Statistics
in Medicine

Research Article

Received 11 December 2012, Accepted 10 February 2014 Published online 6 March 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6129

G-scores: A method for identifying disease-causing pathogens with application to lower respiratory tract infections

Peng Zhang,[a] Peichao Peng,[b] Lu Wang[c*†] and Yu Kang[d]

Lower respiratory tract infections (LRTIs) are well known for the lack of a good diagnostic method. The main difficulty lies in the fact that there are a variety of pathogens causing LRTIs, and their management and treatment are quite different. The development of quantitative real-time loop-mediated isothermal amplification (qrt-LAMP) made it possible to rapidly amplify and quantify multiple pathogens simultaneously. The question that remains to be answered is how accurate and reliable is this method? More importantly, how are qrt-LAMP measurements utilized to inform/suggest medical decisions? When does a pathogen start to grow out of control and cause infection? Answers to these questions are crucial to advise treatment guidance for LRTIs and also helpful to design phase I/II trials or adaptive treatment strategies. In this article, our main contributions include the following two aspects. First, we utilize zero-inflated mixture models to provide statistical evidence for the validity of qrt-LAMP being used in detecting pathogens for LRTIs without the presence of a gold standard test. Our results on qrt-LAMP suggest that it provides reliable measurements on pathogens of interest. Second, we propose a novel statistical approach to identify disease-causing pathogens, that is, distinguish the pathogens that colonize without causing problems from those that rapidly grow and cause infection. We achieve this by combining information from absolute quantities of pathogens and their symbiosis information to form *G*-scores. Change-point detection methods are utilized on these *G*-scores to detect the three phases of bacterial growth—lag phase, log phase, and stationary phase.

Keywords: change point; Gibbs sampling; loop-mediated isothermal amplification; Markov chain Monte Carlo; Tobit model; zero-inflated models

1. Introduction

The diagnosis and treatment of lower respiratory tract infections (LRTIs), one of the most common diseases worldwide, are still rather empirical. The main approach to testing for types of pathogens as the cause of LRTIs is standard culture method (SCM), which however has many drawbacks. First of all, SCM is not accurate enough, particularly in terms of sensitivity, as many factors can affect the growth of pathogens during the culture process. Second, SCM only has qualitative results (presence or non-presence of pathogens) and thus cannot provide quantitative recommendations for treatment, such as dosage. Third, we often cannot obtain timely results from SCM to recommend suitable treatments. These limitations can lead to misuse or overuse of antibiotics and cause the miss of opportunities to treat patients. Therefore, there is a great need to develop new diagnostic tests for pathogens in clinical practice.

Kang *et al.* [1], for the first time, utilized quantitative real-time loop-mediated isothermal amplification (qrt-LAMP) [2] to test for pathogens in the sputum samples. The study was led by the Department

[a]*Department of Surgery, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[b]*Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*
[c]*Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[d]*Department of Respiratory and Critical Care Medicine, Peking University People's Hospital, Beijing 100044, China*
**Correspondence to: Lu Wang, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
†*E-mail: luwang@umich.edu*

of Respiratory and Critical Care Medicine, Peking University People's Hospital, Beijing, with the joint effort of 21 tertiary hospitals in 14 provinces of China. They successfully demonstrated that qrt-LAMP can be used to identify and quantify multiple common pathogens related to LRTIs in sputum samples of patients with great performance from lab tests. LAMP was originally used for infectious diseases caused by a single type of pathogen, such as tuberculosis, malaria, or sleeping sickness [3–5], where statistical analyses are straightforward. However, to our knowledge, there is little statistical literature on analyzing qrt-LAMP data for diseases such as LRTIs, which can be caused by a variety of pathogens. The major challenge lies in the differentiation of infection (pathogens rapidly grow, invade the host system, and cause the disease) from colonization (pathogens inhabit the host system but do not cause problems). The ability to identify such disease-causing bacterial pathogens will have a great impact on clinical diagnosis and treatment.

Towards this goal, we first demonstrate that qrt-LAMP can indeed amplify and quantify the eight most common bacterial pathogens for LRTIs and evaluate its performance. Traditionally, this depends on a 'gold standard' test or a 'resolver' test in discrepant analysis [6]. However, in our case, such tests are not available from the experimental design. As Kang *et al.* [1] suggested that qrt-LAMP has a almost perfect specificity, we employ a zero-inflated mixture model [7–10] to identify false-negative results among all negative results. We then estimate the prevalence of bacterial pathogen carriers in the population as well as the sensitivity of qrt-LAMP test. A separate analysis was conducted on the SCM data in comparison with qrt-LAMP results. In addition, we investigate the effects of important baseline covariates on the prevalence of carriers and on quantities of pathogens for carriers through a zero-inflated Tobit model [11].

Furthermore, we provide an approach to utilizing the measurements from qrt-LAMP to identify the disease-causing pathogens. This is critical for clinical recommendation of choices and dosages of antibiotic treatments. LRTIs are generally associated with a large number of pathogens, and these pathogens can inhabit peacefully within the host system. Therefore, for any single pathogen, a positive result from qrt-LAMP only implies its presence in the host system but does not necessarily mean that this pathogen causes the disease. Only when one or a combination of pathogens rapidly grow and cause infections are they considered disease-causing pathogens. As one can never be sure whether a pathogen causes the disease by merely looking at its qrt-LAMP result, the traditional approaches such as logistic regression or supervised learning cannot be applied. As one particular pathogen grows eventually uncontrollably and causes LRTIs, its initial abundance (or absolute quantity) must be an important factor. Meanwhile, its relative abundance is also crucial because pathogens live together in the hosting environment and compete for limited nutrients available. The growth of one pathogen would consume more nutrients and in turn suppress the growth of others. We combine these two aspects, both the absolute and relative abundance, by novelly resembling the techniques used in parallel tempering from Markov chain Monte Carlo (MCMC) literature [12], and construct G-scores to better represent the growth information compared with the original qrt-LAMP measurements. Change-point detection methods clearly reveal two transition points for these G-scores, which is consistent with the biological fact that there are two transition points for bacterial growth. We further use the piecewise-linear regression to identify the transition points and thus provide candidate criteria for the diagnosis and treatment of LRTIs, which can also serve as guidance to design future phase I or phase II clinical trials.

The remainder of the article is organized as follows: Section 2 gives a brief description of data. Detailed notation, the proposed zero-inflated mixture models, and implementations are presented in Section 3. Section 4 presents how to form the G-scores, determine their transition points, and identify disease-causing pathogens for individuals. Detailed results are shown in Section 5. Finally, we conclude with a brief discussion in Section 6. Technical details are left in the Supporting information.

2. The motivating data

The data were collected from 21 tertiary hospitals in 14 provinces of China. In total, 1533 subjects diagnosed with LRTIs were enrolled in the study. Their personal information, such as age, gender, scores on severity of diseases, and medical prescription history, was recorded. Sputum samples were collected for each patient. Both SCM and qrt-LAMP were performed on the sputum samples to test for eight common bacterial pathogens for LRTIs. SCM was performed three times independently on separate sputum samples. As a qualitative diagnostic test, SCM only reports positive or negative for each

experiment. The number of positive results was recorded. qrt-LAMP assay yields continuous titer values, which represent the concentration of bacterial pathogens. To achieve a high specificity, titer values measured from qrt-LAMP are left censored at 10^3 (values below 10^3 were recorded as 0 [1]).

