Received: 28 July 2023 | Revised: 3 June 2024 | Accepted: 20 June 2024

DOI: 10.1111/sjos.12743

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

# A classical hypothesis test for assessing the homogeneity of disease transmission in stochastic epidemic models

**Georgios Aristotelous**[1] | **Theodore Kypraios**[2] | **Philip D. O'Neill**[2]

[1]Department of Computer Science, University of Nicosia, Nicosia, Cyprus

[2]School of Mathematical Sciences, University of Nottingham, Nottingham, UK

**Correspondence**
Theodore Kypraios, School of Mathematical Sciences, University of Nottingham, Nottingham, UK.
Email: theodore.kypraios@nottingham.ac.uk

**Funding information**
University of Nottingham

## Abstract

This paper addresses the problem of assessing the homogeneity of the disease transmission process in stochastic epidemic models in populations that are partitioned into social groups. We develop a classical hypothesis test for completed epidemics which assesses whether or not there is significant within-group transmission during an outbreak. The test is based on time-ordered group labels of individuals. The null hypothesis is that of homogeneity of disease transmission among individuals, a hypothesis under which the discrete random vector of groups labels has a known sampling distribution that is independent of any model parameters. The test exhibits excellent performance when applied to various scenarios of simulated data and is also illustrated using two real-life epidemic data sets. We develop some asymptotic theory including a central limit theorem. The test is practically very appealing, being computationally cheap and straightforward to implement, as well as being applicable to a wide range of real-life outbreak settings and to related problems in other fields.

KEYWORDS

epidemic model, hypothesis test, infectious disease data

# 1 | INTRODUCTION

This paper is concerned with the problem of assessing the assumption of homogeneity of disease transmission in stochastic epidemic models in populations that are partitioned into social groups. Such groups may correspond to physical locations such as households or classrooms, or to collections of similar individuals such as age groups. Our aim is to assess whether or not transmission within groups differs from that between groups.

We introduce a new classical hypothesis test which conducts this assessment. The test is fundamentally different to existing approaches in the literature (see Section 7.2), requiring no model fitting and not relying on asymptotic approximations when implemented. The null hypothesis of the test is the assumption of homogeneity of disease transmission, so that the groups have no direct bearing on the outbreak. The test is based on the group labels of infected individuals ordered by time, which we assume are available from observed data. The key idea is that greater within-group transmission will lead to greater clustering of group labels. Under the null hypothesis, the discrete random vector of group labels has a known sampling distribution which does not depend on any model parameters. The test statistic has an ordinal interpretation, where lower values provide greater evidence against the null hypothesis. We demonstrate the performance of the test via an extensive simulation study and by application to two real data sets. We also develop some asymptotic theory including a central limit theorem.

The paper is structured as follows. Section 2 contains preliminary information. Section 3 defines the test and its interpretation and implementation. In Section 4, we present a simulation study, and in Section 5 we apply the test to real data. In Section 6, we present some asymptotic results and in Section 7 we give some additional perspectives and commentary. All computer code used to produce our results was written in the statistical programming language R Core Team (2019) and the code used to implement the test is provided in Appendix S1.

# 2 | PRELIMINARIES

## 2.1 | Terminology, notation, and data setting

Stochastic epidemic models are typically defined at an individual level and classify individuals at a given time according to their state. Specifically, susceptible (S) individuals are those at risk of contracting the disease, infective (I) individuals are those who have the disease and can pass it on to others, and removed (R) individuals are those who have had the disease but can neither be infected again nor infect others. In practice the removed state could refer to several conditions including immunity, isolation and death. Models in which individuals can move from S to I to R are known as SIR models. Similarly, SEIR models also include exposed (E) individuals who have been infected but are not yet able to infect others.

Consider an epidemic in a closed population of $C$ individuals which is partitioned into $l$ groups labelled $1, 2, \ldots, l$, with group $m$ consisting of $C_m$ individuals so that $C = \sum_{m=1}^{l} C_m$. For ease of exposition our focus will be on epidemics that have been completed, that is, those where there are no infectives left in the population at the end of the outbreak; however, as explained in Section 7, the methodology we develop could in principle also be applied to ongoing outbreaks.

We write $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)$ to denote the time-ordered event times and $\boldsymbol{g}^{\boldsymbol{e}} = \left(g_1^{\boldsymbol{e}}, g_2^{\boldsymbol{e}}, \ldots, g_n^{\boldsymbol{e}}\right)$ for their corresponding group labels, also ordered by time. Thus the individual with event time $e_k$ belongs to group $g_k^{\boldsymbol{e}}$. If $\boldsymbol{e}$ and $\boldsymbol{g}$ are random vectors, under some specified sampling distribution, we denote them as $\boldsymbol{e}^{\text{sam}} = \left(e_1^{\text{sam}}, e_2^{\text{sam}}, \ldots, e_n^{\text{sam}}\right)$ and $\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}}} = \left(g_1^{\boldsymbol{e}^{\text{sam}}}, g_2^{\boldsymbol{e}^{\text{sam}}}, \ldots, g_n^{\boldsymbol{e}^{\text{sam}}}\right)$, respectively, whereas

if they represent observed data, we write $\boldsymbol{e}^{\text{obs}} = \left(e_1^{\text{obs}}, e_2^{\text{obs}}, \ldots, e_n^{\text{obs}}\right)$ and $\boldsymbol{g}^{\boldsymbol{e}^{\text{obs}}} = \left(g_1^{\boldsymbol{e}^{\text{obs}}}, g_2^{\boldsymbol{e}^{\text{obs}}}, \ldots, g_n^{\boldsymbol{e}^{\text{obs}}}\right)$. Note that the number of observed event times $n$ equals the number of individuals ever contracting the disease and is treated as a fixed nonrandom quantity whose value is known from the observed data.

We consider a generic data observation setting where the event times may refer to any temporal data for which the corresponding time-ordering of the group labels can be considered to convey information on the amount of within-group transmission. For example in an SEIR model, the event times might correspond to exposure times, infection times or removal times. We assume that event times which are not infection times are still able to convey information about transmission itself. In reality this is not unreasonable since SIR and SEIR models typically assume that the times spent in the I and E states are identically distributed for different individuals, and hence there is no systematic distortion of the event ordering if we were to view the order of removal times as a proxy for the order of infection times.

In practice, it is often the case that only case-detection times of individuals are observed, which for modelling purposes are often treated as removal times (e.g. Neal & Roberts, 2005; O'Neill & Roberts, 1999; Xiang & Neal, 2014). This last assumption is not necessary for our purposes since the test can be applied without the specification of a model, a point discussed in Section 7. However, we shall consider a specific model to both motivate the construction of our test and also assess its performance via a simulation study, with the model providing a means of generating suitable data.

For ease of exposition we will restrict our attention to the use of SIR models. This is because SIR models, unlike SEIR models, have no exposure period and it is therefore simpler to describe the amount of distortion that may be introduced when basing the test on different types of event times. However, the test is equally applicable in both SIR and SEIR settings, or indeed any disease transmission model in which individuals are partitioned into groups. In addition, although our motivation and focus is on epidemic models, the test we propose has potential application in other fields, as explained in more detail in section 7.

## 2.2 | Two-level-mixing SIR model

The *two-level-mixing SIR model* (Ball et al., 1997) is defined as follows. Consider a closed population of $C$ individuals, partitioned into groups as described in Section 2.1. Initially, $N$ individuals are susceptible, $K$ are infective and none are removed. Each infective individual remains so for a period of time known as the infectious period before becoming removed. Infectious periods of different individuals are assumed to be independent and distributed according to some random variable $T_D$. During its infectious period, an infective has contacts with each other individual in the population at the time points of a homogeneous Poisson process of rate $\beta_G$. Additionally, each infective has contacts with each other individual in their group at the time points of a homogeneous Poisson process of rate $\beta_L$. If a contacted individual is susceptible at the time of a contact, they immediately become infective. All Poisson processes are assumed to be mutually independent.

In what follows we assume for simplicity that $K = 1$ but this assumption can easily be relaxed. The epidemic ends when no infectives are left in the population.

