1:t})}{\partial \boldsymbol{\theta}} = (1 - H)\left(\frac{\partial P(r_{t-1}, \boldsymbol{Z}_{1:t-1})}{\partial \boldsymbol{\theta}} \pi_t^{(r_t)} + P(r_{t-1}, \boldsymbol{Z}_{1:t-1}) \frac{\partial \pi_t^{(r_t)}}{\partial \boldsymbol{\theta}}\right)$
7: Calculate change point probabilities:
    $P(r_t = 0, \boldsymbol{Z}_{1:t}) = \sum_{r_{t-1}} P(r_{t-1}, \boldsymbol{Z}_{1:t-1})\, \pi_t^{(0)} H$
8: Calculate partial derivatives of change point probabilities w.r.t $\boldsymbol{\theta}$:
    $\frac{\partial P(r_t = 0, \boldsymbol{Z}_{1:t})}{\partial \boldsymbol{\theta}} = \sum_{r_{t-1}} H\left(\frac{\partial P(r_{t-1}, \boldsymbol{Z}_{1:t-1})}{\partial \boldsymbol{\theta}}\ \pi_t^{(0)} + P(r_{t-1}, \boldsymbol{Z}_{1:t-1}) \frac{\partial \pi_t^{(0)}}{\partial \boldsymbol{\theta}}\right)$
9: Calculate marginal probabilities:
    $P(\boldsymbol{Z}_{1:t}) = \sum_{r=0}^{t-1} P(r_t = r, \boldsymbol{Z}_{1:t})$
10: Normalize the run length probabilities:
    $P(r_t = r|\boldsymbol{Z}_{1:t}) = \frac{P(r_t = r, \boldsymbol{Z}_{1:t})}{P(\boldsymbol{Z}_{1:t})}, r = 0, \ldots, t - 1$
11: Calculate partial derivatives of marginal probabilities w.r.t $\boldsymbol{\theta}$:
    $\frac{\partial P(\boldsymbol{Z}_{1:t})}{\partial \boldsymbol{\theta}} = \sum_{r=0}^{t-1} \frac{\partial P(r_t = r, \boldsymbol{Z}_{1:t})}{\partial \boldsymbol{\theta}}$
12: Update sufficient statistics:
    $\boldsymbol{\theta}_{t+1}^{(0)} = \boldsymbol{\theta}_{prior}$;
    $\boldsymbol{\theta}_{t+1}^{(r_t+1)} = \boldsymbol{\theta}_t^{(0)} + U(Z_t)$
end for
```

3.2.3. Rule of detecting change points

The BOCPD algorithm calculates the exact posterior distribution for run length at each time point, and change points are determined from these distributions. However, there is no established convention for identifying a change point when using BOCPD. Because the BOCPD algorithm applies a Bayesian framework at each time point to obtain the posterior distribution for run length, we used the maximum a posteriori (MAP) of the current run length distribution as a basis to formulate our rule of detecting change points as shown in Byrd et al[57]. MAP is defined as the mode of the run length posterior distribution. If the MAP of the current run length at time *t* decreases from the MAP of the current run length at time *t* – 1 by a sufficient amount relative to the current run length at time *t* – 1, then *t* was marked as the point in time when a change point just occurred. That is, there is a change point if the MAP of the current run length at time *t* satisfies,

$$\frac{MAP(t-1) - MAP\ (t)}{MAP(t-1)} > \alpha,$$

for some predetermined α value. Theoretically, if there is no change point at *t*, the MAP should occur at the maximum possible run length, whereas if a change point occurs at *t* the MAP should be at the run length 0. Thus, a sufficiently large difference in MAP between two time points would be evidence for a change point. We evaluated the performance on correctly detecting by varying the α from 0.1 to 0.8 in increments of 0.1 to determine the optimal level of α that should be used in flu surveillance.

3.2.4. Rule of identifying informative change points that may signal the onset of an imminent flu season

As the BOCPD algorithm was not created specifically for flu surveillance, another barrier in applying the BOCPD algorithm is that there is no systematic way to determine which change point should be chosen as the one that may signal the onset of a flu season. That is, the BOCPD process may result in many change points, but there is not an established way to distinguish which one is more informative in signaling the onset of flu season. Thus, we used the following rule to identify this change point.

During week 21 to week 39 (i.e., during the spring and summer), any change points detected were considered to be uninformative, because based on data from previous years, they are almost surely too far from the actual starting date for any regular flu season. However, if the ARGO %ILI during this period was at or above a predetermined epidemic baseline reported by the CDC for two consecutive weeks, we concluded that flu season has already begun outside the conventional time period of flu season (e.g., H1N1 outbreak) and stopped identifying looking for any informative change points.

During week 40 to week 20 of the following year (i.e., where conventional flu season would most likely would occur), if the ARGO %ILI was below the epidemic baseline, and there was a change point detected, this change point was considered as the early signal of the onset of the imminent flu season if it satisfied the following,

$$\mathcal{D} = \frac{epidemic\ baseline - ARGO\ \%ILI}{epidemic\ baseline} \leq p.$$

Smaller $\mathcal{D}$ indicates that the distance between ARGO %ILI and the epidemic baseline are close. If a change point satisfies this criterion, it means that the flu activity has significantly increased toward the epidemic baseline, suggesting that the flu activity level will likely cross the epidemic baseline to signal the onset of flu season. Because the epidemic baselines are not consistent across years, we used the relative distance (rather than absolute difference) to account for the seasonal variation in epidemic baseline. Any change points that do not satisfy this criterion were deemed uninformative. The choice of *p* was varied from 0.5 to 0.1 in increments of 0.1 to determine the optimal level of *p* in identifying informative change points. Once an informative change point has been identified, any subsequent change points were considered uninformative for this season. Therefore, based on our detection rule, only one informative change point will be identified for each season. Lastly, if the ARGO %ILI was at or above the epidemic baseline for two consecutive weeks before any informative change points were identified, then it was concluded that the current time was the onset of flu season.

3.2.5. Evaluation of the performance of the informative change points in correctly predicting the onset of flu season

To evaluate the accuracy of the change points identified, we determined how well they predicted the CDC-reported date of flu season onset each year. For each season, let *t* be the time of the informative change point and t^* be the CDC-reported date of flu season onset. We defined the interval $C(t)$, where

$$C(t) = \{t\colon 0 < t^{*} - t \leq 8 \text{ weeks}\} \qquad t^{*} > t,$$

to retrospectively determine if an informative change point correctly predicted the CDC-reported date of flu season onset. An informative change point was deemed correct in predicting the onset of flu season reported by the CDC if t fell within the interval defined by $C(t)$. For a given season, if there was no official onset of flu season identified by the CDC, an informative change point detected for that season was deemed incorrect in predicting the onset of flu season.

Two benchmarks were used to evaluate the performance of our methods: (1) the proportion of change points that correctly predicted the onset of flu season among all regular seasons, and (2) the average number of weeks between the change point and the CDC-reported date of flu season onset among all correct change points.

3.3. Methods of Aim 2

Aim 2: To apply natural language processing (NLP) techniques to the chronological case notes of adult probationers in Tarrant County, Texas to discover hidden risk factors related to probationers absconding from supervision.

3.3.1. Study population and data source

Our study population included misdemeanors and felony offenders who were 18 years or older at the time of arrest and have been sentenced to community supervision (probation) by local courts as well as those that lived in Tarrant County that receive community supervision in another county/state from January 1st, 2007 through December 31st, 2017. Subjects were excluded from consideration if they had not completed their probation. The total probation population in Tarrant County from 2007 to 2017 was 71,045, consisting of 12% of misdemeanors and felony absconders. Given the heavy computational requirement of using the data from the whole probation population, we randomly selected 2,000 probationers with 500 for each group (i.e., misdemeanors and felony absconders and successfully completers, respectively) without replacement as our study sample. Data used were generated from chronological case notes and provided by the Tarrant County Community Supervision and Correction Department along with information on probation completion status (i.e., complete or abscond). This study has received IRB approval.

3.3.2. Preprocessing text data

Before analyzing the data, we converted all text to lowercase, dropped all website and addresses, punctuations, numbers, stop words (e.g., “a”, “an” those commonly used words, and “absconders”, “absconded”, “absconding” those are highly associated with positive labeling) and extra whitespace. Moreover, due to the severity of crimes the offenders committed, the number of chronological case notes a probationer would have was varied. We combined all notes belong to the same probationer into a single document to take this variation into account. In the final data, each line was corresponding to a single document for each probationer. Let y_i be the label of each document, i = 1, …, N, where N is the total number of probationers. We labeled each document as absconder-related (i.e., y_i = 1) or completer-related (i.e., y_i = -1).

3.3.3. Text classification and features selection using text regression methods

3.3.3.1. Method 1: Concise Comparative Summarization (CCS) method

Miratrix and Ackerman developed a new text regression method known as the concise comparative summarization (CCS) method that can be used to efficiently obtain a concise regression model which is more interpretable than regression models obtained from the traditional text regression method (e.g., least absolute shrinkage and selection operator (LASSO) method).[37] Essentially, the CCS is a machine learning extension to the general linear model,

$$g(\boldsymbol{Y}|\boldsymbol{X}) = f(\boldsymbol{X}),$$

where $\boldsymbol{Y}$ is a binary outcome indicating 1 for absconders and -1 for successful completers, $\boldsymbol{X}$ are regressors which are important words, phrases or themes from the case notes related to the outcome. The core idea of this method is to regress $\boldsymbol{Y}$ on the counts of all words and phrases in the notes.[37]

Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ to be the vector of coefficients for all words and phrases, the CCS method selects features which are important words and phrase related to absconded status by minimizing a regularized loss function (i.e., a sum of a loss function and a penalty function of a vector of parameters),[37]

$$\widehat{\boldsymbol{\beta}} = arg \underset{\beta=(\beta_1,\dots,\beta_p)}{min} \mathcal{L}(\beta), \tag{1}$$

where

$$\mathcal{L}(\boldsymbol{\beta}) = \underbrace{\sum_{i=1}^{N} \left[\left(1 - y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\, \beta_j \right) \right)}_{\text{squared hinge loss function}} \vee 0 \right]^2 + C \underbrace{\sum_{j=1}^{p} |\beta_j|}_{L^1 \text{ norm penalty}},$$

and $a \vee b$ denoting the maximum of *a* and *b*; $x_{ij} = \frac{c_{ij}}{\sqrt{\sum_{i=1}^{N} c_{ij}^2}}$ is used to rescale word *j* differently based on how many times it appears in each document; *N* is total number of documents; C is a regularization parameter and $C \in [0, \infty]$; *p* is total number of features.

The square hinge loss is a loss function very similar to an ordinary least squares (OLS)-type function. Following above notation of the regularized loss function, it can be written as

$$h(y, \hat{y}) = \sum_{i=0}^{N} [max(1 - y_i \hat{y}_i), 0]^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} x_{ij}\, \hat{\beta}_j$, is the predicted value. Figure 4 shows the squared hinge loss for y = 1 with different predicted value $\hat{y}$. When the true $y = \hat{y} = 1$, and when $\hat{y} > 1$ which is an indication that the label is sure that it’s a correct label, the square hinge loss would be 0. When $\hat{y} < 1$ which is an indication that the label is not sure that it’s the correct label, the square hinge loss would be quadratically increased. The square hinge loss has monotonic property which satisfies the assumption of optimization algorithm.

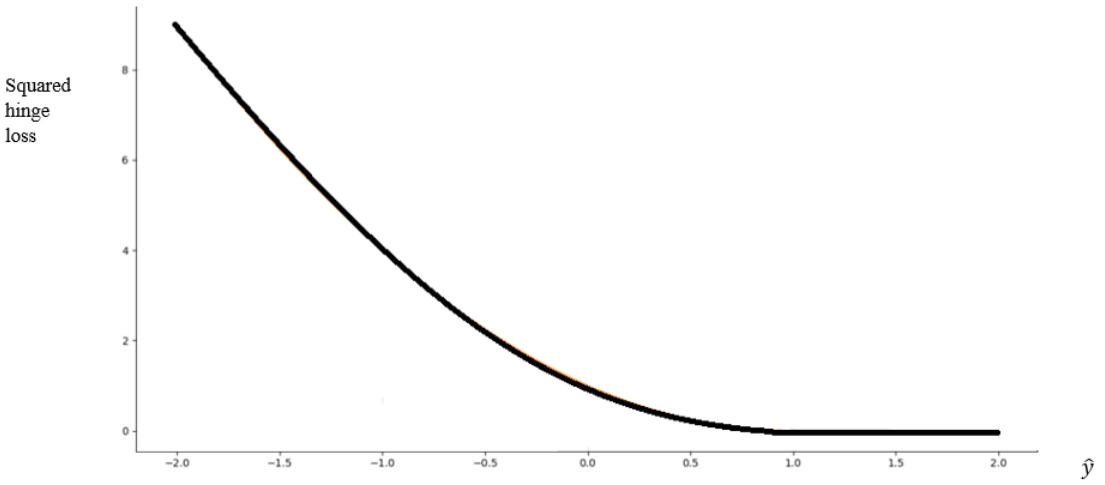

Figure 4. Relationship between squared hinge loss and the predicted value given target y = 1.