For each pathogen, the majority (over 95%) of SCM results report negative, and very few patients have consistent SCM positive results from three experiments, indicating great variability in the bacterial growth in the culture medium. qrt-LAMP, on the contrary, reports more positive results, which range from 6.6% to 15.9% for different pathogens. There can be multiple pathogens present in one patient. Those who carry two or more pathogens account for approximately 30% of the total subjects.

3. Statistical inference on pathogen prevalence

3.1. Notation

We denote sample size as N and number of bacterial pathogens as C ($N = 1533$ and $C = 8$ in our data). For each patient i ($i = 1 \ldots N$) and each bacterial pathogen j ($j = 1 \ldots C$), we use X_{ij} to denote base-10 logarithm of the titer value from qrt-LAMP. As titer values are left censored at 10^3, X_{ij} are left censored at 3. In case of censoring, we set X_{ij} equal to 0. Let Y_{ij} denote the number of positive results from three independent culture experiments for pathogen j of patient i. Among all baseline covariates, we are particularly interested in age, whether having chronic obstructive pulmonary disease (COPD), and whether having bronchiectasis, all of which are clinically known as common risk factors for LRTIs. Hence, we denote $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4}, Z_{i5})^T$, where $Z_{i1} = 1$ if the i-th patient has COPD, and 0 otherwise; $Z_{i2} = 1$ if the i-th patient has bronchiectasis, and 0 otherwise; Z_{i3}, Z_{i4}, or Z_{i5} is 1, if the i-th patient belongs to the age group under 15 years old, 15–69 years old, or 70 years old and over, correspondingly, and is 0 otherwise. The following analyses were conducted separately on the SCM data and the qrt-LAMP data.

3.2. Estimation of prevalence

Our first goal is to estimate the prevalence of carriers for each bacterial pathogen, which represents the percentages of population who carry pathogens. Traditionally, prevalence was estimated from the SCM data by the percentage of positive results, where 'positive' means at least one of three culture results is positive. This implicitly assumes that at least one of culture experiments would report positive if the pathogen is indeed present. However, this assumption is far from the reality. As many factors, such as medium, temperature, and oxygen, influence the growth of bacterial pathogens, it is very likely that we fail to see any positive result among all three culture results, even when bacterial pathogens are indeed present in the sputum samples. These false-negative results lead to inaccurate estimates of the prevalence using the SCM data. False-negative results also exist for qrt-LAMP data because Kang *et al.* [1] choose a high cut-off value, 10^3, to achieve a very high specificity. In this article, we allow the possibility of false-negative results and assume that false-positive rates are zero for both SCM and qrt-LAMP data. As it is impossible to collect sputum samples from healthy people and we do not have a gold standard test by design, we develop novel statistical models to differentiate false-negative results for pathogen carriers undetected by either tests, from true-negative results for non-carriers. We focus on one bacterium each time and investigate how they are distributed marginally. Hence, we suppress the subscript j in the following.

We consider the following zero-inflated mixture models. Specifically, we assume that the measurements from either qrt-LAMP or SCM follow a mixture distribution

$$(1 - \lambda)\delta_0(\cdot) + \lambda f(\cdot; \boldsymbol{\theta}), \tag{1}$$

where λ is the prevalence of pathogen carriers, $\delta_0(\cdot)$ is the degenerated probability distribution at point zero, $f(\cdot; \boldsymbol{\theta})$ is the probability density function or the probability mass function for the measurement of interest among pathogen carriers, and $f(\cdot; \boldsymbol{\theta})$ is unknown subject to a finite-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathcal{R}^d$. Next, we will propose separate choices of f to model qrt-LAMP and SCM data, respectively.

3.2.1. Modeling the qrt-LAMP data. For pathogen carriers, as we know by the design that X is left censored at 3, we assume that it follows a normal distribution with mean μ and variance σ^2 censored at

3. Specifically, f is modeled as the Tobit model [13, 14]:

$$\begin{aligned} f(x;\mu,\sigma^2) &= \Phi\left(\frac{3-\mu}{\sigma}\right)\delta_0(x) + \left\{1-\Phi\left(\frac{3-\mu}{\sigma}\right)\right\}\frac{\varphi\left(\frac{x-\mu}{\sigma}\right)/\sigma I(x\geq 3)}{1-\Phi\left(\frac{3-\mu}{\sigma}\right)} \\ &= \Phi\left(\frac{3-\mu}{\sigma}\right)\delta_0(x) + \varphi\left(\frac{x-\mu}{\sigma}\right)/\sigma I(x\geq 3), \end{aligned} \tag{2}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal random variable. Note that the aforementioned f is itself a zero-inflated mixture model reflecting that the measurement is left censored at 3. The degenerated part of f is the source of false negatives, which we can extrapolate from the positive results that are observed.

3.2.2. Modeling the SCM data. Culture results from SCM are discrete random variables taking values 0, 1, 2, and 3, which represent the number of positive results seen in three independent experiments. We assume those are three independent Bernoulli experiments and hence model it with a binomial distribution, Binom(3, p).

$$f(y;p) = \binom{3}{y} p^y (1-p)^{3-y}. \tag{3}$$

3.3. Adjustment for baseline covariates

Given the quantitative nature of qrt-LAMP, we want to further assess the association between qrt-LAMP and baseline covariates. We extend the Tobit model in (2) and propose a zero-inflated Tobit model, which adjust for the effects of baseline covariates for parameters in both models (1) and (2). Given $\boldsymbol{Z} = (Z_1, Z_2, Z_3, Z_4, Z_5)^T$, we assume

$$X \mid \boldsymbol{Z} \sim \{1-\lambda(\boldsymbol{Z})\}\delta_0(x) + \lambda(\boldsymbol{Z}) f(x \mid \boldsymbol{Z}; \boldsymbol{\theta}(\boldsymbol{Z})) \tag{4}$$

and

$$\begin{aligned} f(x \mid \boldsymbol{Z}; \boldsymbol{\theta}(\boldsymbol{Z})) &= \Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)\delta_0(x) + \left\{1-\Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)\right\}\frac{\varphi\left(\frac{x-\mu(\boldsymbol{Z})}{\sigma}\right)/\sigma I(x\geq 3)}{1-\Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)} \\ &= \Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)\delta_0(x) + \varphi\left(\frac{x-\mu(\boldsymbol{Z})}{\sigma}\right)/\sigma I(x\geq 3), \end{aligned} \tag{5}$$

where σ is common standard deviation, $\delta_0(x)$ is the degenerated probability density at point zero, as defined in Section 3.2, $\lambda(\boldsymbol{Z})$ represents the prevalence of pathogen carriers in the stratified group with covariate $\boldsymbol{Z}$, logit$\{\lambda(\boldsymbol{Z})\}$ is assumed to be linear in $\boldsymbol{Z}$, that is,

$$\lambda(\boldsymbol{Z}) = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{Z})}{1+\exp(\boldsymbol{\beta}^T\boldsymbol{Z})}, \tag{6}$$

$\mu(\boldsymbol{Z})$ represents the mean quantity of the pathogen for carriers in the stratified group with covariate $\boldsymbol{Z}$, and we assume

$$\mu(\boldsymbol{Z}) = \boldsymbol{\gamma}^T\boldsymbol{Z}. \tag{7}$$

Equations (6) and (7) assume additive effects of the status of COPD, bronchiectasis, and age on both logit of prevalence of pathogen carriers and the mean quantity of the pathogen for carriers.