Note that the overall infection process is described by two independent infection processes: one that models contacts at the population level, with governing parameter $\beta_G$, and one modeling contacts at the group level, with associated parameter $\beta_L$. We refer to infections occurring in these processes as *global infections* and *local infections* in the obvious manner.

A key parameter associated with the two-level-mixing model is the reproduction number $R_*$, which is loosely defined as the average number of groups infected by a typically infected group in a totally susceptible population (Ball et al., 1997). In the particular case where all groups are of equal size $C_H$, and the population size $C$ becomes large in such a way that the number of groups $l$ becomes large but the group size $C_H$ remains fixed then $R_* = \mu R_G$, where $\mu$ is the expected number of individuals infected in a within-group epidemic with one initial infective and only local infections, and $R_G$ is the basic reproduction number for the model in which all groups are of size 1 and only global infections occur.

In this paper we consider a specific choice of infectious period distribution, namely Exponential, with rate parameter $\gamma$ and probability density function (p.d.f.) $f(x;\gamma) = \gamma \exp(-\gamma x), x \geq 0;\ \gamma > 0$, and we denote this version of the two-level-mixing model as Exp-2L. This choice of infectious period distribution appears frequently in the epidemic modelling literature.

## 3 | GROUP LABEL TEST

### 3.1 | Rationale

Intuitively, if there is a *within-group-transmission effect* in the data, in which an infective individual is more likely to infect individuals in their group rather than outside of it, then the time-ordered sequence of group labels is likely to include clusters of labels from the same group close together. Conversely, if there is no within-group-transmission effect in the data the labels of each group will typically appear with no specific pattern over time. For example, consider an outbreak of $n = 9$ infection events, occurring in a population with a $l = 4$ groups and group size $C_m = 3$, $m = 1, 2, 3, 4$. In the case of extreme within-group-transmission effect a typical realized group label data set may look like $\boldsymbol{g^e} = (3, 3, 3, 1, 1, 1, 4, 4, 4)$, whereas in the case of no within-group-transmission effect, a realized group label data set may typically look like $\boldsymbol{g^e} = (3, 1, 4, 2, 4, 4, 1, 3, 2)$.

### 3.2 | Null hypothesis and test statistic

The null hypothesis, denoted $H_0$, is the assumption of homogeneity of disease transmission among individuals. A key observation is that under $H_0$, the sampling distribution of the discrete random vector $\boldsymbol{g^e}^{\text{sam}}$ is known and independent of any model parameters, depending only on $n$, $l$, and $C_m$. This follows from the fact that in an SIR or SEIR model with homogeneous mixing, all susceptibles in the population are equally likely to be contacted by an infective. Thus a realization from the sampling distribution of $\boldsymbol{g^e}^{\text{sam}}$ under $H_0$ can be drawn by sampling, without replacement and uniformly at random, a sequence of length $n$ from a set of size $C$ whose elements are the group labels of each individual in the population (see Appendix S2 for the analytic expression of the joint probability mass function of the random vector $\boldsymbol{g^e}^{\text{sam}} \sim H_0$). The fact that this remains true if event times are not infection times but associated event times, such as removal times, follows from the assumptions made in Section 2.1.

We now construct a test statistic $T$ on the space of the $n$-dimensional group label vectors. The test statistic is constructed to quantify the idea of group label clustering, described in Section 3.1, and is therefore given an ordinal nature, where the higher (lower) the within-group-transmission effect in the data, the lower (higher) the value of $T$. Specifically, $T$ is defined as $T(\boldsymbol{g^e}) = \sum_{m=1}^{l} s_{\boldsymbol{g^e}}^{(m)}$,

where $s^{(m)}_{\boldsymbol{g}^e}$ is a measure of spread for the labels of group $m$ that appear in $\boldsymbol{g}^e$, so that higher (lower) levels of within-group-transmission effect in the data are associated with lower (higher) values of $s^{(m)}_{\boldsymbol{g}^e}$. In the following we omit the $\boldsymbol{g}^e$ dependence for notational simplicity.

We specify $s^{(m)}$, $m = 1, 2, \ldots, l$ as follows. Until stated otherwise, suppose that $C_m \geq 2$. Let $\nu^{(m)}$ denote the number of times that the label of group $m$ appears in $\boldsymbol{g}^e$ and, assuming that $\nu^{(m)} \geq 1$, let $\boldsymbol{f}^{(m)} = \left(f_1^{(m)}, f_2^{(m)}, \ldots, f_{\nu^{(m)}}^{(m)}\right)$ denote the vector of indices of $\boldsymbol{g}^e$ at which the labels of group $m$ appear; note that for $\nu^{(m)} = 1$, $\boldsymbol{f}^{(m)}$ reduces to a scalar, i.e. $\boldsymbol{f}^{(m)} = f_1^{(m)}$ where $f_1^{(m)}$ is the index of $\boldsymbol{g}^e$ where the first and only appearance of the label of group $m$ occurs. For example, if $\boldsymbol{g}^e = (3, 1, 4, 2, 4, 4, 1, 3)$, then $\nu^{(3)} = 2$ with $\boldsymbol{f}^{(3)} = (1, 8)$ and $\nu^{(2)} = 1$ with $\boldsymbol{f}^{(2)} = 4$. We consider three cases for $\nu^{(m)}$.

(i) $\nu^{(m)} \geq 2$: First, suppose that the label of group $m$ appears twice or more in $\boldsymbol{g}^e$, so that measuring spread is possible. In this case, we set $s^{(m)} = f_{\nu^{(m)}}^{(m)} - f_1^{(m)} - (\nu^{(m)} - 1)$. For example, if $\boldsymbol{g}^e = (3, 1, 4, 2, 4, 4, 1, 3, 2, 5, 5, 5)$ then $s^{(1)} = 7 - 2 - (2 - 1) = 4$ and $s^{(2)} = 9 - 4 - (2 - 1) = 4$. Note that $s^{(m)}$ can be calculated by counting the number of nongroup-$m$ labels found between the first and last group $m$ label of $\boldsymbol{g}^e$. Thus $s^{(m)}$ can be thought of as penalizing group $m$ according to the extent that it deviates from the most obvious realization of within-group-transmission effect, where its labels appear in consecutive order (such as group 5 in the example above).

If $\nu^{(m)} = 0$ and $\nu^{(m)} = 1$ then measuring the spread of the labels of group $m$ is not possible, and so $s^{(m)}$ is defined differently, but still with the intention that $s^{(m)}$ quantifies the within-group-transmission effect associated with the labels of group $m$.

(ii) $\nu^{(m)} = 0$: If $\nu^{(m)} = 0$, we set $s^{(m)} = 0$. This assignment stems from the idea that, when there is a within-group-transmission effect in the data, given $n$, the label of less rather than more groups should appear in $\boldsymbol{g}^e$. Note that this assignment is also consistent with the ordinal nature we aim to give to $T$, in the sense that it is the only one that allows $T$ to attain its minimum value of 0 when evaluated at realizations of the most obvious within-group-transmission effect. To see this, recall the example realization of section 3.1, $\boldsymbol{g}^e = (3, 3, 3, 1, 1, 1, 4, 4, 4)$, with obvious within-group-transmission effect. In this case $\nu^{(1)} = \nu^{(3)} = \nu^{(4)} = 3$, thus, from the definition of $s^{(m)}$ for $\nu^{(m)} \geq 2$, we have $s^{(1)} = s^{(3)} = s^{(4)} = 0$, and therefore, $T(\boldsymbol{g}^e) = \sum_{m=1}^{4} s^{(4)}$ can only be 0 if group 2, which has $\nu^{(2)} = 0$, has $s^{(2)} = 0$.