The CCS model applies a regularization technique to avoid the risk of over-fitting of the regression model by imposing the *L¹* norm penalty on the objective function (i.e., squared hinge loss) to regularize the vector of coefficients. Over-fitting means that the model includes too many parameters making it too complex to fit the data.[58] It happens when we had more features than the sample size. An overfitted model may fit the idiosyncrasies of a sample data from the population well, but it has poor ability to fit a new sample or the overall population.[58] The coefficients of such models would be misleading. The penalty term gives a larger loss for more complex models and a smaller loss for simpler models and hence helps to control over-fitting problem.[59] Therefore, the CCS method finds the minimum of a regularized loss function by finding the minimum of the sum of a loss function and the L^1 norm penalty function. The L^1 norm penalty is the sum of the absolute values of the vector of parameters. It has the property of shrinking some coefficients of parameters toward zero, and thus producing a sparse (i.e., concise) regression model with a relatively small number of important features (e.g., words and phrases) that with non-zero coefficients.[59,60] Figure 5 geometrically illustrates why L^1 norm penalty would give 0 coefficients which would result in a sparse model, assuming a model with two coefficients (β_1, β_2). As figure 2 shows, the blue elliptical plot represents the contours of the loss function (because we use a square loss function), and the red rotated square plot represents the contours of the L^1 norm penalty. The minimum of the regularized loss function is achieved when these two contour plots are tangent to each other. Because the L^1 norm has sharp corners (aligned with the coordinate axes), there is a large probability that these two contours to be tangent at these corners, corresponding to one or more coefficients set to 0. For example, figure 5 shows that at the corner (i.e., the blue point) where these two contours tangent, $\beta_1$is set to 0. Therefore, using the L^1 norm penalty, we had a sparse text regression model which would be relatively easy to interpret and generate useful knowledge.

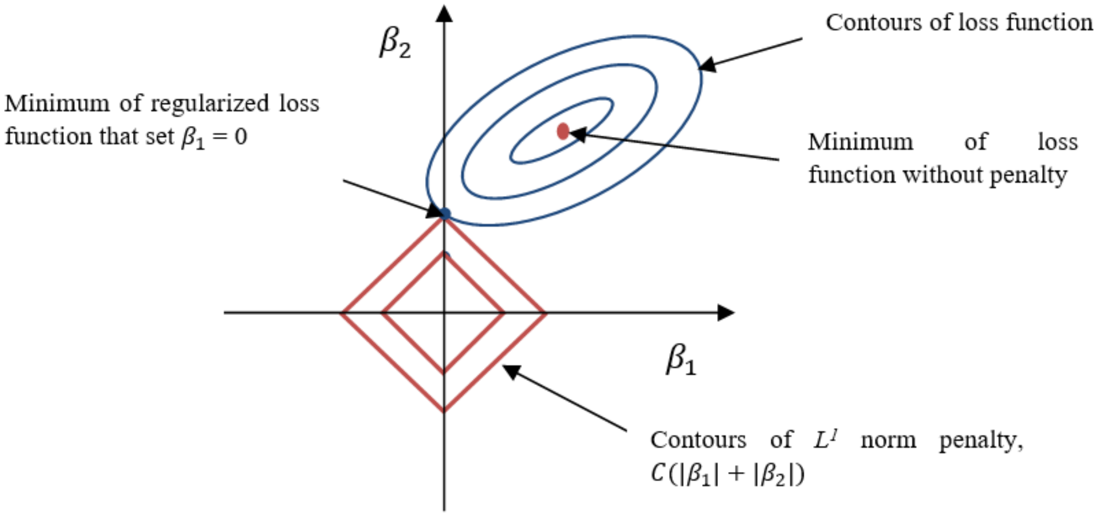
β_2
Contours of loss function
Minimum of regularized loss function that set $\beta_1 = 0$
Minimum of loss function without penalty
β_1
Contours of L^1 norm penalty, $C(|\beta_1| + |\beta_2|)$

Figure 5. Geometric interpretation of using L^1 norm penalty.

Furthermore, in the L^1 norm penalty, the magnitude of the regularization parameter C controls the trade-off between minimizing loss and regularization.[59,61] As the value of C increases, the effect of regularization will be strengthened. Theoretically, a higher value of C used in the CCS method will result in a more concise model with fewer features.[37] As C increases, the penalty function will increase. When we want to minimize the regularized loss function, the values of coefficients needed to be small. Thus, we will shrink more coefficients toward 0 with a larger C and the resulting text model will be more concise with fewer words selected. Conversely, a lower C value will produce a model with more features.[37] However, if the C is too small, for example, C = 0, the penalty term has no effect on controlling the complexity of the model and the estimates are obtained by minimizing the squared hinge loss function only, the resulting model will be at risk of over-fitting. In the CCS method, over-fitting means that the features obtained in the model are detected solely due to random chance in the appearance patterns of words not due to the relationship between the label and the features.[37] To prevent this problem, selecting the appropriate value of C is important.

We used the permutation approach suggested by Miratrix and Ackerman[37] to find an appropriate C that gives a statistically significant summary (i.e., regression model) indicating the presence of systematic differences in the text between the positively and negatively labeled document. Specifically, first, we regressed the data with the same regularized loss function shown in (1) and found the original C^{obs} that gave an empty model which had no selected words/phrases (all coefficients were zero), given the original labels. Then we repeatedly permuted the labeling 100 times across the documents. At each permutation, we regressed the data with the same regularized loss function shown in (1) and found the corresponding C^* that gave a model without no selected words/phrases with the permuted label. These C^*s gave the null distribution of C that indicates what C should be if there were no systematic differences in the text between the positive and negative labels. Finally, we compared the original C^{obs} which obtained with the original label to the distribution of C^*s to calculate a p-value,

$$p\text{-value} = Pr\{C^{obs} \geq C^*\}.$$

If C^{obs} is much larger than the permutation distribution, the corresponding p-value will be small which indicates that there is a real connection between the text and the labeling. Similarly, if we pick a $C = C^*$ that is at the 95th percentile of the permutation distribution, we are 95% confident that by using this value of C, the resulting text regression is due to the relationship of the outcome and the words/phrases, and not due to random chance. This approach provides a useful minimum value (e.g., C at the 95th percentile of the permutation distribution) for the final C. Any C smaller than this bound indicates that the resulting text model could be purely due to conincidence.[37] To find the C value which produced a more interpretable model, we varied the C from the required minimum value (i.e., C at the 95th percentile of the permutation distribution) to the C^{obs} which gave an empty model in increments of 25 percentile.

After specifying the value of C, we used this C in the L^1 norm penalty and minimized the equation (1) by using the greedy coordinate descent optimization algorithm to obtain the regression text model. As with Miratrix and Ackerman,[37] in the use of greedy coordinate descent, we repeatedly found the feature with the highest gradient and then optimized its corresponding β_j with a line search over the regularized loss function (Algorithm 3).

Algorithm 3 Greedy Coordinate Descent

```
$\boldsymbol{\beta}$ = ∅;
features = ∅;
while Not Converged do
$\beta$[intercept] = updateFeature(intercept)
$f$ = findHighestGradient
features. add($f$)
$\boldsymbol{\beta}$ [$f$] = updateFeature($f$)
end while
```

The inner algorithm in the greedy coordinate descent is finding the feature with the largest gradient. To do this, as with Miratrix and Ackerman,[37] we dynamically generated the features by exploiting the nested structure of any multiword phrase having a smaller phrase as a prefix (Algorithm 4). Specifically, we searched all single words and estimate the corresponding coefficients to choose the significant single words (i.e., those with non-zero coefficients) by minimizing the regularized loss function. Then we searched all two-word phrases that begin with the significant single words selected from the previous step and chose the significant two-word phrases using the same optimization algorithm as above. Then we repeated the step for three-word phrases and four-word phrases. Using this method, the final design matrix included all words and phrases with non-zero coefficients selected to be important. Words/phrases not selected through the searching process were not included in the design matrix. The matrix was created on the fly to increase computational efficiency.

Algorithm 4 findHighestGradient

```
$features$ = all non-zero features so far.
$best_f$ = arg $max_{f \in features} gradient(f)$
$u_1, \ldots, u_{p1}$ = all unigrams in dictionary
Q = queue( ), a queue of all features to check which unigram is important
for $u \in u_1, \ldots, u_{p1}$ do
    if gradient($u$) > gradient($best_f$) then
        $bestf = u$
    end if
    Q.add( $u$ )
end for
while not Q do
    $f$ = Q.next()
    if not canPrune($f$, $best_f$) then
        for $c \in$ children($f$) do      // e.g., two-word phrases that begin with the significant single words selected from the previous step
            if gradient($c$) > gradient($best_f$) then
            $best_f = c$
            end if
            Q.add($c$)
        end for
    end if
end while
```

The final design matrix consists of frequencies of all selected one-, two-, three- and four-word phrases which may be informative to generate knowledge about factors associated with probation absconders.

3.3.3.2. Method 2: LASSO Regression

As there are other text regression methods, we applied the classic linear LASSO text regression to the same case note data used in the CCS method to compare the results between two methods. In the LASSO method, we minimized the sum of square errors with a bound on the sum of absolute values of the coefficients (i.e., the L^1 norm penalty),

$$\hat{\beta}_{Lasso} = arg \min_{\beta=(\beta_1,\dots,\beta_p)} \sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|, \tag{2}$$

where $x_{ij} = \frac{c_{ij}}{\sqrt{\sum_{i=1}^{N} c_{ij}^2}}$ is used to rescale word j differently based on how many times it appears in each document; N is total number of documents; λ is a regularization parameter and $\lambda \in [0, \infty]$; p is total number of features. Coordinate descent optimization method was applied to minimize equation (2) to obtain regression coefficients. The hyperparameter λ was estimated using the 5-fold cross-validation method. The LASSO method also uses the L^1 norm penalty to control the complexity of the model. Due to the nature of the L^1 norm penalty explained above, this method shrinks the coefficients of uninfluential words and phrases to exactly zero, which produces a model that only includes the most important words and phrases for explaining the labels of the documents. However, the LASSO method cannot create the text matrix on the fly. We had to convert the raw text into an $m \times p$ document-term matrix with the frequency of p all possible words prior to applying the LASSO method, where m is the number of documents. Generating full document-term matrices with different lengths of phrases is computationally tedious. It is expensive in both time and memory and grows increasingly so with the number of possible phrases. Therefore, we only considered all possible unigrams (i.e., unique single words) when we applied the LASSO method.

In this study, we compared words and phrases which were associated with probation absconding (positive label of documents) and associated with completion (negative label of documents) between the CCS and the LASSO methods.

Chapter 4 Results

4.1. Results of Aim 1

This study retrospectively estimated %ILI and identified change points that correctly predicted the actual date of flu season onset reported by the CDC from 2007 week 21 (i.e., 5/26/2007) to 2015 week 20 (i.e., 5/23/2015). Although already shown to be accurate by others,[15] figure 6 shows the retrospective estimation of ARGO %ILI against the weighted CDC %ILI during the study period.