3.4. Statistical inference

We focus on demonstrating how to make inference for the model in Section 3.3. The model without adjustment of baseline covariates can be similarly implemented. The joint likelihood function from the model formed by Equations (4)–(7) is

$$L(\boldsymbol{\beta},\boldsymbol{\gamma},\sigma^2;\boldsymbol{X},\boldsymbol{Z})=\prod_{i=1}^{N}\left[\left\{1-\lambda(\boldsymbol{Z}_i)+\lambda(\boldsymbol{Z}_i)\Phi\left(\frac{3-\mu(\boldsymbol{Z}_i)}{\sigma}\right)\right\}^{I(X_i=0)}\left\{\frac{\lambda(\boldsymbol{Z}_i)}{\sigma}\varphi\left(\frac{X_i-\mu(\boldsymbol{Z}_i)}{\sigma}\right)\right\}^{I(X_i>0)}\right]. \tag{8}$$

For notation convenience, we shall denote $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\boldsymbol{\gamma}^T,\sigma^2)^T$. Following the classical likelihood theory, the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is asymptotically normally distributed:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}-\boldsymbol{\theta}_0)\xrightarrow{\mathcal{D}}N\left(0,\mathcal{I}_{\boldsymbol{\theta}_0}^{-1}\right),$$

where $\boldsymbol{\theta}_0$ is the true parameters corresponding to data generating process and $\mathcal{I}_{\boldsymbol{\theta}_0}$ is the Fisher information matrix. $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ and an estimate of $\mathcal{I}_{\boldsymbol{\theta}_0}$ can be used to construct confidence intervals of $\boldsymbol{\theta}_0$.

However, obtaining MLE can be difficult, complicated by the fact that the density function is a mixture distribution and hence the likelihood function could have multiple local maxima. Data augmentation and MCMC algorithm can be used to facilitate the computation. For a given prior distribution $\pi(\boldsymbol{\theta})$, for instance $\pi(\boldsymbol{\theta})\propto 1/\sigma^2$, the posterior mean has the same asymptotic property as the MLE because of Bernstein–von Mises theorem [15, Chapter 10]. To sample from the posterior distribution, it can be achieved by introducing vectors of latent variables, $\boldsymbol{I}=(I_1,I_2,\ldots,I_N)$ and $\boldsymbol{X}^*=\left(X_1^*,X_2^*,\ldots,X_N^*\right)$, such that

$$P(I_i=1\mid\boldsymbol{Z}_i)=\lambda(\boldsymbol{Z}_i)$$
$$P(I_i=0\mid\boldsymbol{Z}_i)=1-\lambda(\boldsymbol{Z}_i)$$

and

$$X_i^*\mid I_i,\boldsymbol{Z}_i\sim(1-\lambda(\boldsymbol{Z}_i))\delta_0\left(X_i^*\right)+\lambda(\boldsymbol{Z}_i)\frac{1}{\sigma}\varphi\left(\frac{X_i^*-\mu(\boldsymbol{Z}_i)}{\sigma}\right).$$

I_i indicates whether patient i is bacterial pathogen carrier, and $\boldsymbol{X}^*$ could be thought of as the ideal qrt-LAMP measurements if they were not censored at 3. Let $\boldsymbol{X}=\boldsymbol{X}^*\times I(\boldsymbol{X}^*\geqslant 3)$, and then $(\boldsymbol{X},\boldsymbol{Z})$ would be the observed data. We can implement a Gibbs sampler by sequentially sampling from the conditional distribution $P\{(\boldsymbol{I},\boldsymbol{X}^*)\mid\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z}\}$, $P\{(\boldsymbol{\gamma},\sigma^2)\mid\boldsymbol{\beta},\boldsymbol{I},\boldsymbol{X}^*,\boldsymbol{Z}\}$, and $P(\boldsymbol{\beta}\mid\boldsymbol{\gamma},\sigma^2,\boldsymbol{I},\boldsymbol{X}^*,\boldsymbol{Z})$, which can be similarly derived as in [16]. We present details of MCMC in the Supporting information.

4. Identifying disease-causing bacterial pathogens

Although the prevalence of pathogen carriers estimated in Section 3 has its epidemiological significance, we want to reiterate that, for disease such as LRTIs, the carrying of certain pathogens does not mean that these pathogens cause disease. Unlike diseases caused by a single type of pathogen, over 200 kinds of bacterial pathogens can cause LRTIs, and these bacterial pathogens can inhabit the host system without causing any problems. Therefore, a positive outcome from qrt-LAMP for one pathogen does not necessarily mean that it is the cause of the disease. Only one or a few of the pathogens in the host system would proliferate rapidly, take up most of the nutrients from the host, and cause infections. The eight kinds of bacterial pathogens, measured by our designed qrt-LAMP, are among the most common disease-causing pathogens for LRTIs. In this section, we develop a quantitative standard based on titer values from qrt-LAMP to identify disease-causing pathogens and hence provide guidance and assistance for the diagnosis and treatment among patients with LRTIs.

Biological bacterial growth in batch culture typically consists of three phases—lag phase, log phase, and stationary phase. In the lag phase, bacteria are maturing but are not yet able to rapidly grow. Pathogens in this phase should not be considered as disease causing, as they are not invasive to the host system. The log phase is a period of cell doubling when bacteria start to grow exponentially. In

this phase, the bacteria grow rapidly, and the host immune system can no longer protect the organism from infection. We therefore regard the pathogens at log phase as disease-causing pathogens. It is this transition point from lag phase to log phase that is most interesting to us, as it differentiates disease-causing pathogens from non-disease-causing ones. Finally, when the growth of pathogens is slowed down because of the limited amount of nutrients, it reaches the third phase—stationary phase. The number of pathogens reaches the limit in this phase. Although pathogens in log and stationary phases are both disease-causing pathogens, the second phase transition point would be more helpful to provide guidance for the dosage use of antibiotics and thus minimize antibiotics overuse.

The goal in this section is to construct G-scores, which novelly incorporate both absolute and relative levels of a pathogen in the symbiosis and resemble the growth process described earlier. Details are presented in the next two sections. In Section 4.1, we describe the development and definition of G-scores, and in Section 4.2, we use change-point detection methods to show that G-scores similarly have two transition points. The G-scores that we construct need to reflect the following two principles: (i) the larger the quantity of one pathogen is, the more likely it further grows, and eventually out of control; and (ii) the higher the relative ranking of one pathogen is, the more competitive it is to outgrow others. In both cases, the pathogen is more likely to be disease causing and hence should correspond to larger G-scores.