(iii) $\nu^{(m)} = 1$: Finally we consider the case where $\nu^{(m)} = 1$ for any group $m$ whose label appears in $\boldsymbol{g}^e$. For example, if $n = 5$, $l = 6$, $C_m = 3$ and $m = 1, 2, \ldots, 6$, one such realization might be $\boldsymbol{g}^e = (1, 2, 3, 4, 5)$. This case we refer to as a case of *negative within-group-transmission effect*, meaning that an infective individual is more likely to infect susceptible individuals outside their group rather than within it. In fact the present example is the most extreme case of such negative within-group-transmission effect, in the sense that the outbreak progresses only between and not within groups. In order for $T$ to have the desired ordinal nature, realizations of this type must produce the maximum value of $T$. To this end, considering the fact that $n \geq f_{\nu^{(m)}}^{(m)}$ in all instances, $s^{(m)}$ is specified as in the case of $\nu^{(m)} \geq 2$ (see above) with the difference being that $f_{\nu^{(m)}}^{(m)}$ is replaced by $n$, that is, as $s^{(m)} = n - f_1^{(m)} - (\nu^{(m)} - 1) = n - f_1^{(m)}$. For example for $\boldsymbol{g}^e = (2, 3, 3, 5, 3, 4, 4)$, $s^{(2)} = 7 - 1 = 6$ and $s^{(5)} = 7 - 4 = 3$. Notice that, similarly to the case where $\nu^{(m)} \geq 2$, $s^{(m)}$ can be calculated by counting the number of non-group $m$ labels from the first (and only) group $m$ label until the last index of $\boldsymbol{g}^e$. This counting representation of $s^{(m)}$ highlights how a group $m$, whose label first appears in $\boldsymbol{g}^e$ at index $f_1^{(m)}$, receives the maximum value of $s^{(m)}$ in the instance that its label does not appear again (i.e., in the instance that $\nu^{(m)} = 1$).

We note that from a practical point of view the notion of negative within-group-transmission effect is often not relevant since it is not usually plausible for real-life epidemic outbreaks to have

higher transmission between rather than within groups. Nonetheless, ensuring that $T$ behaves sensibly under such cases is essential in providing it with the required ordinal nature.

We have so far assumed that $C_m \geq 2$. Lastly, suppose that $C_m = 1$. In this case the label of group $m$ can only appear zero or one times, and measuring spread in such cases is neither possible nor meaningful. Thus, the only sensible assignment is $s^{(m)} = 0$.

We summarise the definitions of $T$ and $s^{(m)}$ in Equation (1) below.

$$T(\boldsymbol{g}^{\boldsymbol{e}}) = \sum_{m=1}^{l} s^{(m)}, \quad \text{where } s^{(m)} = \begin{cases} 0 & \text{if } \nu^{(m)} = 0 \text{ or } C_m = 1, \\ n - f_1^{(m)} & \text{if } \nu^{(m)} = 1 \text{ and } C_m \geq 2, \\ f_{\nu^{(m)}}^{(m)} - f_1^{(m)} - (\nu^{(m)} - 1) & \text{if } \nu^{(m)} \geq 2. \end{cases} \tag{1}$$

## 3.3 | Implementation and interpretation

Since $T$ is a deterministic function of $\boldsymbol{g}^{\boldsymbol{e}}$ (see Equation 1), and the distribution of $\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}}} \sim H_0$ is known and independent of any model parameters, the distribution of $T^{\text{sam}} \sim H_0$ is also independent of any model parameters, and independent sampling from $T^{\text{sam}}$ can easily be achieved by first drawing an independent sample $\left\{\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(1)}}, \boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(2)}}, \ldots, \boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(S)}}\right\}$ from $\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}}} \sim H_0$, following the procedure described in Section 3.2, and by then evaluating $T$ at each realization $\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(s)}}$, $s = 1, 2, \ldots, S$, using equation (1). The resulting sample, $\{T^{\text{sam}(1)}, T^{\text{sam}(2)}, \ldots, T^{\text{sam}(S)}\}$, where $T^{\text{sam}(s)} := T\left(\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(s)}}\right)$, $s = 1, 2, \ldots, S$, can be used to test $H_0$ visually, by imposing the observed value $T^{\text{obs}} := T\left(\boldsymbol{g}^{\boldsymbol{e}^{\text{obs}}}\right)$ on a histogram of the sampled values, and quantitatively by calculating the $p$-value, defined as $p\text{-value} := P(T^{\text{sam}} \leq T^{\text{obs}})$, and calculated as

$$\begin{aligned} p\text{-value} &= P(T^{\text{sam}} \leq T^{\text{obs}}) = \mathrm{E}\left(\mathbb{1}_{\{T^{\text{sam}} \leq T^{\text{obs}}\}}\right) \\ &= \int \mathbb{1}_{\{T^{\text{sam}} \leq T^{\text{obs}}\}} \pi\left(\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}}} | H_0\right) d\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}}} \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}_{\{T^{\text{sam}(s)} \leq T^{\text{obs}}\}}, \end{aligned} \tag{2}$$

where $\mathbb{1}_A$ denotes the indicator function of the event $A$. All steps required to implement the test are listed in Algorithm 1.

**Algorithm 1.** Group label test

**input:** $\boldsymbol{g}^{\boldsymbol{e}^{\text{obs}}} = \left(g_1^{\boldsymbol{e}^{\text{obs}}}, g_2^{\boldsymbol{e}^{\text{obs}}}, \ldots, g_n^{\boldsymbol{e}^{\text{obs}}}\right)$ and $C_m$, $m = 1, 2, \ldots, l$;
**output:** $T^{\text{obs}}$, $\left\{T^{\text{sam}(1)}, T^{\text{sam}(2)}, \ldots, T^{\text{sam}(S)}\right\}$ and the $p$-value;

1. Calculate $T^{\text{obs}} := T\left(\boldsymbol{g}^{\boldsymbol{e}^{\text{obs}}}\right)$ using Equation (1);
2. Sample from $T^{\text{sam}} \sim H_0$ :
   **for** $s = 1, 2, \ldots, S$ **do**
   i. Choose uniformly at random a permutation of $n$ out of the total $C$ individuals and record their corresponding group labels $\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(s)}}$;
   ii. Calculate $T^{\text{sam}(s)} := T\left(\boldsymbol{g}^{\boldsymbol{e}^{\text{sam}(s)}}\right)$ using Equation (1);
   **end for**
3. Impose $T^{\text{obs}}$ on the histogram of $\{T^{\text{sam}(1)}, T^{\text{sam}(2)}, \ldots, T^{\text{sam}(S)}\}$ and calculate the $p$-value using Equation (2).

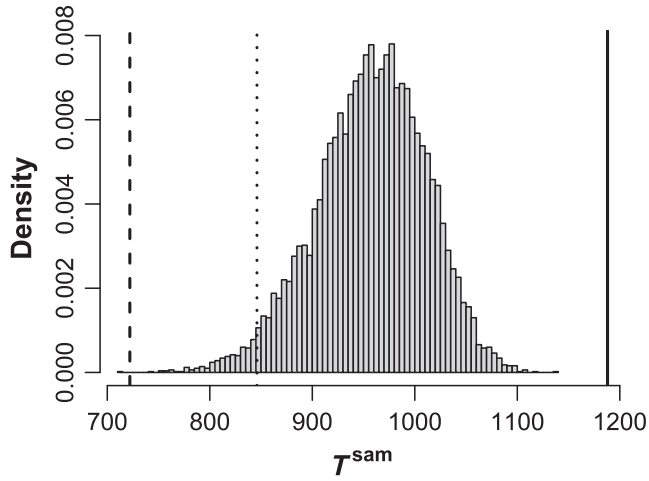


**FIGURE 1** Example of assessing homogeneity of transmission using the classical hypothesis test for group label data. Observed data are generated from an Exp-2L model ($N = 99$, $C_H = 5$, $R_* = 2.5$, $\mu = 1.89$ and $\gamma = 0.1$). The plot is the histogram of 10,000 realizations from the sampling distribution of $T^{\text{sam}} \sim H_0$ with the observed value (based on infections) $T_{\boldsymbol{i}}^{\text{obs}} = 722$ (dashed line), the observed value (based on removals) $T_{\boldsymbol{r}}^{\text{obs}} = 846$ (dotted line) and the maximum value of $T = 1188$ (solid line) imposed. The $p$-values are (based on infections) $p\text{-value}_{\boldsymbol{i}} = .001$ and (based on removals) $p\text{-value}_{\boldsymbol{r}} = .025$.