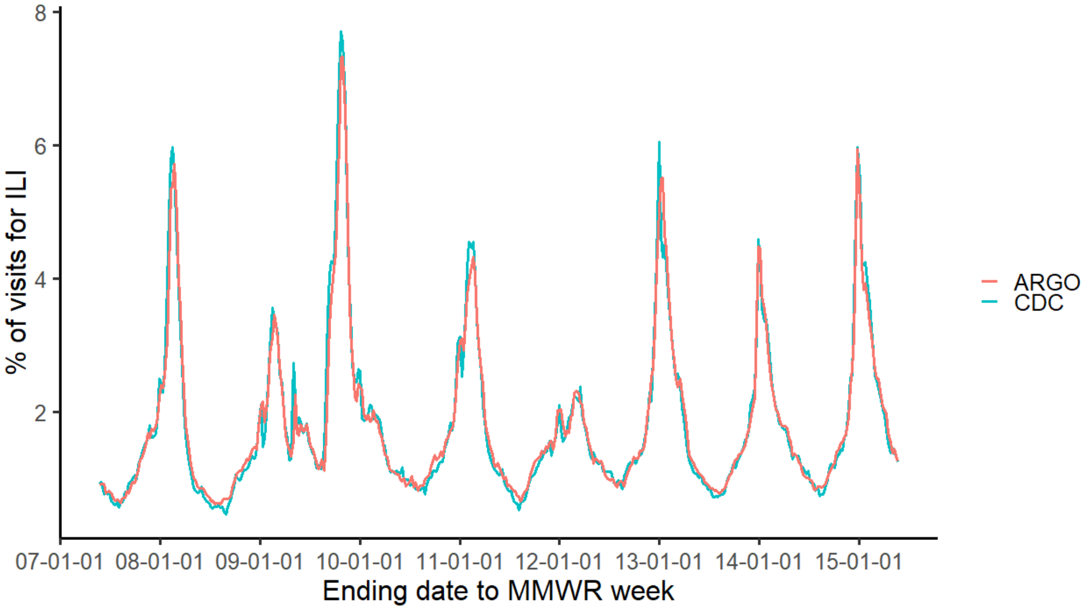

Figure 6. ARGO %ILI against the weighted CDC %ILI from 2007 – 2015 seasons.

In our rule of identifying change points that predicted the onset of the imminent flu season for each year, we compared the ARGO %ILI to the predetermined epidemic baseline reported by the CDC to conclude if the flu season has already begun. Table 1 shows the epidemic baselines reported by the CDC for each season.

Table 1. Predetermined epidemic baselines reported by the CDC, 2007 – 2015 seasons

Seasons	Epidemic baseline reported by the CDC
2007-2008	2.2%
2008-2009	2.4%
2009-2010	2.3%
2010-2011	2.5%

2011-2012	2.4%
2012-2013	2.2%
2013-2014	2.0%
2014-2015	2.0%

To evaluate the performance of a strategy on identifying change points that correctly predicted the actual onset of flu season, we used the actual date of flu season onset reported by the CDC and the corresponding measurement interval *C*(*t*) (Table 2).

Table 2. Interval *C*(*t*) used to identify change points that correctly predicted the onset of flu season, 2007 – 2015 seasons

Seasons	**Start of** ***C*****(*****t*****)**	**End of** ***C*****(*****t*****)**
		(CDC-reported date of onset)
2007-2008	2007 week 44	2007 week 52
2008-2009	2008 week 49	2009 week 4
2009-2010	H1N1 outbreak, flu season begun on 2009 week 34	
2010-2011	2010 week 43	2010 week 51
2011-2012	No official onset based on onset definition	
2012-2013	2012 week 40	2012 week 48
2013-2014	2013 week 40	2013 week 48
2014-2015	2014 week 39	2014 week 47

The following tables show the proportion of change points that correctly predicted the actual onset of flu season among previous 7 regular flu seasons (i.e., 2007 – 2008, 2008 – 2009, 2010 – 2011, 2011 – 2012, 2012 – 2013, 2013 – 2014, 2014 – 2015) included in this study and the average number of weeks between the correct change point and the actual onset reported by the CDC for four methods: (1) without restarting the BOCPD algorithm and using an uninformative prior for the data distribution; (2) restarting the BOCPD algorithm on week 21 (around May 20th) every year with an uninformative prior for the data distribution; (3) without restarting the BOCPD algorithm and using historical CDC %ILI data to estimate prior values; (4) restarting the BOCPD algorithm and re-estimating prior values using historical CDC %ILI data on week 21 every year. Date corresponding to each informative change point for these four strategies are displayed in Appendix A: supplemental results for Aim 1. In the 2009 – 2010 season, the ARGO %ILI started to exceed the predetermined epidemic baseline reported by the CDC on 2009 week 35 (i.e., 2009/9/5) for two consecutive weeks, we concluded that flu season had already begun outside the conventional time period of flu season (e.g., H1N1 outbreak) and stopped identifying any change point thereafter for this year.

For the strategy (1), i.e., without restarting the BOCPD algorithm and using uninformative prior values, the maximum proportion of correct prediction among all regular seasons was 0.43 at $p = 0.5$ over α levels except for the extreme level of α at 0.8 (Table 3). The corresponding average week between the correct change points and the CDC-reported date of flu season onset was 5.3 weeks (Table 4). The lowest proportion of correct prediction was 0.14 at $p = 0.2$ and $p = 0.1$ regardless of α levels.

Table 3. Proportion of correct prediction for strategy (1) without restarting the BOCPD algorithm and using uninformative prior values, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.29
***p* = 0.4**	0.43	0.43	0.43	0.43	0.43	0.43	0.29	0.14
***p* = 0.3**	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.14
***p* = 0.2**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
***p* = 0.1**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

Table 4. Average weeks of correct change points prior to the actual onset of flu season for strategy (1) without restarting the BOCPD algorithm and using uninformative prior, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	5.3	5.3	5.3	5.3	5.3	5.3	4.3	4.3
***p* = 0.4**	3.7	3.7	3.7	3.7	3.7	3.7	3	3
***p* = 0.3**	3	3	3	3	3	3	3	3
***p* = 0.2**	3	3	3	3	3	3	3	3
***p* = 0.1**	3	3	3	3	3	3	3	3

For the strategy (2), i.e., restarting the BOCPD algorithm every year and using uninformative prior values, the maximum proportion of correct prediction was higher than the strategy (1), where the maximum was 0.57 at both p = 0.5 and p = 0.4 across all α levels (Table 5). Meanwhile, the maximum average week between the correct change points and the CDC-reported date of flu season onset was also earlier than that of strategy (1), was 5.8 weeks at p = 0.5 over the α levels except for α at 0.7 and 0.8 (Table 6). Similar to strategy (1), the lowest proportion of correct was also 0.14 and observed at p = 0.2 and p = 0.1 across all α levels.

Table 5. Proportion of correct prediction for strategy (2) restarting the BOCPD algorithm every year with uninformative prior values, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
***p* = 0.4**	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
***p* = 0.3**	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
***p* = 0.2**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
***p* = 0.1**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

Table 6. Average weeks of correct change points prior to the actual onset of flu season for strategy (2) restarting the BOCPD algorithm every year with uninformative prior, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	5.8	5.8	5.8	5.8	5.8	5.8	5	5
***p* = 0.4**	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
***p* = 0.3**	3	3	3	3	3	3	3	3
***p* = 0.2**	3	3	3	3	3	3	3	3
***p* = 0.1**	3	3	3	3	3	3	3	3

For both strategies (3) and (4), prior values of hyperparameters were estimated using the historical flu activity data collected by the CDC. For strategy (3), i.e., without restarting the BOCPD algorithm and using prior values estimated from the historical CDC %ILI data, the hyperparameters of the data distribution were estimated only once at the beginning of the study period, and the same initial values were used throughout the entire study period. And those prior values were, μ_0 = 3.116287, k_0 = 0.095282, α_0 = 0.730475, β_0 = 0.009435. For strategy (4), i.e., restarting the BOCPD algorithm every year and using prior values of the hyperparameters estimated from the historical

CDC %ILI data up to each season. The prior values of hyperparameters were re-estimated using all historical CDC %ILI data up to each season (Table 7). Because the 2009 – 2010 season was not a conventional flu season, %ILI data of this season were excluded when we estimated the prior values used in the detection process for the subsequent seasons.

Table 7. Prior values of hyperparameters used in the strategy (4) for each season, 2007 – 2015 seasons

Seasons	$\boldsymbol{\mu_0}$	$\boldsymbol{k_0}$	$\boldsymbol{\alpha_0}$	$\boldsymbol{\beta_0}$
2007 – 2008	3.116287	0.095282	0.730475	0.009435
2008 – 2009	3.056904	0.093371	0.705951	0.009147
2009 – 2010 (H1N1)	2.937169	0.090925	0.695762	0.008629
2010 – 2011	2.937169	0.090925	0.695762	0.008629
2011 – 2012	3.000586	0.099443	0.677238	0.007925
2012 – 2013	3.028302	0.096231	0.741366	0.00841
2013 – 2014	3.075587	0.10095	0.708239	0.007428
2014 – 2015	3.119509	0.096047	0.720333	0.007635

Table 8 presents the proportion of correct prediction for the strategy (3), i.e., without restarting the BOCPD algorithm and using prior values estimated from the historical CDC %ILI data. For this strategy, because the BOCPD algorithm does not restart every year, the prior values of the hyperparameters were estimated only once using all historical CDC %ILI data up to the beginning date of the study and used for the entirety of the process. This strategy had consistently good predictions of the onset of flu season with the proportion of correct being 0.86 at p = 0.4 across all α levels except for α = 0.8. The corresponding average number of weeks between the correct change point and the actual date of flu season onset was 3.3 weeks (Table 9). The poorest prediction was observed at p = 0.1 regardless of α levels.

Table 8. Proportion of correct prediction for strategy (3) without restarting the BOCPD algorithm and using prior values estimated from the historical CDC ILI data, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.29

***p* = 0.4**	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.29
***p* = 0.3**	0.86	0.86	0.86	0.71	0.71	0.71	0.71	0.29
***p* = 0.2**	0.71	0.71	0.71	0.57	0.57	0.57	0.57	0.29
***p* = 0.1**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

Table 9. Average weeks of correct change points prior to the actual onset of flu season for strategy (3) without restarting the BOCPD algorithm and using prior values estimated from the historical CDC ILI data, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	3.2	3.2	3.2	3.2	3.2	3.2	3.2	2.5
***p* = 0.4**	3.3	3.3	3.3	3.3	3.3	3.3	**3.3**	2.5
***p* = 0.3**	2.7	2.7	2.7	3	3	3	3	2.5
***p* = 0.2**	1.8	1.8	1.8	2	2	2	2	2.5
***p* = 0.1**	1	1	1	1	1	1	1	1

For strategy (4), i.e., restarting the BOCPD algorithm and re-estimating prior values every year using historical CDC %ILI data up to each season, the optimal proportion of correct prediction was 0.86 and was observed at p = 0.4 and p = 0.3 over all α levels except for α = 0.7 and α = 0.8 (Table 10), while the corresponding average number of weeks of the correct change point prior to the actual flu season onset was 3.2 weeks (Table 11). The prediction was again worst at p = 0.1 regardless of α levels.

Table 10. Proportion of correct prediction for strategy (4) restarting the BOCPD algorithm every year with prior values estimated from the historical CDC ILI data, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
***p* = 0.5**	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.14
***p* = 0.4**	0.86	0.86	0.86	0.86	0.86	0.86	0.71	0.14
***p* = 0.3**	0.86	0.86	0.86	0.86	0.86	0.86	0.71	0.14
***p* = 0.2**	0.57	0.57	0.57	0.57	0.57	0.57	0.29	0.14
***p* = 0.1**	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0

Table 11. Average weeks of correct change points prior to the actual onset of flu season for strategy (4) restarting the BOCPD algorithm every year with prior values estimated from the historical CDC ILI data, 2007 – 2015 seasons

	α = 0.1	α = 0.2	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7	α = 0.8
***p* = 0.5**	4.2	4.2	4.2	4.2	4.2	4.2	3.4	4
***p* = 0.4**	3.2	3.2	3.2	3.2	3.2	**3.2**	3.4	4
***p* = 0.3**	3.2	3.2	3.2	3.2	3.2	**3.2**	3.4	4
***p* = 0.2**	2.3	2.3	2.3	2.3	2.3	2.3	2.3	4
***p* = 0.1**	1	1	1	1	1	1	1	None

Compared to strategies (1) and (2), the performance on identifying change points that correctly predicted the actual onset of flu season were much better using strategies (3) and (4). However, the average number of weeks between the correct change points and the CDC-reported date of flu season onset of strategies (3) and (4) were later than that of strategies (1) and (2).

4.2. Results of Aim 2

In our sample case notes data, there were 50,933 unique words and 1,391,750 bigrams. The number of trigrams would appear more times than bigrams. However, the size of the frequency matrix of bigrams was about 20GB which has already exceeded the limit of processing long vector of the statistical software R. Therefore, we limited our analysis using LASSO method with unigrams only.