4.1. Defining G-scores that identify disease-causing pathogens

Recall that X^*_{ij}, for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, C$, defined in Section 3, are the augmented data as if qrt-LAMP were not left censored. These latent titer values reflect the 'absolute' level of pathogens. Larger quantity of X^* means that the pathogen might be at the log or stationary phase, or more likely to transit to these two phases. However, pathogens live together in the host system and compete for limited nutrients available. The 'competitiveness' is also important for them to take up more nutrients and outgrow others. Thus, both 'absolute' levels and 'relative' levels play important roles for a pathogen to cause the disease. Therefore, the desired scores should incorporate both pieces of information in identifying the disease-causing pathogens.

Towards this goal, we first normalize X^*_{ij} to facilitate further comparisons among different pathogens in order to determine their relative levels. As different pathogens have different concentrations within the community, we transform X^*_{ij} into U_{ij}, which we call U-score, according to their distributions among pathogen carriers. Specifically,

$$U_{ij} = \Phi\left\{\frac{E(X^*_{ij} \mid X_{ij}, \boldsymbol{Z}_i) - \hat{\mu}_j}{\hat{\sigma}_j}\right\},$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the estimated parameters from Equations (7) and (8), and $E\left(X^*_{ij} \mid X_{ij}, \boldsymbol{Z}_i\right)$ is the conditional expectation. The U-scores are monotone transformation of X_{ij}, which still reflect information of absolute quantities of pathogen titer values from qrt-LAMP. Meanwhile, U-scores justify comparisons among different pathogens from the same subject, as we do not compare the abundance of pathogens directly but through their standardized values, U-scores.

Next, we want to incorporate the 'relative' information as well. We achieve this by resembling the techniques used in parallel tempering from MCMC literature and 're-weighting' $U_i = (U_{i1}, U_{i2}, \ldots, U_{iC})$ according to their relative levels. Notice that parallel tempering is a simulation method to improve the efficiency of MCMC by 'enhancing' the global mode [12]. We utilize the same transformation on U-scores to combine both 'absolute' and 'relative' information, which leads to an increased score for the pathogen with a larger U-score. Specifically, given a positive constant κ, we define

$$G^*_{ij}(\kappa) = \frac{e^{\kappa U_{ij}} - 1}{\sum_{k=1}^{C}(e^{\kappa U_{ik}} - 1)} \times \sum_{k=1}^{C} U_{ik}$$

for subject i and pathogen j. For any κ, $\sum_{j=1}^{C} G^*_{ij}(\kappa) = \sum_{j=1}^{C} U_{ij}$, and these G^*-scores try to re-weight U-scores to incorporate the relative level information of pathogens. κ controls how much adjustment we want to make, and when $\kappa \to 0$, $G^*_{ij}(\kappa) \to U_{ij}$, thus going back to the original U-scores.

Clearly, G^*-score does reflect the two principles that we mentioned earlier. G^*_{ij} is increasing with U_{ij}, and the right derivative of G^*_{ij} at 0 is

$$\lim_{\kappa\to 0+}\frac{dG^*_{ij}(\kappa)}{d\kappa}=\frac{U_{ij}}{2\sum_{k=1}^{C}U_{ik}}\left(U_{ij}-\overline{U}_i\right),$$

where $\overline{U}_i=\left(\sum_{k=1}^{C}U_{ik}^2\right)/\left(\sum_{k=1}^{C}U_{ik}\right)$. Note $\overline{U}_i=0\sum_{j=1}^{C}\{U_{ij}\left(\sum_{k=1}^{C}U_{ik}\right)^{-1}\}U_{ij}$ is a weighted average of U_{ij}. For $U_{ij}>\overline{U}_i$, its G^*-score increases with κ in the neighborhood of zero, and for $U_{ij}<\overline{U}_i$, its G^*-score decreases with κ in the neighborhood of zero. This 'rich-get-richer' phenomenon mimics the growth of pathogens in the host when they consume more nutrients and then in the meanwhile suppress the growth of others.

Furthermore, for fixed i, the aforementioned G^*-scores preserve the ordering of the corresponding U scores. To be specific, $G^*_{ia}>G^*_{ib}$ if and only if $U_{ia}>U_{ib}$, for any $a\neq b\in\{1,2,\dots C\}$. However, for fixed j, the ordering can change. That is, G^*_{aj} can be smaller than G^*_{bj} even if $U_{aj}>U_{bj}$, for $a\neq b\in\{1,2,\dots N\}$. The reason is that pathogen j can be suppressed in patient a yet proliferate in patient b depending on their relative levels in the patients, which could reverse their orders of comparisons across different subjects. This again shows that G^*-scores successfully combine two parts of information from qrt-LAMP together: the absolute quantities of pathogens and their relative level/ranking in the pathogen community.

Let $j_i=\arg\max_j\{U_{ij}\}$, and then $U_{ij_i}>U_{ij}$ for all $j\neq j_i$. One can easily show that $G^*_{ij_i}(\kappa)$ is a strictly increasing function of κ with the limit $\sum_{j=1}^{C}U_{ij}$ as $\kappa\to\infty$, and thus $G^*_{ij}(\kappa)\to 0$ as $\kappa\to\infty$ for all other $j\neq j_i$. Therefore, when κ is large enough, G^* represents only the dominant pathogen that has the largest U-score. However, when there are a few pathogens all with large U-scores, one probably should treat them together as the cause of disease instead of any individual one. Hence, we restrict that G^* can only go up to 1 and define G-scores as follows:

$$G_{ij}(\kappa)=\begin{cases}G^*_{ij}(\kappa) & \text{if } \kappa\leqslant\kappa_i^{\max}\\ G^*_{ij}\left(\kappa_i^{\max}\right) & \text{if } \kappa>\kappa_i^{\max}\end{cases},$$

where

$$\kappa_i^{\max}=\sup\left\{\kappa:\max_j G^*_{ij}(\kappa)\leqslant 1\right\}.$$

Note that $\kappa_i^{\max}=\infty$, if $\sum_{j=1}^{C}U_{ij}<1$. These G-scores retain the properties of G^*-scores but also allow multiple positive values even for larger κ. G-scores form the basis of our method to identify the disease-causing pathogens in Section 4.2.

4.1.1. Choices of κ. Different κ's lead to different transformations. The choice of κ should be up to the physicians as they need to balance 'absolute' and 'relative' information from qrt-LAMP based upon their experience and knowledge in deciding G-scores, which would be closely related to the bacterial growth. However, when no such expert opinions are available, we recommend to choose κ in the following way. For subject i, if we choose κ such that $\kappa>\kappa_i^{\max}$, then one of the pathogens has its G-score reach 1 and thus multiple pathogens among these eight might contribute to the disease development. Denote $M(\kappa)=\frac{1}{N}\sum_{i=1}^{N}I\left(\kappa\geqslant\kappa_i^{\max}\right)$, which represents the percentage of subjects whose largest G-scores have reached 1. We recommend to use κ such that $M(\kappa)$ equals to the percentage of subjects in the population whose LRTIs are caused by multiple pathogens among these eight pathogens of interest, which is around 10%.