Figure 1 shows sampled values from $T^{\text{sam}} \sim H_0$ when the test is applied to an example simulated data set generated from the Exp-2L model, having mild to moderate within-group-transmission effect. The observed values of $T$, along with their corresponding $p$-values, are interpreted as follows. Values that fall well within the support of $T^{\text{sam}} \sim H_0$ (i.e., closer to the mode rather than the tails of the histogram of $T^{\text{sam}} \sim H_0$) are consistent with $H_0$ and provide no evidence against it; in such cases the associated p-value is not too close to the extreme values of 0 or 1. As values move to the left tail (and beyond) of the histogram of $T^{\text{sam}} \sim H_0$ and toward the minimum value of $T$ which is 0 (which represents the most obvious case of within-group-transmission effect) they become inconsistent with $H_0$, and provide increasing evidence against it and in favor of the hypothesis $H_L$, that there is within-group-transmission effect in the data; the corresponding $p$-value being close to or equal to 0.

We note that a histogram of $T^{\text{sam}} \sim H_0$, with the observed value of $T$ imposed, is more informative than the $p$-value alone. For example, a p-value of 0 might correspond to the observed value of $T$ being very near to the left tail of the histogram of $T^{\text{sam}} \sim H_0$ or very far from it, closer to the minimum value of $T = 0$. Although the $p$-value would be the same in these two cases, the amount of evidence against $H_0$ and in favor of $H_L$ would be higher in the second case.

We now discuss the results of the test on the example data set. Notice that we conduct two assessments, one in which the event times are infection times and one in which they are removal times. To distinguish between the two, the observed group label vector, the observed value of $T$, and the associated $p$-value, are respectively denoted $\boldsymbol{g}^{\boldsymbol{i}^{\text{obs}}}$, $T_{\boldsymbol{i}}^{\text{obs}}$ and $p\text{-value}_{\boldsymbol{i}}$ when event times are infection times, and as $\boldsymbol{g}^{\boldsymbol{r}^{\text{obs}}}$, $T_{\boldsymbol{r}}^{\text{obs}}$ and $p\text{-value}_{\boldsymbol{r}}$ when event times are removal times. We adopt this notational convention in subsequent parts of this paper. For the infection-based assessment, we have that $T_{\boldsymbol{i}}^{\text{obs}} = 722$ and $p\text{-value}_{\boldsymbol{i}} = .001$, and for the removal based, $T_{\boldsymbol{r}}^{\text{obs}} = 846$ and $p\text{-value}_{\boldsymbol{r}} = .025$. Both $p$-values are small enough to provide evidence against $H_0$ and in favor of $H_L$ which is sensible given that the data set in question has mild to moderate

within-group-transmission effect. The amount of evidence is higher in the infection-based assessment, which again is to be expected since the removal times correspond to an i.i.d. random shift of the infection times as described in Section 2.2.

The results of this example are indicative of the test's ability to identify a within-group-transmission effect when present. They also provide an indication of how the test results might differ when based on different types of event times. A much more detailed exploration of the test's performance is given in the simulation study of the following section.

# 4 | SIMULATION STUDY

## 4.1 | Simulation and run conditions

To assess the performance of the test on data of varying levels of within-group-transmission effect, we conduct a large-scale simulation study. We generate data from the Exp-2L model under four simulation scenarios, the conditions of which are given in Table 1.

The parameters for the four simulation scenarios are set as follows. All groups are set to have equal size $C_H = 5$. Having equal group sizes is not necessary for the test, as illustrated below in Section 5, but is useful for the simulation study since it provides us with a simple form for the reproduction number parameter $R_*$, which in turn allows us to clearly quantify the within-group-transmission effect in each scenario, as we now explain.

The global and local infection rate parameters, $\beta_G$ and $\beta_L$, are set so that the four simulation scenarios are of increasing within-group-transmission effect. Specifically, recall the three quantities $R_*$, $\mu$ and $R_G$ from Section 2.2, which can roughly be thought of as quantifying overall, within-group and between-group transmission, respectively. We keep $R_*$ fixed at 2.5 in all scenarios. Given a value for $R_*$, $R_G$, and $\mu$ are inversely proportional and the within-group-transmission effect in the data increases as $\mu$ increases (and $R_G$ decreases).

To guide our choices for the model parameters, recall that the within-group outbreak can be seen as an outbreak from a standard SIR model, with individual-to-individual infection rate $\beta_L$, $C_H - 1$ initial susceptibles and basic reproduction number $R_0^H = \beta_L(C_H - 1)/\gamma$. Choosing a value of $R_0^H$ therefore specifies $\beta_L$. In turn, given $\beta_L$, $\mu$ can be determined from eq. (2.25) of Ball (1986). Finally, given $\mu$ and $R_*$, $R_G$ and $\beta_G$ are specified by the equations $R_* = \mu R_G$ and $R_G = N\beta_G/\gamma$, respectively.

Scenarios 1–4 are such that $R_0^H = 0.5, 1.0, 2.5, 5.0$ and in turn, $\mu = 0.61, 1.32, 2.44$ and $3.88$, respectively. To see the extent of the within-group-transmission effect in each scenario we also calculate the mean proportion of local infections (over total infections) $\overline{p_L}$ under the sampling

**TABLE 1** Simulation conditions for the simulation study.

| | Data generating process | Parameter values | $\overline{p_L}$ |
|---|---|---|---|
| Scenario 1 | Exp-2L | $R_* = 2.5, \gamma = 0.1, \mu = 0.61$ | 0.09 |
| Scenario 2 | Exp-2L | $R_* = 2.5, \gamma = 0.1, \mu = 1.32$ | 0.27 |
| Scenario 3 | Exp-2L | $R_* = 2.5, \gamma = 0.1, \mu = 2.44$ | 0.51 |
| Scenario 4 | Exp-2L | $R_* = 2.5, \gamma = 0.1, \mu = 3.88$ | 0.70 |

*Notes*: Each simulation scenario consists of four rounds, where the number of initial susceptibles $N$ is set at 99, 199, 499, and 999, respectively. For each round 500 datasets are generated. The number of individuals in each group is set at $C_H = 5$, in all instances.

distribution of the model. The values of $\overline{p_L}$ for scenarios 1–4 are 0.09, 0.27, 0.51, and 0.70, respectively, and they are also given in Table 1 for reference. Scenarios 1–4 can be interpreted as having very mild, mild, apparent, and very apparent within-group-transmission effect, respectively.

Each simulation scenario consists of four rounds, with the number of initial susceptibles, $N$, being set at 99, 199, 499, and 999, respectively. The number of initial susceptibles can be thought of as quantifying the dimension of the observed data insofar as the total number of ever-infected individuals, $n$, is likely to increase with $N$, given $R_*$. Thus the different rounds are used to examine if and how the performance of the test changes as the dimension of the observed data increases. For each round, we generate 500 data sets, to capture sampling variability. For all simulated data sets, the initial infective is chosen uniformly at random from the population.

Following Algorithm 1, the test is applied to each generated data set twice, once using infection times and once using removal times. Specifically, for each simulated data set, we first obtain an independent sample of size 10,000 from $T^{\text{sam}} \sim H_0$, and then calculate and record the two observed values of $T$, $T_{\boldsymbol{i}}^{\text{obs}}$, and $T_{\boldsymbol{r}}^{\text{obs}}$, and their corresponding $p$-values, $p$-value$_{\boldsymbol{i}}$, and $p$-value$_{\boldsymbol{r}}$. The value of the Exponential infectious period parameter $\gamma$ is set at 0.1 so that the mean of the infectious period is $\mathrm{E}(T_D) = 10$ and its variance is $\mathrm{Var}(T_D) = 100$. Such an infectious period has more variability than is typically encountered in real-life diseases, but it is useful for the simulation study since it creates considerable distortion for the test based on removal data.

## 4.2 | Results

We summarize the results of the simulation study using tables that give the median (95% quantile interval) $p$-value$_{\boldsymbol{i}}$ and $p$-value$_{\boldsymbol{r}}$, for each round and scenario, while plots of all the $p$-values and the test's power from the simulation study are given in Appendices S3 and S4 respectively.