Let C_1 to be the 95th percentile of 100 permutations of C. In the permutation test, we found C_1= 8.8. The C^{obs} that gave an empty model which had no selected phrases (all coefficients were zero), given the original labels was 23.9. The C^{obs} was much larger than C_1, the corresponding p-value from the permutation distribution of C was extremely small indicating that there was a statistically significant relationship between the text and the labeling and the resulting model (i.e. summary of case notes) was informative. To find the C value which produced a more interpretable model, we varied the C from C_1 to C^{obs} in increments of 25 percentile. Tables 12 and 13 display words and phrases associated with probation absconding and completion varied by different C values. We found that in maximum there were 10 and 14 words and phrases were associated with probation absconding and successful completion, respectively.

Table 12. Words and phrases associated with probation absconding varied by different C values

Words and phrases[*]	C_1**= 8.8**	C_2**= 12.5**	C_3**=16.3**	C_4**=20.1**
technical violations				
transfer intake				
technical			NA	NA
due status			NA	NA
fee amount		NA	NA	NA
cannabinoids		NA	NA	NA
failed pay		NA	NA	NA
technicals		NA	NA	NA
reported transfer		NA	NA	NA
attempts contact		NA	NA	NA

NA indicates this word was not selected.
[*] Words and phrases were listed from the highest coefficients to the lowest.

Table 13. Words and phrases associated with probation successful completion varied by different C values

Words and phrases[*]	C_1**= 8.8**	C_2**= 12.5**	C_3**=16.3**	C_4**=20.1**
paid full				
current				NA
current fees				NA
alcohol use			NA	NA
paid forwarded			NA	NA
fees paid full			NA	NA
completed fees		NA	NA	NA
compliance		NA	NA	NA
everything going well		NA	NA	NA
paid fees		NA	NA	NA
payment		NA	NA	NA
reported date		NA	NA	NA
satisfied fees		NA	NA	NA
travel		NA	NA	NA

NA indicates this word was not selected.
[*] Words and phrases were listed from the highest coefficients to the lowest.

Table 14 shows unigrams associated with probation absconding and successful completion obtained by using the LASSO method. In the LASSO method, more words were found to be associated with probation absconding and completion than using the CCS method. There were 120 and 9 words that were associated with probation absconding and successful completion, respectively.

Table 14. unigrams associated with probation absconding and successful completion using LASSO method

Probation absconding	Successful completion
acticitiy, anemic, anything, apmts, appeared, attepted, attn, befo, bogged, bonded, booked, boosted, budet, cannabinoids, cchr, ccl, certifica, certififcate, cetified, colostrophy, conference, confs, congratul, coop, cork, cossabone, cousin, cpts, criminal, ddc, deice, deliverable, develops, diminished, disconnected, distributing, doep, domain, dosent, drawer, dumps, dwli, emphasis, employmt, endless, engineers, failed, failure, field, flash, forwarding, ftc, fugitive, fumbled, funding, galbladder, gethis, gonzalezvazquez, grade, grandmother, hardcopy, harvesting, hhome, hilltop, hung, indegent, index, insructions, intake, intravenously, intrest, invpolvement, isp, journaling, knight, lateand, lethargy, lted, mailed, message, modifications, motives, nect, nop, noting, ntb, nurned, nwfsu, obser, ocked, origionated, painted, pasture, payslip, playi, pleaded, proceed, procees, provide, received, references, regulations, relocates, restitutions, rferrral, rhymes, rolled, rri, salace, sihned, soj, stampred, steep, stopper, surety, suspicion, unsuccessfully, vicious, voc, vop	acceptances, acknowledge, boating, character, clea, coord, enr, fixes, jalapeno

Our results showed that the text regression model obtained from the CCS method is generally more concise (i.e. manageable number of words and phrases were selected) compared to the LASSO method which selected more than 100 unigrams. In addition, the CCS method spent approximately 15 minutes to complete the searching of important phrases including unigrams, bigrams, trigrams, and so forth until there were no more eligible phrases. Meanwhile, the LASSO method spent approximately 2 minutes to complete the analysis. However, this efficiency only limits in searching unigrams. When we try to expend the analysis to bigrams, it was computationally prohibitive on a single computer.

Chapter 5 Discussion

5.1. Discussion: Influenza Surveillance

Our first goal of this dissertation was to create an innovative strategy to improve flu surveillance by applying the modified BOCPD method to the ARGO real-time estimated flu activity data. To find a strategy that has optimal performance on predicting the onset of flu season prior to flu season beginning, we evaluated the performance by varying the α and the p of the following four strategies during the 2007 – 2015 seasons: (1) without restarting the BOCPD algorithm and using an uninformative prior for the data distribution; (2) restarting the BOCPD algorithm every year with an uninformative prior for the data distribution; (3) without restarting the BOCPD algorithm and using historical CDC %ILI data to estimate prior values; (4) restarting the BOCPD algorithm and re-estimating prior values using historical CDC %ILI data every year.

Our uninformative prior was different from a traditional uninformative prior defined in Bayesian inference which usually refers to a prior that assigns equal weight to all possible parameter values. We assumed that the flu activity data followed a known distribution but with unknown parameter values. We used the prior suggested by a previous study[57] as the uninformative prior, representing the situation where the prior was determined without a systematic principle.

We used the $\mathcal{M}$ defined as the follow,

$$\mathcal{M} = \frac{MAP(t-1) - MAP\ (t)}{MAP(t-1)} > \alpha,$$

which represented the relative decrease in the MAP for the current run length between two consecutive time points (i.e., time point t – 1 vs. t) to detect the occurrence of a change point. The α level indicated that if there was a sufficient amount of relative decrease in the MAP for the current run length. If a change point occurs at t, theoretically, the MAP should be at the run length 0 which makes the α level equals to 1, suggesting a change point is detected at the same time when it occurred. However, the posterior probability at the current run length 0 is estimated based on the same prior because the BOCPD algorithm assumes the same prior belief of the new data distribution. The prior belief may not always well describe the new data distribution because the change in the data distribution is random. As a result, the mode of the run length distribution, i.e., MAP, will never happen at run length 0, and thus the BOCPD algorithm cannot detect a change point at the same time when it occurred. Thus, a sufficiently large $\mathcal{M}$ would be evidence for a change point.

Furthermore, we used the following,

$$\mathcal{D} = \frac{epidemic\ baseline - ARGO\ \%ILI}{epidemic\ baseline} \leq p,$$

which represents the relative distance between the %ILI of a change point and the epidemic baseline to identify an informative change point. The threshold *p* indicates that if there is a sufficiently close relative distance between a change point and the epidemic baseline. We assumed that the flu activity tended to increase gradually, and the $\mathcal{D}$ would become smaller, eventually leading to the onset of flu season. Thus, before the flu season onset, a sufficiently small $\mathcal{D}$ of a change point detected would indicate that the flu activity has significantly increased toward the baseline and be the evidence for the onset of an imminent flu season.

For all strategies, our results indicated that the performance tended to become worse at those extremely high α levels, e.g., α = 0.8, given any *p*. It suggests that it should not use the extremely large $\mathcal{M}$ as the evidence for the occurrence of a change point for the flu activity data. For example, if the $\mathcal{M}$ is 0.8, it means that the MAP for the run length decreases from 5 to 1. However, because flu activity tends to increase gradually, the difference between two data distributions before and after a change point may be small. Thus, large $\mathcal{M}$ are unlikely. Therefore, setting α to an extremely large value will restrict the capability of the BOCPD algorithm to identify change points, resulting in the poor performance of prediction.

Moreover, we also found that for all strategies, given any α level, the performance was much better at the large *p* level (e.g., *p* = 0.4) compared to the small *p* level (e.g., *p* = 0.1). Figure 4 shows that if *p* = 0.1 and given the epidemic baseline is 2.2%, the %ILI of a change point detected should be 1.98% to be identified as informative to signal the onset of flu season. However, our results showed that before the flu season onset, the flu activity, which was extremely close to the epidemic baseline, e.g., the red dot in figure 7, was less likely to be detected as a change point. It may be due to the fact that a change point has already occurred before the flu activity becomes extremely close to the epidemic baseline. The flu activity like the red dot in figure 7 may be from the same data distribution as a change point (e.g., the red square dot in figure 7) just detected. Thus, they cannot be the signal (i.e., change points) of the beginning of the new distribution. Therefore, using an extremely small value of *p* will decrease the chance of identifying a change point which is informative to signal the onset of flu season.

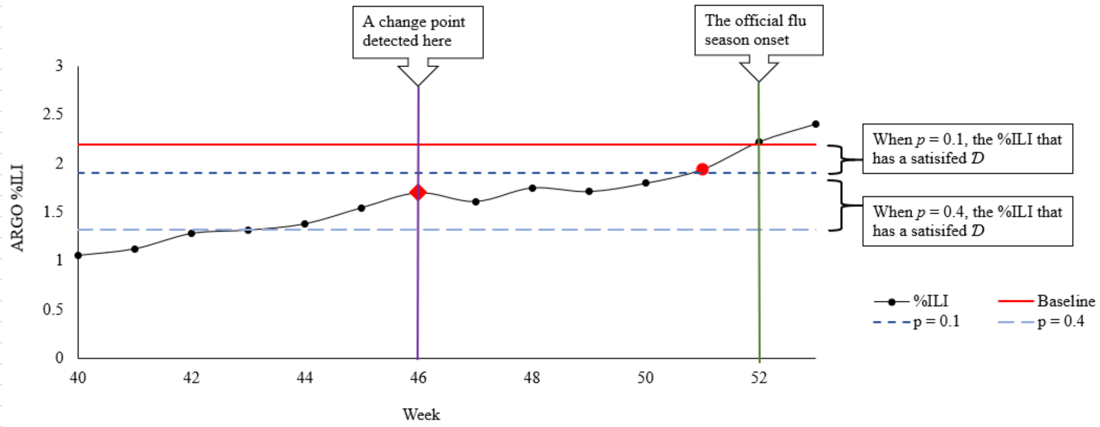

Figure 7. Illustration of the effect of using extremely small *p* level on identifying an informative change point.

Furthermore, our results showed that for strategies (1) and (3) both of which without restarting the BOCPD algorithm, the maximum proportion of correct prediction was 0.43 for the strategy (1) which used uninformative prior values, and was 0.86 for the strategy (3) which used informative prior. For strategies (2) and (4) both of which restarted the BOCPD algorithm every year, the maximum proportion of correct prediction was 0.57 and was 0.86 for the strategy (2) that used uninformative prior values and the strategy (4) that used informative prior, respectively. At their maximum proportion of correct prediction, strategies (1) and (2) incorrectly predicted at least 3 past regular flu seasons during the study period, while strategies (3) and (4) only falsely predicted 1 past regular flu season. The accuracy of predicting the imminent onset of flu season is the most critical benchmark to distinguish if a strategy is pragmatic for flu surveillance. Incorrectly predicting the onset of flu season may lead to insufficient flu season preparedness, resulting in hundreds of thousands of deaths, millions of hospitalizations, and hundreds of billions of dollars in direct and indirect costs.[62] Our findings suggested that strategies that used informative prior values had much better predictions of the onset of flu season compared to strategies that used noninformative prior values. Because the BOCPD algorithm uses the same prior for the entirety of the process, using uninformative prior may result in inaccurate estimates of the predictive probabilities of data, leading to a poor estimate of run length distribution. Without the accurate estimate of run length distribution, the performance of identifying change points that correctly predicted the onset of flu season will be dramatically deteriorated. These findings supported our hypothesis that incorporating the historical flu activity data collected by the CDC is an effective method to determine the prior, and thus make the strategy of providing early detection of flu season onset much more robust.

Comparing strategies (3) and (4), the overall performance of identifying change points that correctly predicted the flu season onset was similar. The maximum proportion of correct prediction was 0.86, which was the same for both strategies. Furthermore, at their maximum proportion of correct prediction, the average number of weeks between the correct change points and the CDC-reported date of flu season onset was almost the same for both strategies, 3.3 weeks for strategy (3) and 3.2 weeks for strategy (4).

For strategies that used informative prior, other than computation efficiency, the main difference between restarting and not restarting the detection process is in the support of the run length distribution. When the process does not restart, the number of the elements in the support is equivalent to the current time point *t*. Whereas, if the process restarts, the number of elements in the support is no more than 52, a 1-year period. Our results suggested that reducing the support of the run length distribution may have a different effect on strategies that use uninformative verse informative prior. While using informative prior, restarting the detection process did not optimize the performance i.e., strategies (3) and (4) performed similarly. However, restarting the detection process improved the performance of prediction while using an uninformative prior. One possible explanation is when using an informative prior, it does not matter that the support of the run length distribution keeps growing, because the probability mass on the extreme points of the support may be negligible. For example, table 15 shows that at t =100, the cumulative probability mass from run length 53 to the maximum possible run length 99 is much smaller in the strategy (3) compared to that of the strategy (1), and thus it is negligible. Figure 8 shows that if the probability mass on the extreme points of the support is negligible, there is a better chance that the run length distribution would be similar up to a certain time point (e.g., *t* = 53) between restarting verse no restarting. Therefore, there was no difference in the performance between strategies that restated or did not restart the process while using informative prior. However, while using uninformative prior, the probability mass on the extreme run lengths was not negligible, and thus the run length distribution would be very different up to a certain time point between restarting verse no restarting (Figure 9), resulting in different performance between these two strategies.