4.2. Determining phase transition points through change-point detection methods

As stated earlier, we are interested in identifying phase-transition points of the bacterial growth process, in particular the transition point from lag phase to log phase, which differentiates between disease-causing pathogens and non-disease-causing pathogens. These phase-transition points have clear clinical significance to provide criteria for diagnosis and treatment. In Section 4.1, we defined G-scores to incorporate important information about bacterial growth. We will now show that the proposed

G-scores have the same attributes that they similarly have two transition points. Consider the following regression model:

$$G(\kappa) = \boldsymbol{\beta}(X^*)^T (1, X^*)^T + \epsilon,$$

where $\boldsymbol{\beta}(X^*) = \{\beta_1(X^*), \beta_2(X^*)\}^T$ and ϵ is the normal random error. The null hypothesis is

$$H_0 : \boldsymbol{\beta}(X^*) = \boldsymbol{\beta}.$$

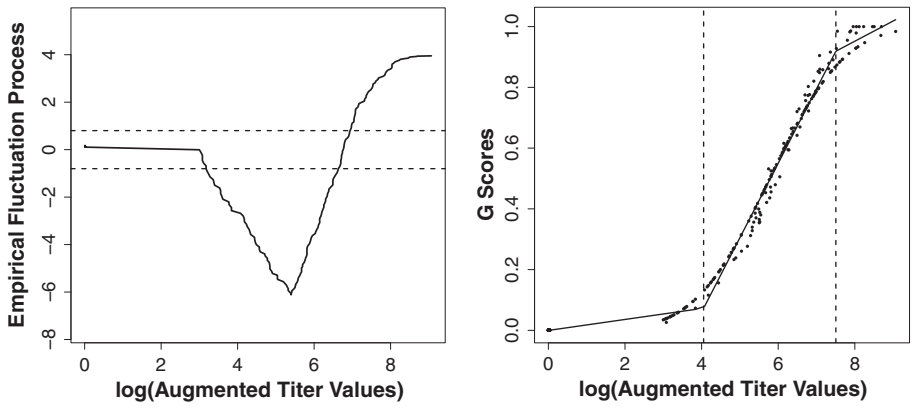
That is, the regression coefficients do not change with X^*, meaning that there are no structural changes. The alternative hypothesis is that $\boldsymbol{\beta}(X^*)$ has structural changes at some points, which correspond to the phase transitions for bacteria growth described earlier. We propose to utilize the process of the partial moving sums (MOSUM process) of residuals from the ordinary least square estimates [17–20] to detect change points for linear regression models, which is visually illustrative. Let $\hat{u}_i$ and $\hat{\sigma}^2$ denote the estimated residuals and variance from ordinary least estimates under the null hypothesis. Let $C_h(X^*)$ denote the interval centered at X^*, which includes $h\%$ of all data, where h is a tuning parameter. Consider the moving sums defined as follows:

$$M(X^*) = \frac{1}{\hat{\sigma}\sqrt{n-2}} \sum_i \hat{u}_i I\left(X_i^* \in C_h(X^*)\right).$$

Under the null hypothesis, this process converges to the increments of a Brownian bridge. Fluctuations beyond its normal variations suggest the existence of change points. In Figure 1(a), it shows a clear evidence that there are two change points, which rejects the null hypothesis and suggests that $\boldsymbol{G}$ scores have three phases as well as the bacterial growth of pathogens. To further identify those two change points, we fit a piecewise linear regression with two change points as follows:

$$G(\lambda) = \beta_0 + \beta_1 X^* + \beta_2 (X^* - c_1)^+ + \beta_3 (X^* - c_2)^+ + \epsilon, \tag{9}$$

where c_1 and c_2 ($c_1 < c_2$) are the two change points, which can be estimated by minimizing L^2 loss function. We employ the iterative algorithm proposed by Muggeo [21] to fit model (9). Detailed results are presented in the next section. We use the statistical software R (http://www.r-project.org) to perform the analysis. The strucchange package is used to draw the MOSUM process, and the segmented package is used to identify change points. Both packages are available at http://cran.r-project.org.



(a) The MOSUM Process (b) Piecewise Linear Regression

Figure 1. Change-point detection. (a) The solid line shows the moving sum of residuals from the ordinary least square estimation. The two dashed lines are 95% level of fluctuations under the null hypothesis when there is no structural change. The MOSUM process going across the fluctuation lines suggests existence of structural changes. (b) The solid line is the fitted piecewise regression line for model (9), and the two dashed lines correspond to the estimated $\hat{c}_1$ and $\hat{c}_2$.

5. Results applied to LRTI data

5.1. Quality of extrapolation

As there is no gold standard reference test, we are unable to identify false-positive or false-negative results by merely looking at the data. Instead, with the knowledge from lab performance [1], we assume that there is no false-positive results, and we parametrically model the distribution of quantities of pathogens for carriers. Therefore, we are able to extrapolate information on the number of false-negative results from positive results that we observed. For this reason, the accuracy of our statistical results depends on the quality of extrapolation when predicting the amount of false-negative observations. We conduct model checking for both (2) and (3).

For qrt-LAMP data, which is continuous, it is straightforward to consider drawing Q-Q plots. Using the parameters as estimated from the MCMC algorithm in Section 3, we calculate the theoretical quantiles for model (2) at levels $5\%, 10\%, \ldots, 95\%$ and compared them with the empirical quantiles calculated directly from the qrt-LAMP measurements. Q-Q plots in the Supporting information clearly show that theoretical quantiles and empirical quantiles closely follow the 45° diagonal line $y = x$, which reflects that the mixture model fits data well. In addition, the Kolmogorov–Smirnov test, which compares theoretical distributions with empirical distributions, fails to reject the null hypotheses for all pathogens, confirming the results shown in Q-Q plots. Therefore, we are quite confident that the predicted numbers of false-negative results are accurate to a certain extent for all eight pathogens.

For SCM data, we similarly calculate the theoretical probability mass function using the parameters as the estimates in Section 3. We then test whether the observed data come from the multinomial distribution with the theoretical probabilities we calculate. This is usually performed through Pearson's chi-square test. However, in our case, the probability that one can see two or more positive results is quite small, and normal approximation needed in chi-square test does not hold. Therefore, we conduct the exact test instead. That is, for the observed data $\boldsymbol{X} = (X_0, X_1, X_2, X_3)$ and parameters for multinomial distribution $\boldsymbol{p} = (p_0, p_1, p_2, p_3)$, we calculate

$$p\text{-value} = \sum_{\boldsymbol{x}:\Pr(\boldsymbol{x})\leq\Pr(\boldsymbol{X})} \Pr(\boldsymbol{x}),$$

where $\Pr(\boldsymbol{x}) = N!\prod_{i=0}^{3} p_i^{x_i}/x_i!$. Unlike qrt-LAMP, four of eight null hypotheses are rejected, meaning that zero-inflated binomial model does not fit the SCM data well. Possible explanations could be that three culture experiments conducted at different time might be subject to different environment factors, and later experiments might have less probability to produce positive results due to deaths of pathogens during storing. These would lead to over-dispersion, and thus, the zero-inflated binomial model fails to fit the data. However, we cannot go one step further to try modeling such as beta-binomial to accommodate over-dispersion, as the model will become saturated, and there is no way to conduct a goodness-of-fit test. We hence stay with the binomial model but remind readers that they should view the SCM results with skepticism if they fail the goodness-of-fit test. Both p-values of the exact multinomial test and the Kolmogorov–Smirnov test are presented in the Supporting information.