### 4.2.1 | Infection-based assessment

We first consider infection based results. Table 2 shows that the results are sensible in all scenarios and rounds and that the effect of increasing $N$ and $\mu$ is the desirable one in the sense that larger values of $N$ ($\mu$) yield smaller $p$-value$_{\boldsymbol{i}}$s, for a given $\mu$ ($N$). For example, in scenario 1 the median (95% quantile interval) $p$-value$_{\boldsymbol{i}}$ is 0.25 (0, 0.90), 0.17 (0, 0.85), 0.05 (0, 0.67), and 0.01 (0, 0.46) for $N = 99$, 199, 499, and 999, respectively, suggesting that even in the presence of a very mild within-group-transmission effect the test would still provide adequate evidence against $H_0$ and in favor of $H_L$, if the dimension of the data is large enough. The power of the test is also evident in

**TABLE 2** Median (95% quantile interval) $p$-value from the household labels test based on observing infection times, $p$-value$_{\boldsymbol{i}}$, for the simulation study.

| | $N = 99$ | $N = 199$ | $N = 499$ | $N = 999$ |
|---|---|---|---|---|
| Scenario 1 | 0.25 (0, 0.90) | 0.17 (0, 0.85) | 0.05 (0, 0.67) | 0.01 (0, 0.46) |
| Scenario 2 | 0.01 (0, 0.40) | 0 (0, 0.10) | 0 (0, 0) | 0 (0, 0) |
| Scenario 3 | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| Scenario 4 | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |

*Note*: Simulation conditions for each scenario are given in Table 1.

scenario 2, where the within-group-transmission effect in the data is still relatively mild, but the sampling distribution of the $p$-value$_i$ is concentrated near 0, even for the smaller values of $N$. It is also worth noticing that for scenarios 3 and 4, where the two-level transmission effect is more apparent, the sampling distribution of the $p$-value$_i$ is consistently a point mass at 0, meaning that in such instances, as appropriate, the evidence against $H_0$ and in favor of $H_L$ would systematically be conclusive.

#### 4.2.2 | Removal-based assessment

As can be seen from Table 3, removal-based results are very similar to infection-based results, with the difference being that $p$-value$_r$ is typically slightly higher than $p$-value$_i$, which is to be expected because of the distortion introduced by the infectious period. As for the infection-based assessment, results are sensible in all scenarios and rounds and the effect of $N$ and $\mu$ is the desirable one, so that the larger the value of $N$ ($\mu$), for a given $\mu$ ($N$), the smaller the $p$-value$_r$. The only scenario in which there is no strong evidence against $H_0$ is scenario 1, with the median (95% quantile interval) $p$-value$_r$ being 0.44 (0.02, 0.96), 0.41 (0.01, 0.94), 0.35 (0.01, 0.95), and 0.26 (0.01, 0.95) for $N = 99$, 199, 499, and 999, respectively. Nonetheless, when considering that scenario 1 represents a case of very mild within-group-transmission effect (where the mean proportion of local infections is only $\overline{p_L} = 0.09$), these results are still very encouraging. In scenario 2, where the within-group-transmission effect becomes a little less mild, the test exhibits increased power and successfully detects the effect in the data, with the median (95% quantile interval) $p$-value$_r$ being equal to 0.08 (0, 0.71), 0.02 (0, 0.53), 0 (0, 0.10) and 0 (0, 0.01) for $N = 99$, 199, 499, and 999, respectively. For scenarios 3 and 4, which, respectively, represent apparent and very apparent within-group-transmission effect, the sampling distribution of the $p$-value$_r$ is consistently a point mass at 0, as appropriate.

## 5 | APPLICATIONS TO REAL DATA

### 5.1 | Abakaliki smallpox outbreak

#### 5.1.1 | Data description

We first consider the data set obtained from a smallpox outbreak in Abakaliki, Nigeria, in 1967 (Bailey, 1975, page 125). This is a widely studied data set, either analyzed to understand the outbreak (see e.g., Eichner & Dietz, 2003; Stockdale et al., 2017) or used to illustrate new statistical methodology (see e.g. O'Neill & Roberts, 1999; Boys & Giles, 2007; Clancy &

**TABLE 3** Median (95% quantile interval) $p$-value from the household labels test based on observing removal times, $p$-value$_r$, for the simulation study.

| | $N = 99$ | $N = 199$ | $N = 499$ | $N = 999$ |
|---|---|---|---|---|
| Scenario 1 | 0.44 (0.02, 0.96) | 0.40 (0.01, 0.96) | 0.35 (0.01, 0.95) | 0.26 (0.01, 0.95) |
| Scenario 2 | 0.08 (0, 0.71) | 0.02 (0, 0.53) | 0 (0, 0.10) | 0 (0, 0.01) |
| Scenario 3 | 0 (0, 0.17) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| Scenario 4 | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |

*Note*: Simulation conditions for each scenario are given in Table 1.

O'Neill, 2008; Kypraios et al., 2017). The data are described in detail in Thompson and Foege (1968) and Eichner and Dietz (2003) and consist, among other things, of information on the 32 individuals who became infected among 251 individuals who were living in nine compounds in the city.

For the purposes of this analysis, we restrict attention to those individuals who lived in a compound. Specifically we consider a population of size $C = 251$, where the individuals are partitioned into $l = 9$ compounds, labeled as $1, 2, \ldots, 9$, with each compound $m$ consisting of $C_m$ individuals, $m = 1, 2, \ldots, 9$, where $(C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9) = (33, 15, 10, 33, 22, 43, 20, 42, 33)$ (see Eichner and Dietz 2003, table 2). Our observed event times are the $n = 32$ case-detection times. These times are given in days and some individuals have the same case-detection day, so there is more than one possible time orderings of the event times. The results we present here are based on the event times being ordered as in Thompson and Foege (1968, table 1), for which the corresponding compound label vector is given by $\boldsymbol{g}^{e^{\text{obs}}} = (1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 4, 5, 1, 1, 1, 1, 5, 2, 1, 2, 6, 5, 2, 7, 4, 2, 2, 8, 3, 9, 5, 2)$, but also consider all other options as explained below.

### 5.1.2 | Results and conclusions

We apply Algorithm 1 with a sample of size 10,000 from $T^{\text{sam}} \sim H_0$, and present the results in Figure 2. The position of the observed value $T^{\text{obs}}$ on the histogram of $T^{\text{sam}} \sim H_0$, and the small $p$-value = .004, suggest that the data are inconsistent with $H_0$ and thus provide strong evidence in favor of the hypothesis of a within-compound-transmission effect. This conclusion is in agreement with that found in the literature (see e.g., Thompson & Foege, 1968; Eichner & Dietz, 2003; Stockdale et al., 2017).

There are 31 additional possible orderings of the compound label vector, and we calculated the $p$-value for each. The range of these 31 $p$-values was from .003 to .006, showing that the test conclusion is not sensitive to the choice of ordering.

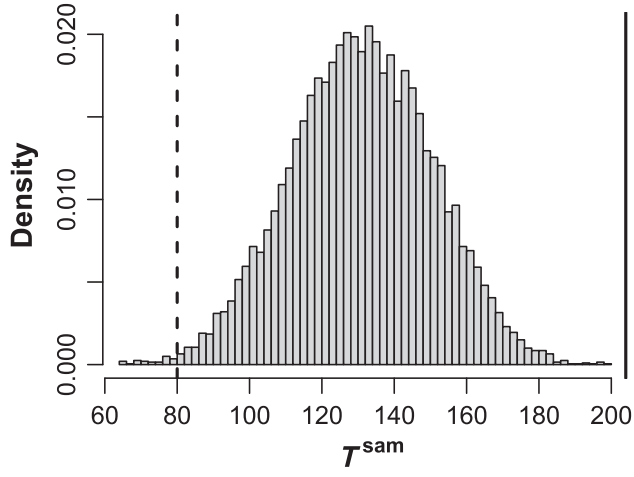


**FIGURE 2** Application of the classical hypothesis test for compound label data on the Abakaliki outbreak data. The plot is the histogram of 10,000 realizations from the sampling distribution of $T^{\text{sam}} \sim H_0$ with the observed value $T^{\text{obs}} = 80$ (dashed line) and the maximum value of $T = 204$ (solid line) imposed. The test $p$-value is equal to 0.004.