Table 15. Cumulative probability mass on the extreme points of support of run length for strategy (1) and strategy (3)

Strategy	P (53 ≤ r ≤ 99 | *t*=100)	P (53 ⩽ r ⩽ 199 | *t*=200)	P (53 ⩽ r ⩽ 299 | *t*=300)	P (53 ⩽ r ⩽ 399 | *t*=400)
(1)	1.169322e-14	7.91916e-21	2.679189e-22	1.275764e-09
(3)	2.011787e-17	2.628032e-27	2.21735e-28	1.014374e-23

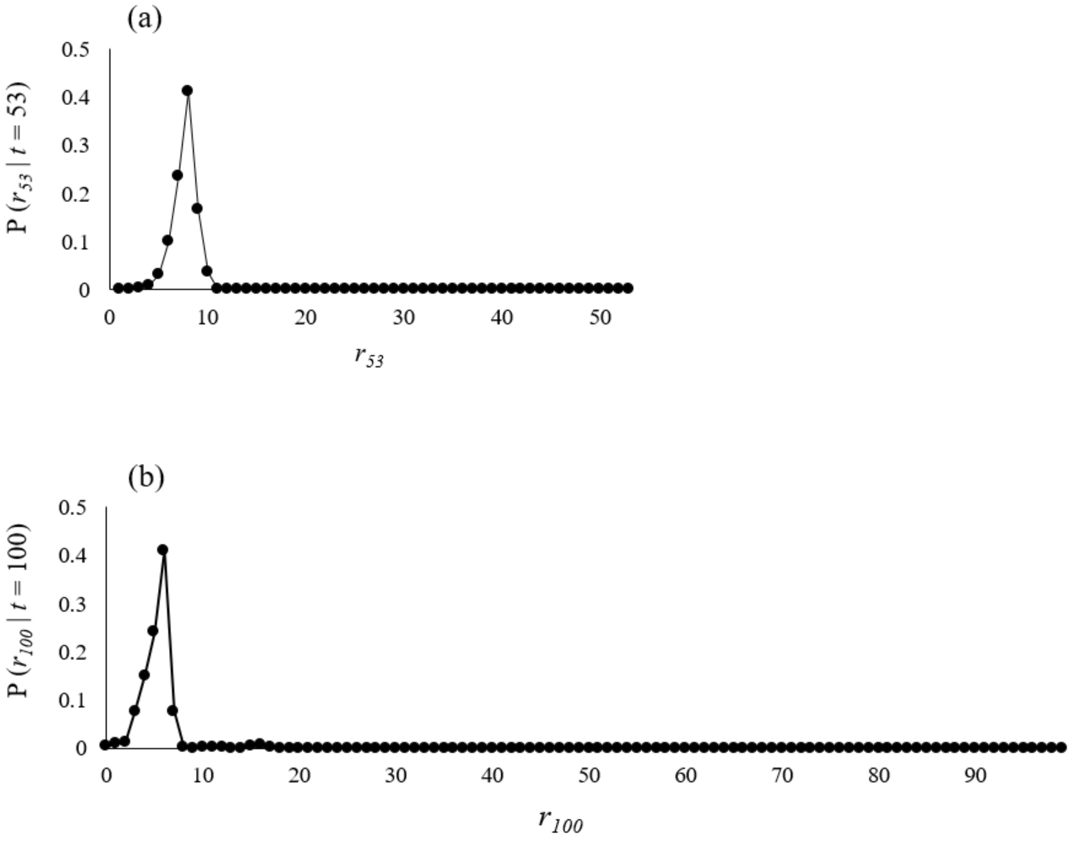

Figure 8. Posterior probability mass of the run length distribution for strategies that used informative prior: (a) strategy of restarting at *t* = 53; (b) strategy of not restarting at *t* =100.

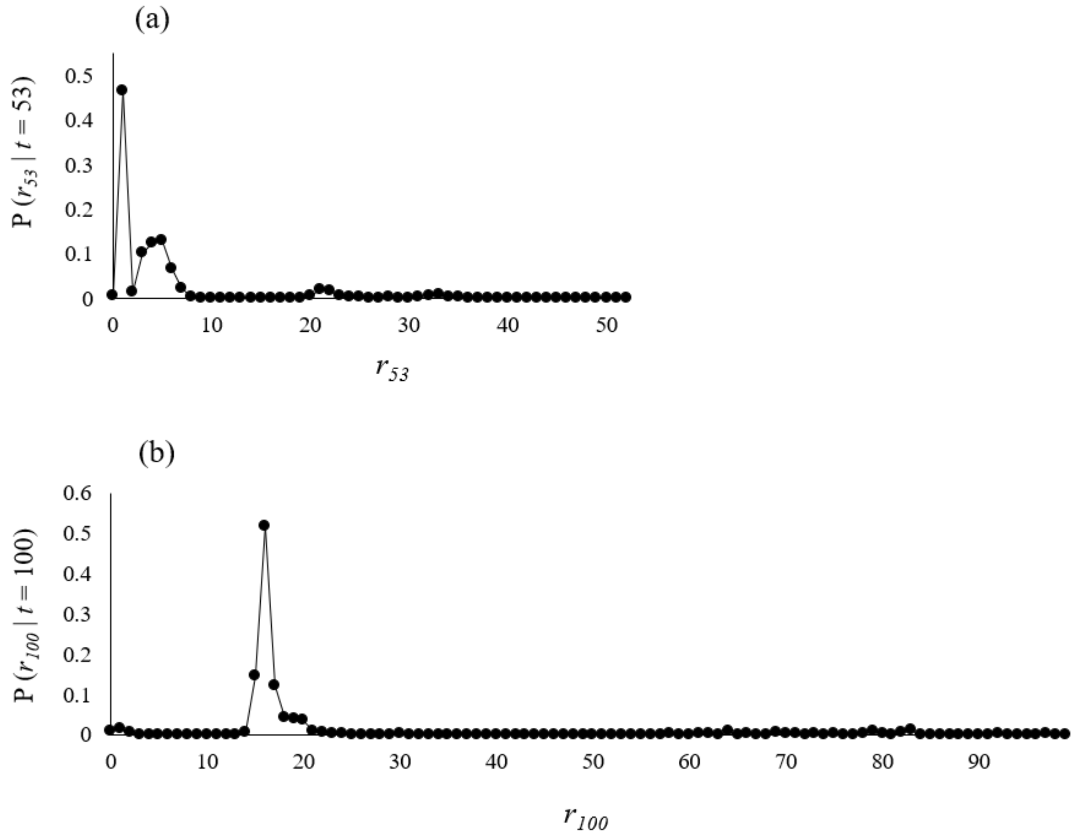

Figure 9. Posterior probability mass of the run length distribution for strategies that used uninformative prior: (a) strategy of restarting at t = 53; (b) strategy of not restarting at t =100.

Another possible explanation of why restarting the detection process did not optimize the performance may be the prior used for each season were relatively consistent. For this strategy, by just adding a few years to update the estimate of prior may not be helpful because flu activity pattern may be relatively similar from season to season during our study period. If compared to the flu activity pattern during centuries ago, the current flu activity pattern would be very different due to changes in climate, types of viruses, the effectiveness of vaccination, prevention intervention and so on. It may suggest that updating the prior with one additional year and using a cumulative sum of years may not dramatically change the estimate compare to the estimate of prior from the recent past. Thus, restarting the process and updating the prior did not improve the performance. In the future, studies will be needed to test if there is a benefit of restarting the process when prior values change more dramatically. However, restarting the detection process would make much more sense for flu surveillance because we will always have a chance to update the prior. Without restarting the process is not practically meaningful because we will keep using the same prior and losing the chance to use more appropriate prior, especially our results have showed the importance of using informative prior. The next of the study will try to explore different methods to estimate the prior to see if it is beneficial to optimize the prediction performance.

In this study, our optimal strategy of early detection of the imminent flu season onset exhibits a high accuracy of prediction. Meanwhile, the early warning signal detected in our optimal strategy has an average of 3-week lead time prior to the official onset of flu season. This lead time provides a critical period for health authorities to prepare and respond to a new flu season. Previous studies indicated that the rapid activation of flu interventions, such as reducing social and community contacts and increasing home isolation, may significantly prevent flu epidemic development.[3,63] Previous studies found that if the flu interventions was implemented within 2 weeks after the introduction of the first infectious case into the community, it will reduce the peak daily illness attack rates from 474 to below 35 cases per 10,000.[63] Moreover, our detection result is practically meaningful to improve public awareness of the current risk of flu to increase vaccination rates because flu vaccines are not fully effective until about 2 weeks after the shot.[44] Furthermore, previous studies also indicated that a 20% stockpile of a pre-prepared vaccine against the source avian virus could significantly reduce, even if its efficacy was low.[64] Applying our strategy, the health institutes will have a valuable lead time to stockpile the vaccine prior to a flu season begins.

There are several strengths to this study. First, we applied the ARGO model to the Google flu-related search queries data to obtain the real-time estimates of flu activity. Because the aggregated Google search query data are publicly available on a near real-time basis, and the ARGO model can accurately estimate flu activity using this data source, the resulting flu activity estimates are pragmatic for flu season onset detection compared to the gold standard of flu activity reporting. Second, we applied the modified BOCPD algorithm to the ARGO real-time estimated flu activity data to create a more practical strategy that can provide timely information on flu season onset. Previously, several attempts have been made to explore the feasibility of applying change point detection methods to provide early detection of a flu epidemic.[65,66] [67] However, most of these applications were practically meaningless for flu surveillance because the data sources used were not available in real-time and/or the change point detection algorithms used were only be able to identify the change points retrospectively, leading to the flu epidemic can be detected after the fact. Moreover, although many flu forecasting models existed which may potentially provide more information about impending epidemics, including the duration of the season, the overall burden, and the timing and magnitude of the epidemic peak.[68-70] However, these models were not typically designed for early warning or for detecting the onset of flu season. Thus, our study may potentially fill these knowledge gaps. Third, we established a systematic rule for detecting change points and a rule of identifying informative change points that may signal the onset of an imminent flu season. In current, there is no existing studies have ever attempted to tailor the BOCPD algorithm for flu surveillance. Our contributions show the feasibility of using the BOCPD algorithm to improve flu surveillance. Fourth, the general framework of our strategy can be extended to provide early detection of flu season onset at the regional level in the U.S. and even other countries. The seasonal flu has become one of the global health concerns due to the heavy burden it costs.[71] Worldwide, seasonal flu is estimated to result in at least 3 million cases of severe illness, and 290,000 respiratory deaths annually.[71] In addition, the effectiveness of the flu surveillance system in many other countries is also being impacted by the same limitations identified in the U.S. flu surveillance system.[72] Our framework provides a possible solution that can be adapted to improve flu surveillance globally.

There are limitations to this study that should be noted. First, our detection rule may not work well for a year like 2011 – 2012 season, which did not have the official onset of flu season at the national level based on the definition of flu season onset because the %ILI exceed the epidemic baseline for only one week (i.e., 2012 week 11). Ideally, no change points that signaled the onset of flu season should be detected for such seasons. However, for this season, our strategy was able to detect the single time point when the %ILI was above the baseline. Second, we only included 100 Google flu-related search terms to estimate flu activity. However, on average, the ARGO model only selected 14 terms out of 100 each week to estimate flu activity, suggesting that the most frequently used flu-related search terms may have included.[15]

In conclusion, this study provided evidence to support the feasibility of applying our modified BOCPD algorithm to the internet-based data to improve the surveillance of emerging seasonal flu in the U.S. We expanded the application of the BOCPD algorithms to flu surveillance. We created a flu surveillance strategy that combines search engine query data with the online change point detection algorithm has the power to predict the onset of the imminent flu season with valuable lead time. Such a strategy may provide valuable support for public health officials to take appropriate actions to prevent and control the spread of the seasonal flu epidemics. In the future, further improvements in our strategy may come from utilizing multiple internet-based data sources and extending our framework to the regional level.