We have demonstrated that the new method, qrt-LAMP, provides more credible results than SCM. To quantify its performance, one can calculate its sensitivities and false-negative rates as

$$\text{Sensitivity} = \Phi\left(\frac{\mu - 3}{\sigma}\right)$$

and

$$\text{False-negative rate} = \frac{\lambda\Phi\left(\frac{3-\mu}{\sigma}\right)}{1 - \lambda + \lambda\Phi\left(\frac{3-\mu}{\sigma}\right)}.$$

In our data, sensitivities range from 82% to 96%, and false-negative rates range from 0.6% to 3.6% for different pathogens. Details are presented in the Supporting information.

5.2. Results from the mixture model

Parameter λ in models (2) and (3) is the prevalence of pathogen carriers. Table I presents estimates of prevalence from SCM and qrt-LAMP. One can see that qrt-LAMP mostly gives larger estimates

Table I. Estimates of prevalence of pathogen carriers.

(a) SCM			(b) qrt-LAMP marginal model			(c) qrt-LAMP regression model		
Pathogen	Est.	CI	Pathogen	Est.	CI	Pathogen	Est.	CI
Spn	0.148	(0.044, 0.504)	Spn	0.130	(0.112, 0.147)	Spn	0.128	(0.111, 0.165)
Sau	0.056	(0.025, 0.123)	Sau	0.101	(0.084, 0.118)	Sau	0.099	(0.077, 0.126)
Eco	0.031	(0.019, 0.049)	Eco	0.082	(0.063, 0.101)	Eco	0.079	(0.057, 0.109)
Kpn	0.042	(0.029, 0.061)	Kpn	0.189	(0.162, 0.215)	Kpn	0.188	(0.156, 0.226)
Pae	0.090	(0.070, 0.116)	Pae	0.139	(0.120, 0.159)	Pae	0.137	(0.110, 0.170)
Aba	0.063	(0.042, 0.096)	Aba	0.157	(0.136, 0.178)	Aba	0.156	(0.127, 0.191)
Sma			Sma	0.179	(0.153, 0.205)	Sma	0.174	(0.142, 0.213)
Hin	0.053	(0.014, 0.198)	Hin	0.136	(0.117, 0.154)	Hin	0.138	(0.111, 0.165)

(a) and (b) show the estimated λ in Equation (1). (c) shows the estimated $E[\lambda(Z)]$ in Equation (4). 95% confidence intervals (CI) are given in parenthesis.

on the prevalence compared with SCM. The reason is that bacterial pathogens might die during the storing process, so they are not able to grow when SCM are conducted. In this case, we observe negative results even when pathogens are indeed present in the sputum samples. In contrast, qrt-LAMP amplifies and quantifies DNA sequences and thus manage to detect pathogens regardless whether they are dead or alive during the qrt-LAMP process. This explains why SCM underestimates the prevalence compared with qrt-LAMP. In addition, qrt-LAMP also shows better precision than SCM, as SCM sometimes gives abnormally broad confidence intervals, for example, in the results for *Streptococcus pneumoniae* (Spn), *Staphylococcus aureus* (Sau), and *Haemophilus influenzae* (Hin). Also notice that only four pathogens in the bold font pass the goodness-of-fit test for SCM. Therefore, the SCM results for Sau, Escherichia coli (Eco), and *Pseudomonas aeruginosa* (Pae) are suspicious. For pathogen *Stenotrophomonas maltophilia* (Sma), there are no patients who have two or more positive results from SCM; therefore, we are not able to extrapolate the number of false-negative results. That is why we leave it blank for that particular entry in the table.

In model (4) where we adjust for baseline covariates, the regression coefficients have their clinical interpretations. $\exp(\beta_1)$ and $\exp(\beta_2)$ are the odds ratios of carrying pathogens between patients with or without COPD and with or without bronchiectasis, respectively. γ_1 and γ_2 are differences of quantities of pathogens for carriers between patients with or without COPD and with or without bronchiectasis, respectively. Table II shows these estimates and their 95% confidence intervals. The most statistically significant finding is that Pae is very different between groups with versus without bronchiectasis in terms of both odds ratio of carrying Pae and differences of quantities of pathogens for Pae-carriers. The odds ratio is almost fivefold, and Cohen's d for the standardized effect size is 3.98 (both p-values are less than 0.001). The results remain statistically significant even after multiple comparison adjustments. Other findings are that Spn has a higher prevalence in the group of patients with COPD compared with the group of patients without COPD, and *Klebsiella pneumoniae* (Kpn) has a lower concentration among Kpn carriers with bronchiectasis compared with those without bronchiectasis.

To compare among three age groups, $\exp(\beta_i - \beta_j)$ and $\gamma_i - \gamma_j$ $(i, j = 3, 4, 5$ and $i > j)$ are odds ratios of carrying pathogens and differences of quantities of pathogens for carriers, respectively, which were presented in Table III. For most pathogens, age group 15–69 years old and age group 70 years old and over have similar odds ratios and differences, while statistically significant differences are present between age group under 15 years old and the other two age groups. Patients 15 years old and over have threefold to fourfold odds to carry *Acinetobacter baumannii* (Aba) and twofold to threefold odds of carrying Sma compared with patients under 15 years old. Older patients also have larger concentrations of Sau, Kpn, and Hin for pathogens carriers, as shown in Table III(b).

In addition, from the regression model (4), we can re-estimate prevalence, sensitivity, and false-negative rate by

$$\text{Prevalence} = E\{\lambda(\boldsymbol{Z})\}$$

$$\text{Sensitivity} = E\left\{\Phi\left(\frac{\mu(\boldsymbol{Z}) - 3}{\sigma}\right)\right\}$$

Table II. Odds ratio of carrying pathogens and difference of quantities of pathogens for carriers between patients with or without COPD (the first column) and with or without bronchiectasis (the second column).

(a) Odds ratios of carrying pathogens					(b) Difference of quantities of pathogens for carriers				
	$\exp(\beta_1)$		$\exp(\beta_2)$		Pathogen	γ_1		γ_2	
Pathogen	Est.	CI	Est.	CI		Est.	CI	Est.	CI
Spn	1.524	(1.068, 2.173)*	0.721	(0.383, 1.360)	Spn	0.025	(−0.567, 0.617)	−0.409	(−1.413, 0.594)
Sau	1.034	(0.683, 1.564)	0.712	(0.349, 1.451)	Sau	−0.306	(−1.100, 0.488)	−0.155	(−1.503, 1.193)
Eco	0.901	(0.540, 1.504)	0.908	(0.367, 2.245)	Eco	−0.553	(−1.338, 0.232)	−1.063	(−2.364, 0.239)
Kpn	0.846	(0.606, 1.180)	1.089	(0.619, 1.916)	Kpn	−0.419	(−0.889, 0.051)#	−1.033	(−1.797, −0.270)**
Pae	1.312	(0.923, 1.865)	4.900	(3.294, 7.288)***	Pae	−0.154	(−0.714, 0.407)	1.205	(0.6100, 1.799)***
Aba	1.262	(0.902, 1.766)	0.826	(0.469, 1.455)	Aba	−0.431	(−0.924, 0.063)#	−0.452	(−1.299, 0.396)
Sma	0.973	(0.686, 1.381)	0.681	(0.384, 1.206)	Sma	0.444	(−0.191, 1.080)	0.186	(−0.865, 1.236)
Hin	0.824	(0.574, 1.181)	1.378	(0.864, 2.200)	Hin	0.255	(−0.312, 0.822)	0.192	(−0.501, 0.885)

***Significant level less than 0.001.
**Significant level between 0.001 and 0.01.
*Significant level between 0.01 and 0.05.
#Significant level between 0.05 and 0.1.