## 5.2 | Tristan da Cunha respiratory disease outbreak

### 5.2.1 | Data description

We next consider a data set described and analyzed in Becker and Hopper (1983) and Hayakawa et al. (2003). The data set contains case-detection times of individuals during a common cold outbreak that took place from in 1967 on the island of Tristan da Cunha. The total population of the island was 255 individuals, partitioned into three age-groups, namely infants (age 0–4; group 1), children (age 5–14; group 2), and adults (age 15 or above; group 3). There was one unidentified case, and so we assume a population of $C = 254$ individuals divided into the three age groups as $(C_1, C_2, C_3) = (25, 36, 193)$. The outbreak led to a total of $n = 40$ cases, for which case-detection times were reported in days. We used these dates as the observed event times. As for the Abakaliki dataset some individuals had the same case-detection days, leading to multiple possible time orderings of the event times. We present results based on a randomly selected ordering of the case-detection times, and investigate the sensitivity of the results over different orderings below. The age group label vector corresponding to our chosen ordering is given by $\boldsymbol{g}^{e^{\text{obs}}} = (3, 3, 2, 3, 1, 1, 2, 1, 1, 2, 3, 3, 3, 1, 1, 3, 3, 1, 1, 2, 3, 2, 3, 3, 3, 3, 3, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3)$.

### 5.2.2 | Results and conclusions

We apply Algorithm 1 with a sample of size 10,000 from $T^{\text{sam}} \sim H_0$. The results are given in Figure 3. The $p$-value is equal to .447, suggesting that there is no indication of within-age-group-transmission effect. This result is plausible since in previous analyses, the age groups are often used as a way of differentiating different susceptibility or infectivity (e.g., Hayakawa et al., 2003), rather than as social groups in which individuals mix together.

There are 345,599 possible orderings of the age group label vector in addition to the one considered above, over which the $p$-value ranged from .391 to .507, suggesting no sensitivity of the test conclusion to the choice of ordering.

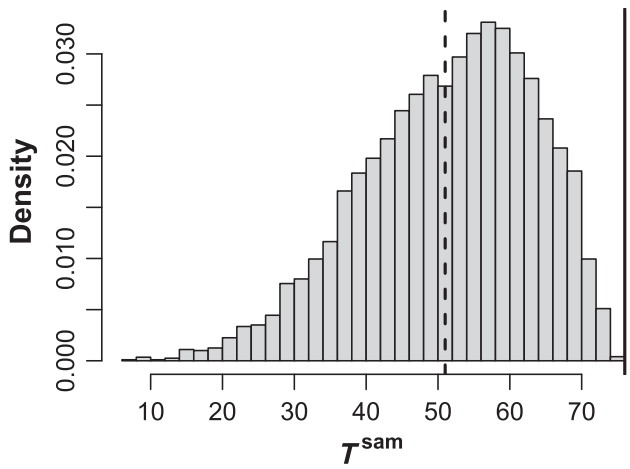


**FIGURE 3** Application of the classical hypothesis test for age group label data on the Tristan da Cunha outbreak data. The plot is the histogram of 10,000 realizations from the sampling distribution of $T^{\text{sam}} \sim H_0$ with the observed value $T^{\text{obs}} = 51$ (dashed line) and the maximum value of $T = 76$ (solid line) imposed. The test $p$-value is equal to .447.

# 6 | ASYMPTOTIC RESULTS

Although the examples of the previous section illustrate that our hypothesis test can be very easily implemented in practice, it is nonetheless of interest to consider its theoretical behavior. We now briefly consider the asymptotic behavior of $T$ as the population size becomes large, under the null hypothesis $H_0$ that the individuals in the outbreak are drawn at random without replacement from the population. In general this is a challenging problem so in what follows we assume that the population consists of groups of equal size. Our main focus will be on the case in which the population size $C$ grows linearly with the outbreak size $n$, the reason being that the latter is known to be $O(C)$ in the event of a major outbreak (Andersson & Britton, 2000, chap. 4), although we briefly consider other regimes. We write $T = T_n$ in the obvious manner. All proofs can be found in Appendix S5.

## 6.1 | Degeneracy of $T$

It is possible for the distribution of $T$ to be a point mass, which is clearly not desirable for hypothesis testing. In what follows, we establish conditions under which this degeneracy can occur. In the case of most practical concern in which $C = O(n)$, the probability of degeneracy is asymptotically zero as $n \to \infty$. In practical terms, our results show the importance of a suitably large outbreak relative to the population size.

Let $\mathcal{G}$ denote the set of groups containing individuals who are infected during the epidemic, so that $T_n = \sum_{m\in\mathcal{G}} s^{(m)}$. For $n = 1, 2, \ldots$ let $\mathcal{E}_n$ denote the event that all of the $n$ individuals in the outbreak belong to different groups. If $\mathcal{E}_n$ occurs then $\nu^{(m)} = 1$ for all $m \in \mathcal{G}$ and $\left\{f_1^{(m)} : m \in \mathcal{G}\right\} = \{1, 2, \ldots, n\}$. Thus

$$T_n = \sum_{m\in\mathcal{G}} (n - f_1^{(m)}) = n^2 - \sum_{j=1}^{n} j = n(n-1)/2,$$

so that $T_n$ collapses to a point mass. Conversely if $\mathcal{E}_n$ does not occur then there exists a $k$ such that $\nu^{(k)} \geq 2$ and

$$s^{(k)} = f_{\nu^{(k)}}^{(k)} - f_1^{(k)} - (\nu^{(k)} - 1) \leq n - f_1^{(k)} - 1 < n - f_1^{(k)},$$

and since $s^{(m)} \leq n - f_1^{(m)}$ for all $m \neq k$, $m \in \mathcal{G}$ it follows that

$$T_n = \sum_{m\in\mathcal{G}} s^{(m)} < \sum_{m\in\mathcal{G}} (n - f_1^{(m)}) < n(n-1)/2,$$

the final strict inequality arising from the fact that there are fewer than $n - 1$ groups in $\mathcal{G}$ apart from $k$. We have thus established that $T_n = n(n-1)/2$ if and only if $\mathcal{E}_n$ occurs.

**Lemma 1.** *If $P(\mathcal{E}_n) \to 1$ as $n \to \infty$ then $2T_n/n(n-1)$ converges in probability to 1.*

It remains to establish the behavior of $P(\mathcal{E}_n)$. Let $f(n) \sim g(n)$ denote that $f(n)/g(n) \to 1$ as $n \to \infty$.

**Lemma 2.** *Suppose that the population consists of $l$ groups of size $m \geq 2$ so that $C = ml$. If $C \sim \theta n^\beta$ with $\theta > 0$ and $\beta \geq 1$ then*

$$\lim_{n\to\infty} P(\mathcal{E}_n) = \begin{cases} 0 & \text{if } 1 \leq \beta < 2, \\ \exp(-(m-1)/2\theta) & \text{if } \beta = 2, \\ 1 & \text{if } \beta > 2. \end{cases}$$

We conjecture that corresponding results apply if the assumption of equal-sized groups is relaxed. Such results appear harder to prove, as explained in Appendix S5.

## 6.2 | A central limit theorem for groups of size two

Since $T_n = \sum_{m=1}^{l} s^{(m)}$, it is natural to ask if $T_n$ can be asymptotically Gaussian. Under the null hypothesis, the $s^{(m)}$ are random variables with a non-trivial dependence structure, and also $n$-dependent. They are also exchangeable, but we were unable to find any results in the literature on central limit theorems that could be applied to $T_n$, the main difficulty being that such results usually consider a sum that is made up of a subset of exchangeable random variables $Y_1,\ldots,Y_n$ rather than all $n$ of them. However, in the special case in which the population consists of groups of size two, it is possible to write $T_n$ as a doubly indexed permutation statistic, and then apply a result from Barbour and Eagleson (1986) to yield a central limit theorem, as we now describe.