5.2. Discussion: Probationers Absconding Surveillance

Our second goal of this dissertation was to explore words and phrases associated with probation absconders by applying natural language processing (NLP) techniques to official chronologic case notes written by probation officers. To find a practically useful strategy to explore the case notes data, we compared two text regression methods and applied them to the text data generated from the case notes of a random sample of adult misdemeanors and felony offenders who have received probation in Tarrant County, Texas. One method was the concise comparative summarization (CCS) method, and the other was the least absolute shrinkage and selection operator (LASSO) text regression. The LASSO method was chosen for comparison because it is one of the more commonly used text regression methods. In contrast, the CCS methods is relatively new, and its use and practicality are still being explored.

Our results indicated that the CCS method was much more practically useful to generate knowledge regarding probation absconders and completers compared to the LASSO method. The CCS method was designed to dynamically select words and phrases that are unimportant by exploiting the nested structure of any multiword phrase having a smaller phrase as a prefix. For example, if a unigram (single word) was found to be unimportant, all two-word phrases that begin with this unigram (i.e., its children) would not be considered. This method would result in an early stop of searching for texts that were not important to achieve a greater computational advantage. Meanwhile, the LASSO method requires that the full document-term matrix, i.e., matrix with frequency of all words and/or phrases being considered for the analysis, to be created before deeming words and phrases as important or unimportant. Because generating the full document-term matrix with different lengths of phrases is computationally tedious and the size of the resulting matrix increases exponentially, it makes the LASSO method rather impractical without considerable computational power. Thus, the computational efficiency of the LASSO is usually limited to unigrams or at best, very short phrases. For example, in our study, if we extended the application of the LASSO method to all bigrams (two-word phrases), both time and required computer memory would dramatically increase because the full document-term matrix would be over 20 gigabytes (GB). As such, we were not able to consider bigrams or above using the LASSO method.

Moreover, the final text regression model generated from the CCS method was more concise and easier to be interpreted than the LASSO method. The number of important words and phrases selected from the CCS method was 24 at maximum whereas the LASSO method selected more than 100 single words. As the model obtained from the CCS method contains not only single words but also multiword phrases, it provides richer information to summarize information that are associated with probation absconder and successful completer. However, due to the computational limitations of using the LASSO method, we were only able to capture single words, making it very difficult to generate useful knowledge. Because every single word has numerous possibilities to make up different meanings, it is hard to understand what it actually implies in the case notes, e.g., “screened” can be followed with “positive” or “negative” which would indicate two completely different scenarios. Therefore, compared to the LASSO method, our results suggested that the application of the CCS method is a much more practical strategy that can be used to discover commonalities in the case notes of probation absconders and completers.

The CCS method selects important words and phrases related to absconding status by minimizing a regularized loss function (i.e., a sum of a loss function and a penalty function of a vector of parameters),[37]

$$\widehat{\boldsymbol{\beta}} = arg \min_{\beta=(\beta_1,\ldots,\beta_p)} \mathcal{L}(\beta),$$

where

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\underbrace{\left(1 - y_i\left(\beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j\right)\right)}_{\text{squared hinge loss function}} \vee 0 \right]^2 + C \underbrace{\sum_{j=1}^{p} |\beta_j|}_{L^1 \text{ norm penalty}},$$

$\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ is the vector of coefficients for all words and phrases, and $a \vee b$ denoting the maximum of a and b; $x_{ij} = \frac{c_{ij}}{\sqrt{\sum_{i=1}^{N} c_{ij}^2}}$ is used to rescale word j differently based on how many times it appears in each document; N is total number of documents; C is a regularization parameter and $C \in [0, \infty]$; p is total number of all words and phrases.

The choice of the regularization parameter C plays an important role in finding a statistically significant text model. In the CCS method, a statistically significant text model means that the words and phrase selected are due to the relationship of the label and the text not due to the random chance.[37] Therefore, in this study we varied the C from C_1 to the C^{obs} in increments of 25 percentiles, where $C_1 = 8.8$, C_2= 12.5, C_3=16.3, and C_4=20.1, to explore different text models and its corresponding results. Here C^{obs} is the upper bound of C that gave an empty text regression model with no selected phrases (all coefficients were zero), given the original labels. C_1 was the minimum bound of C which indicated that we are 95% confident that the resulting non-empty text regression was due to the relationship of the outcome and the text, and not due to random chance.

When C was equal to the minimum bound $C_1 = 8.8$, the text model obtained from the CCS method provided the longest list of words and phrases compared to the other text models with higher C. It contained 10 and 14 words associated with probation absconders and completers, respectively. However, as C increases, the resulting text model will become shorter with fewer words to be selected. For example, when $C = C_4$=20.1, there were only 2 and 1 phrases that were associated with absconder and completer, respectively. For an exploratory study such as this, using a higher value of C, we may lose information in terms of commonalities in contents of case notes that may be critical to generate knowledge about probationers who tended to be absconding and successfully complete. Moreover, because there is no study that has analyzed the case notes data to explore words/phrases associated with probation outcomes, we do not have much guidance in ruling out words that are practically meaningless. Therefore, to provide a more conservative interpretation, we chose the CCS model using the smallest $C = C_1 = 8.8$ which produced the longest list of words/phrases.

In most text analysis, there is somewhat of an art form in interpreting the results. We compared the meanings of words and phrases found in this model between probation absconders and completers based on subjective, yet educated, judgement (Table 16), because there is no established standard way to interpret key words/phrases found in the case notes.

In probation absconders, words/phrases of "*technical violations*", "*technical*", "*technicals*" were found to be absconder-related. It suggested that probationers who had violated any probation conditions during probation were more likely to be absconding from supervision. This finding was consistent with a previous study which also found that more than half of probation absconders had violation experiences during their supervision.[23] Moreover, we also found words/phrases which may indicate what specific violations that the probation absconders tended to commit during their probation period. For example, "*cannabinoids*" suggested that absconders were more likely to violate substance use. This finding indicates that it is critical to develop and provide effective substance abuse/use treatment programs along with the probation supervision to probationers to correct such risky behaviors, and thus preventing probation absconding behavior. "*failed pay*", "*due status*", "*fee amount*" suggested that absconders tended to fail to pay the court-ordered fee by the due day. These findings suggested that probationers who had not stable economic status to pay court-ordered fees and fines were more likely to abscond. One possible explanation was the employment status. There was an existing study indicated that an unemployed probationer was more likely to abscond.[22] This may due to the fact that probationers who were unemployed were more likely to have financial issues, and thus more likely to fail to pay the fee as scheduled. Because they were afraid of getting additional punishment from court due to failure to pay, they chose to flee. To address this issue, supervision officers may need to encourage probationers to openly discuss their financial situation so that the officers can assist them to meet financial obligations, such as providing budgeting classes and special payment plans and asking the court to reduce or waive fees which may help reduce the incidences of absconding. Moreover, we also found "*attempts contact*" to be absconder-related. If a probationer failed to report as scheduled, the supervision officers would attempt to make contact with the probationer immediately by making phone calls, sending warning letters, and doing home visits. This finding suggested that the experience of failure to report may be an indicator of probation absconding. Therefore, officers may need to pay more attention to those who had experiences of failure to report by increasing face-to-face meetings or phone call contacts to prevent absconding. Furthermore, "*transfer intake*" and "*reported transfer*" were found to be associated with probation absconders. Transfer cases refer to the probationers who transfer their probation supervisions from the original counties or states where they were sentenced to probation to other places. The reasons for a probationer requested for transfer would be varied, such as change employers and residency. It is very interesting to find that probationers who were transfer cases tended more likely to be absconders. It may be due to the fact that those transfer probations lied to the officers that they moved to other places and needed to transfer their probation. However, in fact, they did not move to the new county where they were transferred to, and just took chances to run away. Further studies are needed to investigate what reasons cause those transfer cases to be absconding.

In probation completers, we found "*compliance*" as one of the keywords. It was contrary to the keyword of "*technical violations*" found in the absconders. It supports our finding that absconders were more likely to fail to comply with the probation conditions compare to completers. Moreover, we also found "*paid full*", "*current*", "*current fees*", "*paid forwarded*", "*fees paid full*", "*completed fees*", "*paid fees*", "*payment*", "*reported date*" and "*satisfied fees*" as keywords which suggested that completers tended to have a stable income to pay the court-order fees and thus they were more likely to complete their supervision. This finding supported our conclusion that probationers who failed to pay the fee were more likely to be absconding. It indicates that the financial status is a critical factor to affect the success of probation. Moreover, "*everything going well*", and "*travel*" were found as indicators of probation completion. It suggested that completers may have more positive attitudes and willingness to share their personal lives and feelings. This finding may contribute to informing the probation department that developing programs that promote the psychological health of probations is critical to help them to build a healthy and positive attitude towards life. Furthermore, it was very interesting that we found "*alcohol use*" as the key phrase associated with completers. It may indicate that completers tended to be more honest to report to the officers about their alcohol use behavior during probation. This may be another way to reflect the fact that completers may trust probation officers more to tell their stories compared to absconders.

Table 16. Comparison of words/phrases found in the CCS method with $C = C_1 = 8.8$ between absconders and completers

Probation absconders	Probation completers
technical violations, technical, technical, cannabinoids, failed pay, due status, fee amount, attempts contact, transfer intake, reported transfer,	paid full, current, current fees, paid forwarded, fees paid full, completed fees, compliance, payment, paid fees, satisfied fees, everything going well, travel, reported date, alcohol use

In this study, our interpretation of the CCS model primarily focused on the sign of the coefficient instead of the magnitude. Conventionally, a model-based interpretation for the coefficients β_j in the regression model would be that for a 1 unit change in the count of feature j, the predictive outcome would change by β_j holding other words and phrases constant. However, applying this model-based interpretation to a text regression model would be problematic and misleading. Because the case notes data has a free-text nature, the frequency of words/phrases used in each document would be less likely to be constant across the sample. Therefore, the interpretation of the magnitude of coefficient would be meaningless. Moreover, there is no previous evidence to suggest that what words/phrases would be important to predict the outcomes. Although words/phrases found in the CCS model may have small magnitudes of coefficients, it may not necessarily indicate that such words/phrases are less important; it may be due to the fact that such words/phrases were less likely to be used in the case notes, or vice versa. Therefore, in our case, interpretation of the sign instead of the magnitude of coefficients would be more useful. Our interpretation would provide general knowledge to probation officers about what words and phrases were “red flags”.

However, Miratrix and Ackerman indicated that blindly interpreting the sign of the coefficients could be problematic due to the potential multicollinearity issue in the data. For example, a negative coefficient for a word/phrase would offset the positive coefficient for its highly correlated alternate word/phrase, when in fact both words/phrases may have a positive association with the outcome.[37] In this case, interpreting the negative sign would be inappropriate. We applied the proposed solution suggested by Miratrix and Ackerman[37] to verify those words/phrases we found were not affected by the multicollinearity issue by considering only the set of positive coefficient. Table 17 showed that a majority of positive words/phrases found in the CCS model after forcing negative coefficients to not exist were the same as those found in the uncontrolled CCS model (allowing negative words/phrases). It implies that those words/phrases we found would be the truly distinct language used in the documents.

Table 17. The CCS models only considering positive words and phrases varied by different C values.

Words and phrases	$\boldsymbol{C_1}$**= 8.5**	$\boldsymbol{C_2}$**= 12.4**	$\boldsymbol{C_3}$**=16.2**	$\boldsymbol{C_4}$**=20.1**
transfer intake				
technical violations				
due status			NA	NA
reported transfer			NA	NA
attempts contact		NA	NA	NA
fee amount		NA	NA	NA
technicals		NA	NA	NA
arrears supervision fees amount		NA	NA	NA
failure pay failure		NA	NA	NA

NA indicates this word was not selected.
* Words and phrases were listed from the highest coefficients to the lowest.

Furthermore, in this study, we did not compare the accuracy of prediction between the CCS and the LASSO methods. The accuracy of prediction measures how well a predictive model predicts future outcomes. The goal of a predictive model is to use the associations between predictors and the outcome to generate good predictions for future outcomes. However, the CCS method was developed to summarize the important words and phrases were related to the outcome, not for prediction. Therefore, it is not surprising that the predictive accuracy is low. Miratrix and Ackerman[37] have shown that even with substantially large data to estimate the CCS model, the predictive accuracy was only about 20%. The CCS model is more like explanatory modeling that focuses on identifying variables (e.g., words/phrases) that have statistically significant relationships with an outcome (e.g., probation absconding). This type of model is helpful to find out the explanations of why absconding behavior happen and inform probation officers what possible actions should be carried out to prevent absconding.