Table III. Odds ratio of carrying pathogens and difference of quantities of pathogens for carriers between different age groups.

	(a) Odds ratios of carrying pathogens							(b) Difference of quantities of pathogens for carriers						
	$\exp(\beta_4-\beta_3)$		$\exp(\beta_5-\beta_3)$		$\exp(\beta_5-\beta_4)$			$\gamma_4-\gamma_3$		$\gamma_5-\gamma_3$		$\gamma_5-\gamma_4$	
Pathogen	Est.	CI	Est.	CI	Est.	CI	Pathogen	Est.	CI	Est.	CI	Est.	CI
Spn	0.627	(0.412, 0.954)*	0.676	(0.438, 1.043)#	1.078	(0.738, 1.574)	Spn	−1.010	(−1.664, −0.356)**	−0.899	(−1.617, −0.181)*	0.111	(−0.607, 0.829)
Sau	0.585	(0.347, 0.985)*	0.720	(0.427, 1.212)	1.230	(0.790, 1.915)	Sau	1.494	(0.513, 2.476)**	1.341	(0.37, 2.312)**	−0.153	(−1.124, 0.818)
Eco	1.555	(0.772, 3.132)	2.156	(1.044, 4.454)*	1.387	(0.843, 2.281)	Eco	0.376	(−0.803, 1.554)	0.114	(−1.048, 1.276)	−0.262	(−1.424, 0.9)
Kpn	0.802	(0.451, 1.426)	1.023	(0.576, 1.815)	1.275	(0.918, 1.770)	Kpn	2.132	(1.473, 2.791)***	2.211	(1.52, 2.903)***	0.08	(−0.612, 0.771)
Pae	1.056	(0.645, 1.729)	1.480	(0.889, 2.466)	1.402	(0.965, 2.038)#	Pae	−0.262	(−1.079, 0.555)	−0.463	(−1.265, 0.339)	−0.201	(−1.003, 0.6)
Aba	3.180	(1.788, 5.654)***	3.642	(2.037, 6.511)***	1.145	(0.810, 1.619)	Aba	−0.034	(−0.921, 0.853)	0.248	(−0.647, 1.143)	0.282	(−0.613, 1.177)
Sma	2.207	(1.310, 3.717)**	2.832	(1.669, 4.806)***	1.283	(0.896, 1.839)	Sma	−0.510	(−1.527, 0.506)	−0.215	(−1.166, 0.736)	0.296	(−0.655, 1.246)
Hin	1.418	(0.848, 2.372)	0.956	(0.547, 1.671)	0.674	(0.466, 0.976)*	Hin	1.161	(0.358, 1.963)**	1.703	(0.837, 2.568)***	0.542	(−0.323, 1.407)

***Significant level less than 0.001.
**Significant level between 0.001 and 0.01.
*Significant level between 0.01 and 0.05.
#Significant level between 0.05 and 0.1.

Table IV. Estimates and standard errors of phase transition (P-T) points for different pathogens and different κ's through model (9).

(a) Results when $M(\kappa) = 5\%$					(b) Results when $M(\kappa) = 10\%$					(c) Results when $M(\kappa) = 15\%$				
	1st P-T point		2nd P-T point			1st P-T point		2nd P-T point			1st P-T point		2nd P-T point	
Pathogen	Est.	SE	Est.	SE	Pathogen	Est.	SE	Est.	SE	Pathogen	Est.	SE	Est.	SE
Spn	4.07	0.016	7.48	0.024	Spn	4.41	0.024	7.26	0.034	Spn	4.54	0.033	7.12	0.046
Sau	3.83	0.017	7.66	0.018	Sau	4.19	0.030	7.71	0.032	Sau	4.33	0.042	7.63	0.042
Eco	3.27	0.018	6.46	0.024	Eco	3.35	0.028	6.38	0.041	Eco	3.37	0.038	6.24	0.044
Kpn	3.34	0.019	6.71	0.022	Kpn	3.53	0.028	6.47	0.032	Kpn	3.95	0.035	5.95	0.039
Pae	4.14	0.017	7.81	0.021	Pae	4.43	0.026	7.73	0.034	Pae	4.52	0.039	7.45	0.053
Aba	3.66	0.013	6.69	0.017	Aba	3.80	0.022	6.61	0.028	Aba	3.88	0.031	6.49	0.038
Sma	3.54	0.021	6.76	0.019	Sma	3.96	0.025	6.54	0.027	Sma	4.14	0.035	6.46	0.038
Hin	3.91	0.015	7.35	0.015	Hin	4.17	0.024	7.13	0.024	Hin	4.28	0.033	7.09	0.032

$$\text{False-negative rate} = E\left\{\frac{\lambda(\boldsymbol{Z})\Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)}{1-\lambda(\boldsymbol{Z})+\lambda(\boldsymbol{Z})\Phi\left(\frac{3-\mu(\boldsymbol{Z})}{\sigma}\right)}\right\}.$$

The results are shown in the corresponding tables along with estimates from the marginal model.

5.3. Phase transition points for identifying disease-causing pathogens

Figure 1(a) shows the moving sum process for G-scores for Spn, where the horizontal axis is logarithm (base 10) of the augmented titer values, the vertical axis is the MOSUM process, and κ is chosen to have $M(\kappa) = 10\%$. The solid line stands for the ordinary least square-based MOSUM process, and the dashed lines are the boundaries of fluctuation such that the probability that the MOSUM process is entirely contained within the boundaries is 0.95 under the null hypothesis.

Wherever the MOSUM process exceeds the upper or lower boundaries, the fluctuation of the empirical MOSUM process is improbably large, which suggests that there should be a change point at that location. For this reason, Figure 1(a) strongly indicates that G-scores have two change points for the chosen κ, which is consistent with the biological phase transitions. To identify these transition points, we fit a piecewise linear regression as shown in Equation (9). Figure 1(b) presents the results, where the horizontal axis is logarithm (base 10) of the augmented titer values, and the vertical axis is the G-score. The solid dots correspond to the observed data, and the solid line is the fitted piecewise linear curve. The vertical dashed lines are for fitted $\hat{c}_1$ and $\hat{c}_2$, which are the first and second transition points correspondingly.

For our recommended κ and two other possible values such that $M(\kappa)$ is 5% and 15%, respectively, the estimates of both phase transition points and their standard errors are shown in Table IV. One can see that the first phase transition point is generally increasing and the second phase transition point is generally decreasing as κ increases. This is because larger κ up-weights more on larger U scores and down-weights more on smaller U scores.