### 6.2.1 | Representing $T$ as a doubly indexed permutation statistic

Suppose that the population consists of groups of size two. We can construct $T_n$ as follows. The individuals in the population can be represented by the vector $v = (1, 2,\ldots, C)$, where $C = 2l$. We assume that the groups consist of pairs of individuals $(1, 2), (3, 4),\ldots, (C - 1, C)$.

Consider a permutation $\pi$ of $v$ drawn uniformly at random from all possible permutations. Then the first $n$ elements of $\pi(v)$ are an ordered sample, chosen uniformly at random without replacement, from the individuals in the population. In particular, this sample is identically distributed to the individuals sampled under the null hypothesis. Note that individual $i$ appears in the sample if and only if $1 \le \pi(i) \le n$.

If a group has no individuals in this sample, it makes no contribution to $T_n$. If a group has exactly one individual in the sample, $i$ say, it contributes $n - \pi(i)$ to $T_n$. If a group has two individuals in the sample, $i$ and $j$ say, then it contributes $|\pi(i) - \pi(j)| - 1$ to $T_n$.

We encode this construction by summing over all possible ordered pairs of individuals $(i, j)$ in the population. If $i$ and $j$ are in different groups then $(i, j)$ and $(j, i)$ make no contribution to $T_n$. Conversely if $i$ and $j$ are in the same group then both $(i, j)$ and $(j, i)$ make the same contribution to $T_n$, which we take account of by dividing each separate contribution by 2. Specifically, we define

$$T_n = \sum_{\substack{i,j=1 \\ i \neq j}}^{C} d_{ij}\, y_{\pi(i)\pi(j)}, \tag{3}$$

where for $1 \le i, j \le C$,

$$d_{ij} = d_{ji} = \begin{cases} \frac{1}{2} & \text{if } (i,j) \in \{(1,2),(2,1),(3,4),(4,3),\ldots,(C,C-1)\}, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

and

$$y_{ij} = y_{ji} = \begin{cases} |i-j| - 1 & \text{if } 1 \le i,j, \le n \text{ and } i \neq j, \\ n - \min(i,j) & \text{if } 1 \le \min(i,j) \le n < \max(i,j), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

By way of example suppose that $C = 8$, so the population vector of individuals is $v = (1, 2, \ldots, 8)$, that $n = 4$, and $\pi(v) = (4, 1, 3, 7, 6, 2, 5, 8)$. Then the sample of $n$ is $(4, 1, 3, 7)$, consisting of two individuals from the same group (3,4) and two individuals from two other groups $(1, 7)$. Individuals 3 and 4 contribute 1 to $T_n$ since there is one individual (namely 1) between them. Individual 1 is in position 2 and so contributes $n - 2 = 2$ to $T_n$. Finally individual 7 is in position 4 and so contributes $n - 4 = 0$ to $T_n$. Thus $T_n = 1 + 2 = 3$.

To see that this agrees with (3), note that from (4) we need only consider the pairs $(1, 2)$, $(2, 1)$, $(3, 4)$, $(4, 3)$ … $(7, 8)$, $(8, 7)$, all of which have $d_{ij} = 1/2$. For $(1, 2)$ and $(2, 1)$, $\pi(1) = 2$ and $\pi(2) = 6$ so that $y_{\pi(1)\pi(2)} = y_{\pi(2)\pi(1)} = 4 - 2 = 2$. Similarly $\pi(3) = 3$ and $\pi(4) = 1$ so $y_{\pi(3)\pi(4)} = y_{\pi(4)\pi(3)} = |3 - 1| - 1 = 1$, while $\pi(5) = 7$ and $\pi(6) = 5$ and so $y_{\pi(5)\pi(6)} = y_{\pi(6)\pi(5)} = 0$. Finally $\pi(7) = 4$ and $\pi(8) = 8$ yielding $y_{\pi(7)\pi(8)} = y_{\pi(8)\pi(7)} = 4 - 4 = 0$, and we thus obtain $T_n = (1/2)(2 + 2 + 1 + 1 + 0 + 0 + 0 + 0) = 3$ as before.

### 6.2.2 | Central limit theorem

Statistics of the form (3) are known as doubly-indexed permutation statistics, and a number of papers give sufficient conditions under which such statistics are asymptotically Gaussian, as described in Zhao et al. (1997). Only one of these results appears to be applicable to our setting, namely Thm 2. of Barbour and Eagleson (1986), which can be used to obtain the following result.

> **Theorem 1.** *Suppose that the population consists of groups of size two and that* $C = C(n) \sim \theta n$, *where* $\theta \geq 1$. *Define* $s_n^2 = \mathrm{var}(T_n)$ *and* $W_n = (T_n - E[T_n])/s_n$. *Let* $\mathcal{L}(W_n)$ *and* $\mathcal{N}$ *denote respectively the probability distribution of* $W_n$ *and of a standard Gaussian random variable. Then as* $n \to \infty$, $\mathcal{D}(\mathcal{L}(W_n), \mathcal{N}) \to 0$ *where* $\mathcal{D}$ *denotes the distance*
>
> $$\mathcal{D}(F, G) = \sup_{h \in C_1} \left\{ \left| \int h \, dF - \int h \, dG \right| (\|h\|_1)^{-1} \right\},$$
>
> *and* $\|h\|_1 = \sup_x |h(x)| + \sup_x |h'(x)|$.

Expressions for $E[T_n]$ and $s_n^2$ can be found in the proof of Theorem 1 in Appendix S4, along with some numerical illustrations.

It is possible that a central limit theorem still holds if $C \sim \theta n^\beta$ for $1 < \beta < 2$, but the method of proof used for the $\beta = 1$ case appears to fail. Also it seems likely that a central limit theorem will also hold for a population with arbitrary group structure, although deriving such a result would require a different approach to that we have adopted.

## 7 | DISCUSSION

### 7.1 | Limitations, extensions, and general remarks

Event time outbreak data are often daily or weekly, a consequence of which is that there may be multiple orderings of the group label vector. In the two real outbreak examples considered in this paper, we calculated the $p$-value for all possible orderings. However, this approach might not be

feasible if the number of orderings is too large. A possible solution is to calculate the $p$-value for a random sample of these orderings, as opposed to calculating it for all. It may also be possible to identify particular orderings that are as extreme as possible in the sense of clearly favoring $H_0$ or not, and focus attention on these.

An important property of $T$ is that it is a statistic of the data only. In this sense our test is a nonparametric test. Thus testing $H_0$ does not assess the plausibility of a specific set of parameters of a proposed model which has $H_0$ as an assumption, but rather provides a more generic test for the plausibility of the family of models that share $H_0$ as an assumption, under any set of parameters. In practice, outbreak data analyses are usually conducted with one or more transmission models, but our test could be used in conjunction with any model that assumes the population of at-risk individuals is partitioned in some way. Furthermore, the test implementation involves no parameter estimation or simulations of a model, procedures that are typically computationally intensive to perform. Thus the test is computationally extremely cheap, with a sample size of 10,000 from $T^{\text{sam}} \sim H_0$ typically taking less than a second of computer time to obtain.

In developing our test, we have assumed that an epidemic outbreak has finished. In principle it is also possible to apply the test to an ongoing outbreak in which $n$ events have been observed to date. The extent to which this is sensible will depend on the situation at hand, two potential problems being group-varying reporting delays in the collection of data and the fact that infected individuals may not have yet been detected. Nevertheless the test could be helpful in giving some indication of between-group or within-group transmission in the early stages of an emerging disease.

Our test requires outbreak data at an individual level. There are many studies in the literature where outbreaks of both human and animal disease have taken place in a population partitioned into groups, perhaps by type of individual or by a spatial region, and the available data consist of event times that are proxies of the infection times such as symptom-onset times, case-detection times, or lab test results (see, e.g., Auranen et al., 2000; Black et al., 2017; Cauchemez et al., 2009; de Greeff et al., 2012; McKinley et al., 2020; Neal & Roberts, 2004; Seymour et al., 2021). Under the assumption that these event times are able to convey information about transmission itself our test could also be implemented in these settings.