In addition to the above, there are several other limitations to this study. First, we did not take the timeline of each case note into account. The timeline could be an important factor to tell us how probationers' behaviors change over time. However, to date, there is no exiting text regression method that is developed for longitudinal application. In the future, studies are needed to extend the NLP techniques to the longitudinal text data. Second, our data was only limited to one county in the U.S. Our results may not be generalizable to other counties or states. However, our study provides a strategy that probation officers from all over the country can adopt to utilize the case notes as data to conduct research of the probation population systematically.

Despite some of the limitations, this study has the following strengths. First, this is the first study in Texas and maybe nationwide, that applied NLP techniques to text data generated from chronological case notes to explore contents associated with probation absconders and completers. To our knowledge, no study has utilized this type of data from the probation system with the application of NLP techniques to investigate factors are associated with probation outcomes. Thus, our study discovered commonalities in the case notes of absconders and completers to fill these knowledge gaps. Second, we applied an innovative text regression method, the CCS, which can analyze large-scale text data with extremely high computational efficiency. Previously, several attempts have been made to develop an efficient text regression method. However, most of them have expensive computation cost and are not flexible to analyze multiword phrases, such as the LASSO text method we used in this study.[73] Third, many of our findings were consistent with previous studies that utilized numerical data. In addition, our results provided critical hints to officers about previously untapped factors that were potentially linked to absconders and completers, such as transfer status. Thus, our study provides a possible strategy for the probation department to use case notes data systematically.

In conclusion, this study provided evidence to support the feasibility of applying the CCS method as a strategy for exploring risk factors related to probation outcomes. Currently, the case notes are kept only for record-keeping purposes. Developing a strategy of utilizing the case notes systematically is critically meaningful to contribute to a new surveillance system to prevent the incidence of probation absconding. The commonalities found in the case notes of absconders and completers play an important role in understanding what makes the probations successful or failing, and thus inform priorities to improve supervision practice to achieve the goals of reintegration and public safety. In the future, we may focus on exploring contents associated with absconders and completers stratified by felony and misdemeanor offenders.

Appendix A: Supplemental results for Aim 1

Table S1a. Results for strategy (1) without restarting the BOCPD algorithm and using uninformative prior by alpha level, *p* = 0.5, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)
	2008 - 2009	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)	wk 41 (incorrect)
	2011 - 2012	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)
	2012 - 2013	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)
	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	None	None	None	None	None	None	None	None

Proportion of correct prediction		0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.29
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	6	6	6	6	6	6	3	3
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	7	7	7	7	7	7	7	7
	2013 - 2014	3	3	3	3	3	3	3	None
	2014 - 2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		5.3	5.3	5.3	5.3	5.3	5.3	4.3	4.3

Table S1b. Results for strategy (1) without restarting the BOCPD algorithm and using uninformative prior by alpha level, *p* = 0.4, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**

Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)
	2012 - 2013	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	None	None
	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.43	0.43	0.43	0.43	0.43	0.43	0.29	0.14
Distance between correct change	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	3	3	3	3	3	3	3	3
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None

points and the official date of onset (weeks)	**2012 - 2013**	5	5	5	5	5	5	None	None
	2013 - 2014	3	3	3	3	3	3	3	None
	2014 - 2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3.7	3.7	3.7	3.7	3.7	3.7	3	3

Table S1c. Results for strategy (1) without restarting the BOCPD algorithm and using uninformative prior by alpha level, *p* = 0.3, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)
	2012 - 2013	None	None	None	None	None	None	None	None

	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	3	3	3	3	3	3	3	3
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	3	3	3	3	3	3	3	None
	2014 - 2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S1d. Results for strategy (1) without restarting the BOCPD algorithm and using uninformative prior by alpha level, *p* = 0.2, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Distance	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	3	3	3	3	3	3	3	3

between correct change points and the official date of onset (weeks)	**2010 - 2011**	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S1e. Results for strategy (1) without restarting the BOCPD algorithm and using uninformative prior by alpha level, *p* = 0.1, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 10	wk 10	wk 10	wk 10	wk 10	wk 10	wk 10	None

in each season		(incorrect)	(incorrect)	(incorrect)	(incorrect)	(incorrect)	(incorrect)	(incorrect)	
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	3	3	3	3	3	3	3	3
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S2a. Results for strategy (2) restarting the BOCPD algorithm every year and using uninformative prior for the data distribution, *p* = 0.5, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)	wk 43 (incorrect)
	2008 - 2009	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 51 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)	wk 42 (incorrect)
	2012 - 2013	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)
	2013 - 2014	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)
	2014 - 2015	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)

Proportion of correct prediction		0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008-2009	6	6	6	6	6	6	3	3
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	7	7	7	7	7	7	7	7
	2013-2014	7	7	7	7	7	7	7	7
	2014-2015	3	3	3	3	3	3	3	3
Average of distance (weeks)		5.8	5.8	5.8	5.8	5.8	5.8	5	5

Table S2b. Results for strategy (2) restarting the BOCPD algorithm every year and using uninformative prior for the data distribution, *p* = 0.4, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)
	2012 - 2013	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)	wk 43 (correct)
	2013 - 2014	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)	wk 41 (correct)
	2014 - 2015	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)

Proportion of correct prediction		0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008-2009	3	3	3	3	3	3	3	3
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	5	5	5	5	5	5	5	5
	2013-2014	7	7	7	7	7	7	7	7
	2014-2015	3	3	3	3	3	3	3	3
Average of distance (weeks)		4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5

Table S2c. Results for strategy (2) restarting the BOCPD algorithm every year and using uninformative prior for the data distribution, *p* = 0.3, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)

Proportion of correct prediction		0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008-2009	3	3	3	3	3	3	3	3
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	None	None	None	None	None	None	None	None
	2013-2014	None	None	None	None	None	None	None	None
	2014-2015	3	3	3	3	3	3	3	3
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S2d. Results for strategy (2) restarting the BOCPD algorithm every year and using uninformative prior for the data distribution, *p* = 0.2, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	2009 week 1	2009 week 1	2009 week 1	2009 week 1	2009 week 1	2009 week 1
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)	wk 51 (incorrect)
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct		0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

prediction									
Distance between correct change points and the official date of onset (weeks)	2007 - 2008	None	None	None	None	None	None	None	None
	2008-2009	3	3	3	3	3	3	3	3
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	None	None	None	None	None	None	None	None
	2013-2014	None	None	None	None	None	None	None	None
	2014-2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S2e. Results for strategy (2) restarting the BOCPD algorithm every year and using uninformative prior for the data distribution, *p* = 0.1, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)	wk 1 (correct)
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	wk 10 (incorrect)	wk 10 (incorrect)	wk 10 (incorrect)	wk 10 (incorrect)	wk 10 (incorrect)	wk 10 (incorrect)	wk 10 (incorrect)	None
	2012 - 2013	None	None	None	None	None	None	None	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct		0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

prediction									
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008-2009	3	3	3	3	3	3	3	3
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	None	None	None	None	None	None	None	None
	2013-2014	None	None	None	None	None	None	None	None
	2014-2015	None	None	None	None	None	None	None	None
Average of distance (weeks)		3	3	3	3	3	3	3	3

Table S3a. Results for strategy (3) without restarting the BOCPD algorithm and using historical CDC ILI data to estimate prior values, *p* = 0.5, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2008 - 2009	wk 48 (incorrect)	wk 48 (incorrect)	wk 48 (incorrect)	wk 48 (incorrect)	wk 48 (incorrect)	wk 48 (incorrect)	wk 48 (incorrect)	wk 53 (correct)
	2010 - 2011	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2011 - 2012	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 52 (incorrect)
	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)
	2013 - 2014	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2014 - 2015	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
Proportion of correct prediction		0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.29

Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008 - 2009	None	None	None	None	None	None	None	4
	2010 - 2011	5	5	5	5	5	5	5	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	1	1	1	1	1	1	1	1
	2013 - 2014	2	2	2	2	2	2	2	None
	2014 - 2015	1	1	1	1	1	1	1	None
Average of distance (weeks)		3.2	3.2	3.2	3.2	3.2	3.2	3.2	2.5

Table S3b. Results for strategy (3) without restarting the BOCPD algorithm and using historical CDC ILI data to estimate prior values, *p* = 0.4, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
	2007 - 2008	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None

Informative change point identified in each season	**2008 - 2009**	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2011 - 2012	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 46 (incorrect)	wk 52 (incorrect)
	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)
	2013 - 2014	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2014 - 2015	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
Proportion of correct prediction		0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.29
Distance between correct	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008 - 2009	4	4	4	4	4	4	4	4
	2010 - 2011	5	5	5	5	5	5	5	None

change points and the official date of onset (weeks)	**2011 - 2012**	None	None	None	None	None	None	None	None
	2012 - 2013	1	1	1	1	1	1	1	1
	2013 - 2014	2	2	2	2	2	2	2	None
	2014 - 2015	1	1	1	1	1	1	1	None
Average of distance (weeks)		3.3	3.3	3.3	3.3	3.3	3.3	3.3	2.5

Table S3c. Results for strategy (3) without restarting the BOCPD algorithm and using historical CDC ILI data to estimate prior values, *p* = 0.3, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2008 - 2009	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)2008 week 53
	2010 - 2011	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	None	None	None	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)

	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)
	2013 - 2014	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2014 - 2015	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
Proportion of correct prediction		0.86	0.86	0.86	0.71	0.71	0.71	0.71	0.29
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008 - 2009	4	4	4	4	4	4	4	4
	2010 - 2011	1	1	1	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	1	1	1	1	1	1	1	1
	2013 - 2014	2	2	2	2	2	2	2	None
	2014 - 2015	1	1	1	1	1	1	1	None

Average of distance (weeks)	2.7	2.7	2.7	3	3	3	3	2.5

Table S3d. Results for strategy (3) without restarting the BOCPD algorithm and using historical CDC ILI data to estimate prior values, *p* = 0.2, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	None	None	None	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)
	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)
	2013 - 2014	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None
	2014 - 2015	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	None

Proportion of correct prediction		0.71	0.71	0.71	0.57	0.57	0.57	0.57	0.29
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	4	4	4	4	4	4	4	4
	2010 - 2011	1	1	1	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	1	1	1	1	1	1	1	1
	2013 - 2014	2	2	2	2	2	2	2	None
	2014 - 2015	1	1	1	1	1	1	1	None
Average of distance (weeks)		1.8	1.8	1.8	2	2	2	2	2.5

Table S3e. Results for strategy (3) without restarting the BOCPD algorithm and using historical CDC ILI data to estimate prior values, *p* = 0.1, 2007 – 2015 seasons

	α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**

Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	None	None	None	None	None	None	None	None
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
Distance between correct change points and the official date of	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	None	None	None	None	None	None	None	None
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 - 2012	None	None	None	None	None	None	None	None
	2012 - 2013	1	1	1	1	1	1	1	1
	2013 - 2014	None	None	None	None	None	None	None	None

onset (weeks)	**2014 - 2015**	None	None	None	None	None	None	None	None
Average of distance (weeks)		1	1	1	1	1	1	1	1

Table S4a. Results for strategy (4) restarting the BOCPD algorithm every year and using historical CDC ILI data estimate prior values, *p* = 0.5, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2008 - 2009	wk 47 (incorrect)	wk 47 (incorrect)	wk 47 (incorrect)	wk 47 (incorrect)	wk 47 (incorrect)	wk 47 (incorrect)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	wk 44 (correct)	None	None
	2011 - 2012	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 52 (incorrect)
	2012 -	wk 46	wk 46	wk 46	wk 46	wk 46	wk 46	wk 47	None

	2013	(correct)	(correct)	(correct)	(correct)	(correct)	(correct)	(correct)	
	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
Proportion of correct prediction		0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008-2009	None	None	None	None	None	None	4	4
	2010-2011	7	7	7	7	7	7	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	2	2	2	2	2	2	1	None
	2013-	3	3	3	3	3	3	3	None

	2014								
	2014-2015	2	2	2	2	2	2	2	None
Average of distance (weeks)		4.2	4.2	4.2	4.2	4.2	4.2	3.4	4

Table S4b. Results for strategy (4) restarting the BOCPD algorithm every year and using historical CDC ILI data estimate prior values, *p* = 0.4, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2008 - 2009	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	None	None
	2011 - 2012	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 45 (incorrect)	wk 52 (incorrect)

	2012 - 2013	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 47 (correct)	None
	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
Proportion of correct prediction		0.86	0.86	0.86	0.86	0.86	0.86	0.71	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008-2009	4	4	4	4	4	4	4	4
	2010-2011	1	1	1	1	1	1	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	2	2	2	2	2	2	1	None

	2013-2014	3	3	3	3	3	3	3	None
	2014-2015	2	2	2	2	2	2	2	None
Average of distance (weeks)		3.2	3.2	3.2	3.2	3.2	3.2	3.4	4

Table S4c. Results for strategy (4) restarting the BOCPD algorithm every year and using historical CDC ILI data estimate prior values, *p* = 0.3, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2008 - 2009	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)

	2012 - 2013	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 47 (correct)	None
	2013 - 2014	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
	2014 - 2015	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
Proportion of correct prediction		0.86	0.86	0.86	0.86	0.86	0.86	0.71	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	7	7	7	7	7	7	7	None
	2008-2009	4	4	4	4	4	4	4	4
	2010-2011	1	1	1	1	1	1	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	2	2	2	2	2	2	1	None

	2013-2014	3	3	3	3	3	3	3	None
	2014-2015	2	2	2	2	2	2	2	None
Average of distance (weeks)		3.2	3.2	3.2	3.2	3.2	3.2	3.4	4

Table S4d. Results for strategy (4) restarting the BOCPD algorithm every year and using historical CDC ILI data estimate prior values, *p* = 0.2, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each season	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)	wk 53 (correct)
	2010 - 2011	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	wk 50 (correct)	None	None
	2011 - 2012	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)	wk 52 (incorrect)

	2012 - 2013	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 46 (correct)	wk 47 (correct)	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	wk 45 (correct)	None
Proportion of correct prediction		0.57	0.57	0.57	0.57	0.57	0.57	0.29	0.14
Distance between correct change points and the official date of onset (weeks)	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008-2009	4	4	4	4	4	4	4	4
	2010-2011	1	1	1	1	1	1	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	2	2	2	2	2	2	1	None

	2013-2014	None	None	None	None	None	None	None	None
	2014-2015	2	2	2	2	2	2	2	None
Average of distance (weeks)		2.3	2.3	2.3	2.3	2.3	2.3	2.3	4

Table S4e. Results for strategy (4) restarting the BOCPD algorithm every year and using historical CDC ILI data estimate prior values, *p* = 0.1, 2007 – 2015 seasons

		α = 0.1	**α = 0.2**	**α = 0.3**	**α = 0.4**	**α = 0.5**	**α = 0.6**	**α = 0.7**	**α = 0.8**
Informative change point identified in each	**2007 - 2008**	None	None	None	None	None	None	None	None
	2008 - 2009	None	None	None	None	None	None	None	None
	2010 - 2011	None	None	None	None	None	None	None	None
	2011 -	None	None	None	None	None	None	None	None

season	**2012**								
	2012 - 2013	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	wk 47 (correct)	None
	2013 - 2014	None	None	None	None	None	None	None	None
	2014 - 2015	None	None	None	None	None	None	None	None
Proportion of correct prediction		0.14	0.14	0.14	0.14	0.14	0.14	0.14	None
Distance between correct change points and the official date of onset (weeks)	**2007 -2008**	None	None	None	None	None	None	None	None
	2008-2009	None	None	None	None	None	None	None	None
	2010-2011	None	None	None	None	None	None	None	None
	2011-2012	None	None	None	None	None	None	None	None
	2012-2013	1	1	1	1	1	1	1	None
	2013-2014	None	None	None	None	None	None	None	None
	2014-2015	None	None	None	None	None	None	None	None

Average of distance (weeks)		1	1	1	1	1	1	1	None

References