Figure 2 shows how varying κ affects estimates of the first and second phase transition points for Sma, where the horizontal axis is κ and the vertical axis is logarithm of the augmented titer values. The same trends are reflected in Figure 2. The estimated first phase transition point (lower solid line) is increasing, while the estimated second phase transition point (upper solid line) is decreasing as κ increases. Point-wise 95% confidence intervals are given by the dashed lines. The vertical dot-dashed lines correspond to κ such that $M(\kappa)$ is 5%, 10%, and 15%, respectively.

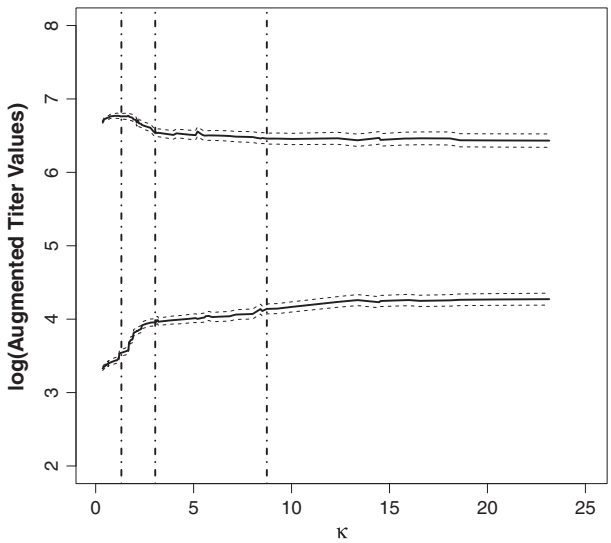


Figure 2. Phase transition points of pathogen Sma. The lower and upper solid lines are the estimated first and second phase transition points, respectively, for different κ. The dashed lines are the corresponding point-wise 95% confidence intervals.

6. Concluding remarks

In this article, we propose statistical methods to provide quantitative recommendations for diagnosis and treatment of LRTIs from qrt-LAMP outcomes of eight common bacterial pathogens in the sputum samples of patients. This is the first time that qrt-LAMP is used to test for bacterial pathogens in the research of LRTIs. We first utilize zero-inflated mixture models to address the issue that false-negative results might occur. We employ data augmentation algorithm and MCMC to facilitate the statistical computation and provide analyses on both SCM data and qrt-LAMP data. Baseline covariates are adjusted for LAMP data as well. Further, we investigate how to utilize qrt-LAMP data to identify disease-causing pathogens. Many pathogens live in the host system without causing problems, with very few of them turning into disease-causing pathogens. We propose G-scores, whose definition resembles the techniques used in parallel tempering from MCMC literature, to combine both absolute and relative information of pathogens. These G-scores form the basis of our method to identify the disease-causing pathogens. The change points of G-scores are consistent with phase transitions of biological bacterial growth in the culture medium. Phase transition points are confirmed by the process of the moving sum of partial residuals and identified by fitting the piecewise linear regression model. The results obtained from this observational study can be used to help design phase I or phase II trials, or adaptive treatment strategies towards personalized health care. The statistical methods developed in this article are especially useful when analyzing diseases caused by multiple pathogens and can also be adapted to other similar circumstances.

References

1. Kang Y, Deng R, Wang C, Deng T, Peng P, Cheng X, Wang G, Qian M, Gao H, Han B, Chen Y, Hu Y, Geng R, Hu C, Zhang W, Yang J, Wan H, Yu Q, Wei L, Li J, Tian G, Wang Q, Hu K, Wang S, Wang R, Du J, He B, Ma J, Zhong X, Mu L, Cai S, Zhu X, Xing W, Yu J, Deng M, Gao Z. Etiologic diagnosis of lower respiratory tract bacterial infections using sputum samples and quantitative loop-mediated isothermal amplification. *PLoS ONE* 2012; **7**(6):e38743. DOI: 10.1371/journal.pone.0038743.
2. Mori Y, Kitao M, Tomita N, Notomi T. Real-time turbidimetry of lamp reaction for quantifying template dna. *Journal of Biochemical and Biophysical Methods* 2004; **59**(2):145–157.
3. Poon L, Wong B, Ma E, Chan K, Chow L, Abeyewickreme W, Tangpukdee N, Yuen K, Guan Y, Looareesuwan S. *et al.* Sensitive and inexpensive molecular test for falciparum malaria: detecting *Plasmodium falciparum* DNA directly from heat-treated blood by loop-mediated isothermal amplification. *Clinical Chemistry* 2006; **52**(2):303–306.
4. Njiru Z, Mikosza A, Matovu E, Enyaru J, Ouma J, Kibona S, Thompson R, Ndung'u J. African trypanosomiasis: sensitive and rapid detection of the sub-genus trypanozoon by loop-mediated isothermal amplification (lamp) of parasite DNA. *International Journal for Parasitology* 2008; **38**(5):589–599.
5. Geojith G, Dhanasekaran S, Chandran S, Kenneth J. Efficacy of loop mediated isothermal amplification (LAMP) assay for the laboratory identification of *Mycobacterium tuberculosis* isolates in a resource limited setting. *Journal of Microbiological Methods* 2011; **84**(1):71–73.
6. Hawkins D, Garrett J, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Statistics in Medicine* 2001; **20**(13):1987–2001.
7. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**(1):1–14.
8. Welsh A, Cunningham R, Donnelly C, Lindenmayer D. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling* 1996; **88**(1–3):297–308.
9. Agarwal D, Gelfand A, Citron-Pousty S. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* 2002; **9**(4):341–355.
10. Cunningham R, Lindenmayer D. Modeling count data of rare species: some statistical issues. *Ecology* 2005; **86**(5):1135–1142.
11. Amemiya T. Tobit models: a survey. *Journal of Econometrics* 1984; **24**(1-2):3–61.
12. Liu J. *Monte Carlo Strategies in Scientific Computing*. Springer Verlag: Cambridge, 2008.
13. Moulton L, Halsey N. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; **51**(4):1570–1578.
14. Sigelman L, Zeng L. Analyzing censored and sample-selected data with Tobit and Heckit models. *Political Analysis* 1999; **8**(2):167–182.
15. Van der Vaart A. *Asymptotic Statistics*. Cambridge Univ Pr: Cambridge, 2000.
16. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian Data Analysis*, 2 edn. Chapman and Hall/CRC: Boca Raton, 2004.
17. MacNeill I. Limit processes for sequences of partial sums of regression residuals. *The Annals of Probability* 1978; **6**(4):695–698.
18. MacNeill I. Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *The Annals of Statistics* 1978:422–433.
19. Chu C, Hornik K, Kaun C. MOSUM tests for parameter constancy. *Biometrika* 1995; **82**(3):603–617.

20. Chu C, Stinchcombe M, White H. Monitoring structural change. *Econometrica: Journal of the Econometric Society* 1996; **64**(5):1045–1065.
21. Muggeo V. Estimating regression models with unknown break-points. *Statistics in Medicine* 2003; **22**(19):3055–3071.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.