Although our focus has been on epidemiological applications, our test is also potentially applicable to the study of other kinds of transmission. Examples include the spread of information, awareness, rumors, social media content or financial crises among populations that can be partitioned into relevant groups, for example, of individuals or countries. Potential groupings include the age of individuals, geographical locations, or other demographic categories.

## 7.2 | Related literature

The problem of assessing transmission homogeneity assumptions in epidemic models using classical hypothesis testing is not new in the literature. For example, the tests in Britton (1997a, 1997b, 1997c, 1997d) all conduct this type of assessment. The general features of these tests are as follows. As in our case the population is partitioned into groups, and a particular epidemic model is considered, which assumes global infection contacts at rate $\lambda > 0$, and local infection contacts at rate $\delta \geq 0$. The model is formulated so that when $\delta = 0$ it reduces to a homogeneously mixing model. A test for additional within-group transmission is conducted by testing the null

hypothesis $\delta = 0$ against the alternative hypothesis $\delta > 0$. This is achieved using the likelihood ratio (LR) statistic $L(\delta)/L(0)$, where $L$ is the likelihood function. Calculation of the observed value of the LR test statistic and derivation of its sampling distribution under the null hypothesis are based on a Taylor series approximation of $L(\delta)$ around $\delta = 0$ and on asymptotic results when the number of groups is large. An advantage of our test, and a fundamental difference, is that these LR tests heavily rely on approximations (numeric and asymptotic), whereas our test does not. Despite the fundamental differences, we have compared our test against the test of Britton (1997c). The results of this comparison are presented in Appendix S4 which reveal that whilst both tests perform equally well in the case of strong within-group-transmission and large population size $N$, ours has greater power than Britton's when there is very mild within-group-transmission (and hence harder to detect) and $N$ is small.

Another test that can be used in our setting, which has the same null hypothesis $H_0$ as our test, is a Chi-squared goodness-of-fit test, based on the number of individuals infected in each group, that is, on group label counts (a similar approach is described in, Becker, 1989, pp. 21-22). This test is based on the fact that, under $H_0$, and given the total number of events $n$, the group label counts follow a Multinomial distribution. The main difference between this test and ours is that the former only uses final outcome information whereas the latter uses temporal information which can be crucial in detecting within-group-transmission effect. For example, if an epidemic infected almost all individuals in group 1, then almost all individuals in group 2, and so on then the time-ordering would strongly suggest within-group-transmission, but just looking at the final numbers infected in each group would not do so.

Besides classical testing, there are other ways possible to assess the assumption of homogeneity of disease transmission in populations having a group structure. For example, in a Bayesian framework, it is possible to look at the posterior distribution of the number of local infections and compare it to that of global infections, as described in Alharthi (2016). The idea of this approach is that the higher the number of local infections suggested by the model, the more the evidence of within-group-transmission effect in the observed data. Although this is a natural way of conducing this assessment, it requires model fitting via Markov chain Monte Carlo methods or similar numerical methods, and is far more computationally expensive to implement than our group label test.

## ACKNOWLEDGMENTS

We are grateful to two anonymous reviewers and an Associate Editor whose suggestions have greatly improved the paper. The first author is grateful to the University of Nottingham for providing financial support during his PhD studies.

## ORCID

*Theodore Kypraios* https://orcid.org/0000-0002-6190-4762

## REFERENCES

Alharthi, M. (2016). *Bayesian model assessment for stochastic epidemic models* [PhD thesis]. University of Nottingham.

Andersson, H., & Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Springer.

Auranen, K., Arjas, E., Leino, T., & Takala, A. K. (2000). Transmission of pneumococcal carriage in families: A latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association*, *95*(452), 1044–1053.

Bailey, N. T. J. (1975). *The mathematical theory of infectious diseases and its applications* (2nd ed.). Griffin.

Ball, F. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability*, *18*(2), 289–310.

Ball, F., Mollison, D., & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability*, *7*(1), 46–89.

Barbour, A., & Eagleson, G. (1986). Random association of symmetric arrays. *Stochastic Analysis and Applications*, *4*(3), 239–281.

Becker, N. G. (1989). *Analysis of infectious disease data*. Chapman and Hall.

Becker, N. G., & Hopper, J. L. (1983). Assessing the heterogeneity of disease spread through a community. *American Journal of Epidemiology*, *117*(3), 362–374.

Black, A. J., Geard, N., McCaw, J. M., McVernon, J., & Ross, J. V. (2017). Characterising pandemic severity and transmissibility from data collected during first few hundred studies. *Epidemics*, *19*, 61–73.

Boys, R. J., & Giles, P. R. (2007). Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *Journal of Mathematical Biology*, *55*(2), 223–247.

Britton, T. (1997a). Limit theorems and tests for within family clustering in epidemic models. *Communications in Statistics - Theory and Methods*, *26*(4), 953–976.

Britton, T. (1997b). A test of homogeneity versus a specified heterogeneity in an epidemic model. *Mathematical Biosciences*, *141*(2), 79–99.

Britton, T. (1997c). A test to detect within-family infectivity when the whole epidemic process is observed. *Scandinavian Journal of Statistics*, *24*(3), 315–330.

Britton, T. (1997d). Tests to detect clustering of infected individuals within families. *Biometrics*, *53*(1), 98.

Cauchemez, S., Donnelly, C. A., Reed, C., Ghani, A. C., Fraser, C., Kent, C. K., Finelli, L., & Ferguson, N. M. (2009). Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. *New England Journal of Medicine*, *361*(27), 2619–2627.

Clancy, D., & O'Neill, P. D. (2008). Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, *3*(4), 737–757.

de Greeff, S., de Melker, H., Schellekens, A. W. J., Mooi, F., & Michiela van Boven, M. (2012). Estimation of household transmission rates of pertussis and the effect of cocooning vaccination strategies on infant pertussis. *Epidemiology*, *23*, 852–860.

Eichner, M., & Dietz, K. (2003). Transmission potential of smallpox: Estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, *158*(2), 110–117.

Hayakawa, Y., O'Neill, P. D., Upton, D., & Yip, P. S. (2003). Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. *Australian & New Zealand Journal of Statistics*, *45*(4), 491–502.

Kypraios, T., Neal, P., & Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Mathematical Biosciences*, *287*, 42–53.

McKinley, T. J., Neal, P., Spencer, S. E. F., Conlan, A. J. K., & Tiley, L. (2020). Efficient Bayesian model choice for partially observed processes: With application to an experimental transmission study of an infectious disease. *Bayesian Analysis*, *15*(3), 839–870.

Neal, P., & Roberts, G. (2005). A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, *15*(4), 315–327.

Neal, P. J., & Roberts, G. O. (2004). Statistical inference and model selection for the 1861 hagelloch measles epidemic. *Biostatistics*, *5*(2), 249–261.

O'Neill, P. D., & Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(1), 121–129.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Seymour, R. G., Kypraios, T., O'Neill, P. D., & Hagenaars, T. J. (2021). A Bayesian nonparametric analysis of the 2003 outbreak of highly pathogenic avian influenza in the Netherlands. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *70*(5), 1323–1343.

Stockdale, J. E., Kypraios, T., & O'Neill, P. D. (2017). Modelling and Bayesian analysis of the Abakaliki smallpox data. *Epidemics*, *19*, 13–23.

Thompson, D., & Foege, W. (1968). *Faith tabernacle smallpox epidemic*. World Health Organization.

Xiang, F., & Neal, P. (2014). Efficient MCMC for temporal epidemics via parameter reduction. *Computational Statistics & Data Analysis*, *80*, 240–250.

Zhao, L., Bai, Z., Chao, C.-C., & Liang, W.-Q. (1997). Error bound in a central limit theorem of double-indexed permutation statistics. *The Annals of Statistics*, *25*(5), 2210–2227.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Aristotelous, G., Kypraios, T., & O'Neill, P. D. (2024). A classical hypothesis test for assessing the homogeneity of disease transmission in stochastic epidemic models. *Scandinavian Journal of Statistics*, 1–19. https://doi.org/10.1111/sjos.12743