1. Rolfes MA, Foppa IM, Garg S, et al. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. 2016.
2. Buehler JW. CDC's vision for public health surveillance in the 21st century. Introduction. 2012(2380-8942 (Electronic)).
3. Organization WH. WHO guidance for surveillance during an influenza pandemic. 2017.
4. Initiative EP. FluSight 2017–2018; 2018. *URL: https://predict phiresearchlab org/post/59973fe26f7559750d84a843.*
5. Control CfD, Prevention. Overview of Influenza Surveillance in the United States. 2013. In:2013.
6. Sharpe JD, Hopkins RS, Cook RL, Striley CWJJph, surveillance. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. 2016;2(2).
7. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. Paper presented at: AMIA Annual Symposium Proceedings2006.
8. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant LJN. Detecting influenza epidemics using search engine query data. 2009;457(7232):1012.
9. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RAJCid. Using internet searches for influenza surveillance. 2008;47(11):1443-1448.
10. Fox S, Duggan M. Health online 2013. *Health.* 2013;2013:1-55.
11. Carneiro HA, Mylonakis EJCid. Google trends: a web-based tool for real-time surveillance of disease outbreaks. 2009;49(10):1557-1564.
12. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CHJPo. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. 2011;6(4):e18687.
13. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JSJPcb. Combining search, social media, and traditional data sources to improve influenza surveillance. 2015;11(10):e1004513.
14. Santillana MJCID. Perspectives on the Future of Internet Search Engines and Biosurveillance Systems. 2016:ciw660.
15. Yang S, Santillana M, Kou SCJPotNAoS. Accurate estimation of influenza epidemics using Google search data via ARGO. 2015;112(47):14473-14478.
16. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SJBid. Using electronic health records and Internet search information for accurate influenza forecasting. 2017;17(1):332.
17. Aminikhanghahi S, Cook DJJK, systems i. A survey of methods for time series change point detection. 2017;51(2):339-367.
18. Texier G, Farouh M, Pellegrin L, et al. Outbreak definition by change point analysis: a tool for public health decision? 2016;16(1):33.
19. Mayzer R, Gray MK, Maxwell SRJJoCJ. Probation absconders: A unique risk group? 2004;32(2):137-150.
20. May TP. Probation: Politics, policy and practice. 1990.
21. Kaeble D. Probation and Parole in the United States, 2016. 2018.
22. Stevens-Martin K, Liu JJFP. Fugitives from Justice: An Examination of

Felony and Misdemeanor Probation Absconders in a Large Jurisdiction. 2017;81:41.

23. Stevens-Martin K, Oyewole O, Hipolito CJFP. Technical revocations of probation in one jurisdiction: Uncovering the hidden realities. 2014;78:16.
24. Finn MA, Prevost JP, Braucht GS, et al. Home visits in community supervision: a qualitative analysis of theme and tone. 2017;44(10):1300-1316.
25. Liddy ED. Natural language processing. 2001.
26. Chowdhury GGJArois, technology. Natural language processing. 2003;37(1):51-89.
27. Munot N, Govilkar SSJIJoCA. Comparative study of text summarization methods. 2014;102(12).
28. Blei D, Carin L, Dunson DJIspm. Probabilistic topic models. 2010;27(6):55-65.
29. Nsubuga P, White ME, Thacker SB, et al. Public health surveillance: a tool for targeting and monitoring interventions. 2006;2:997-1018.
30. Groseclose SL, Buckeridge DLJAroph. Public health surveillance systems: recent advances in their use and evaluation. 2017;38:57-79.
31. Lee LM, Thacker SB, Louis MES. *Principles and practice of public health surveillance.* Oxford University Press, USA; 2010.
32. Mirza N, Reynolds T, Coletta M, et al. Steps to a Sustainable Public Health Surveillance Enterprise A Commentary from the International Society for Disease Surveillance. 2013;5(2):210.
33. Morse SSJB, bioterrorism: biodefense strategy p, science. Public health surveillance and infectious disease detection. 2012;10(1):6-16.
34. Martin-Sanchez F, Verspoor KJYomi. Big data in medicine is driving big changes. 2014;9(1):14.
35. Habl C, Renner A, Bobek J, Laschkolnig AJSoBDiPH, Telemedicine, Report. HF. Study on Big Data in Public Health, Telemedicine and Healthcare. Final Report. 2016.
36. Santillana M, Nguyen A, Louie T, et al. Cloud-based electronic health records for real-time, region-specific influenza surveillance. 2016;6:25732.
37. Miratrix L, Ackerman RJSA, Journal DMTADS. Conducting sparse feature selection on arbitrarily long phrases in text corpora with a focus on interpretability. 2016;9(6):435-460.
38. Baseman JG, Revere D, Painter I. Big Data in the Era of Health Information Exchanges: Challenges and Opportunities for Public Health. Paper presented at: Informatics2017.
39. Liang Y, Kelemen AJAJoB, Biostatistics. Big Data science and its applications in health and medical research: Challenges and opportunities. 2016;7(3).
40. Manogaran G, Lopez D. Disease surveillance system for big climate data processing and dengue transmission. In: *Climate Change and Environmental Concerns: Breakthroughs in Research and Practice.* IGI Global; 2018:427-446.
41. Organization WH. Influenza (seasonal). Fact sheet no. 211. *World Health Organization, Geneva, Switzerland http://www who int/mediacentre/factsheets/fs211/en/index html.* 2009.
42. Organization WH. Influenza (seasonal). fact sheet no 211. 2014. *Available from: left angle bracket http://www who int/mediacentre/factsheets/fs211/en/right angle bracket.* 2014.

43. Control CfD, April PJU. Estimated influenza illnesses, medical visits, hospitalizations, and deaths averted by vaccination in the United States. 2017;19.

44. Control CfD, Prevention. Key facts about seasonal flu vaccine. 2015. In:2015.

45. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science.* 2014;343(6176):1203-1205.

46. Joh EE. The new surveillance discretion: Automated suspicion, big data, and policing. *Harv L & Pol'y Rev.* 2016;10:15.

47. Krisberg BA, Marchionna S, Hartney CJ. *American corrections: Concepts and controversies.* Sage Publications; 2018.

48. Levinson D. *Encyclopedia of crime and punishment.* Vol 1: Sage; 2002.

49. Domestic VNA, Violence S. Vermont Department of Corrections. 2010.

50. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad NJJJoAIDS. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. 2018;77(2):160-166.

51. Mohebbi M, Vanderkam D, Kodysh J, Schonberger R, Choi H, Kumar S. Google correlate whitepaper. 2011.

52. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS one.* 2011;6(8):e23610.

53. Adams RP, MacKay DJJapa. Bayesian online changepoint detection. 2007.

54. Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis.* Chapman and Hall/CRC; 2013.

55. Murphy KPJd. Conjugate Bayesian analysis of the Gaussian distribution. 2007;1(2σ2):16.

56. Turner R, Saatci Y, Rasmussen CE. Adaptive sequential Bayesian change point detection. 2009.

57. Byrd M, Nghiem L, Cao J. Lagged exact Bayesian online changepoint detection. *arXiv preprint arXiv:171003276.* 2017.

58. Hawkins DM. The problem of overfitting. *Journal of chemical information and computer sciences.* 2004;44(1):1-12.

59. Tibshirani RJJotRSSSB. Regression shrinkage and selection via the lasso. 1996:267-288.

60. Tibshirani R, Wasserman L. A Closer Look at Sparse Regression. In:2016.

61. Moore R, DeNero J. L1 and L2 regularization for multiclass hinge loss models. Paper presented at: Symposium on Machine Learning in Speech and Language Processing2011.

62. Griffin BA, Jain AK, Davies-Cole J, et al. Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system. 2009;9(1):483.

63. Kelso JK, Milne GJ, Kelly H. Simulation suggests that rapid activation of social distancing can arrest epidemic development due to a novel strain of influenza. *BMC public health.* 2009;9(1):117.

64. Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. *Nature.* 2006;442(7101):448.

65. Baron M, Antonov V, Huber C, et al. Early detection of epidemics as a sequential change-point problem. 2004:7-9.

66. Kass-Hout TA, Xu Z, McMurray P, et al. Application of change point analysis to daily influenza-like illness emergency department visits. 2012;19(6):1075-1081.

67. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. *JMIR public health and surveillance.* 2016;2(2):e161.

68. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch MJNc. Real-time influenza forecasts during the 2012–2013 season. 2013;4:2837.

69. Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology*. 2015;11(8):e1004382.

70. Ertem Z, Raymond D, Meyers LA. Optimal multi-source forecasting of seasonal influenza. *PLoS computational biology*. 2018;14(9):e1006236.

71. Nair H, Brooks WA, Katz M, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet.* 2011;378(9807):1917-1930.

72. Clemente L, Lu F, Santillana M. Improved Real-Time Influenza Surveillance: Using Internet Search Data in Eight Latin American Countries. *JMIR public health and surveillance.* 2019;5(2):e12214.

73. Conroy B, Sajda P. Fast, exact model selection and permutation testing for l2-regularized logistic regression. Paper presented at: Artificial Intelligence and Statistics2012.