Predicting COVID-19 cases with unknown homogeneous or heterogeneous resistance to infectivity

Ramalingam Shanmugam,^a Gerald Ledlow,^b and Karan P. Singh^c

^a Honorary Professor of International Studies

School of Health Administration

Texas State University, San Marcos, TX 78666, USA.

Email: rs25@txstate.edu

ORCID: [0000-0002-3388-1014](https://orcid.org/0000-0002-3388-1014)

^b Professor and Dean, Department of Healthcare Policy, Economics and Management

School of Community and Rural Health

The University of Texas Health Sciences Center at Tyler

Tyler, TX 11937, USA

Email: gerald.ledlow@uthct.edu

^c Professor and Chair, Department of Epidemiology and Biostatistics

School of Community and Rural Health

The University of Texas Health Sciences Center at Tyler

Tyler, TX 11937, USA

Email: karan.singh@uthct.edu

ORCID: [0000-0002-1841-846X](https://orcid.org/0000-0002-1841-846X)

Abstract

This article constructs a restricted infection rate inverse binomial-based approach to predict COVID-19 cases after a family gathering. The traditional inverse binomial (IB) model is unqualified to match the reality of COVID-19, because the data contradicts the model's requirement that variance should be greater than expected value. A refined version of the IB model is a necessity to predict COVID-19 cases after family gatherings. Our refined version of an IB model is more appropriate and versatile, as it accommodates all potential data scenarios: equal, lesser, or greater variance than expected value.

Application of the approach is based on a restricted infectivity rate and methodology on Fan et al.'s COVID-19 data, which exhibits two clusters of infectivity. Cluster 1 has a smaller number of primary cases and exhibits larger variance than the expected cases with a negative correlation of 28%, implying that the number of secondary cases is lesser when the number of primary cases increases and vice versa. The traditional inverse binomial (IB) model is appropriate for Cluster 1. The probability of contracting COVID-19 is estimated to be 0.13 among the primary, but is 0.75 among the secondary in Cluster 1, with a wider gap. Conversely, Cluster 2, exhibits smaller variance than the expected cases with a correlation of 79%, implying the number of primary and secondary cases increase or decrease together. Cluster 2 disqualifies the traditional IB model and demands its refined version. Probability of contracting COVID-19 is estimated to be 0.74 among the primary, but is 0.72 among the secondary in Cluster 2, with a narrower gap.

The model's ability to estimate the community's health system memory for future policies to be developed is an asset of this approach. The current hazard level to be infected with COVID-19 among the primary and secondary groups are estimable and interpretable.

46 **Keywords:** Restricted infectivity; informatics; domino effect; risk analysis; hazard rate;
 47 community's memory; odds ratio; primary group; secondary group; Poisson model; inverse
 48 binomial model; homogeneity; heterogeneity; family union.

49
 50 **Short title:** Predicting COVID-19 cases with unknown resistance

52

Author Summary

Current statistical models are not able to accurately predict disease infection spread in the COVID-19 pandemic. We have applied a widely-used inverse binomial method to predict rates of infection after small gatherings, going from primary (original) cases to secondary (later) cases after family gatherings or social events, using the data from the Wuhan and Gansu provinces in China, where the virus first spread. The advantages of the proposed approach include that the model's ability to estimate the community's health system memory for future policies to be developed, as such policies might reduce COVID's spread if not its control. In our approach, as demonstrated, the current hazard level of becoming infected with COVID-19 and the odds of contracting COVID-19 among the primary in comparison to the secondary groups are estimable and interpretable. We hope the proposed approach will be used in future epidemics.

Introduction

COVID-19 is the third-leading cause of death in 2020 in the USA, Belgium, France, Sweden, and the UK, behind only heart disease and cancer (www.kff.org). Predicting a pandemic like COVID-19 is challenging (reasons in Ioannidis et al., 2020). Major practical reasons that have been cited include poor data, insensitive parameter estimates, and imprecise epidemiologic features. A discussion of an underlying model for the data is missing from this list. What is a model? The model is an abstraction of reality; the better the model, the better it represents reality. The occurrence and spread of COVID-19 are too complex in reality to be well matched by any known model in the literature. It is not surprising that the efforts to predict COVID-19 result in a failure, due to utilizing a wrong model for the data. To avoid failure, we must start by refining the model to suit the complexities that exist in the dynamic nature of COVID-19. That is exactly the research theme of this article. Of course, the main objective of any data collection is to predict future incidences as accurately as possible in order to be prepared for any emergencies and contingencies.

The professionals occasionally hear and/or argue that all models are wrong, but some are useful (Field, 2015; Box, 1976). Recently, Shanmugam (2020, a) presented a probabilistic approach to capturing the impact of healthcare efforts on the prevalence rate of COVID-19's infectivity, hospitalization, recovery, and mortality rates in the USA. Several non-intuitive findings including the existence of imbalance, different vulnerabilities, and risk reduction were noticed. In Shanmugam (2020, b), the number of COVID-19 cases confirmed, recovered/cured, and fatalities across thirty-two of India's states/territories, as of May 1, 2020 were modelled and analyzed. In the end, the attained administrative efficiency by the government was scrutinized.

This knowledge leads to valuable lessons for adaptation for use in any future pandemics like COVID-19.

Hence, via modeling by trial and error, the professionals attempted to catch up and unravel the mysterious nature of the pandemic. This article is an attestation of such a hubristic endeavor to predict COVID-19 infectivity with a refined version of the traditional inverse binomial (IB) model. The traditional IB model possesses a unique property that the variance is larger than its expected value (Stuart and Ord, 2015 for details) and it limits its suitability for COVID-19 data under such variance. A refined version of the IB model is necessary to accommodate data under equal and over variance scenarios as it occurs in COVID-19 data.

Yet, the issue of under variance seems to have not received enough attention compared to over variance in the literature (Barmalzan, 2019; Tin 2008). The Poisson model requires the equality of expected value and variance (Stuart and Ord, 2015). To deal with a deviation from such a requirement in Poisson data, Conwell and Maxwell (1962) initiated a mathematically complex approach by raising a function of the observables to an unknown versatile parameter τ . Conway and Maxwell applied their approach to model the queuing systems with state-dependent service rates.

We now return to discuss a refined version of the IB model for COVID-19 data. At a fixed incidence rate, the number of COVID-19 cases is, of course, a Poisson type random variable. When COVID-19's incidence rate is stochastically changing due to a variety of reasons as a gamma probability pattern, the convolution of Poisson and gamma probability structures results in the IB model for an occurrence of a random number of COVID-19 cases (Shanmugam and Chattamvelli, 2015 for details on the convolution). However, in this article, a refined and viable alternative to dealing with equal/under/over variance COVID-19 data, we start at its *jump*

start probability. Shanmugam and Radhakrishnan (2011) introduced a concept of jump start probability in health/medical data analysis. The method in this article is simple, easy, and versatile, but originates from the jump probability. The unsteady nature of the variance is recognized here as the *heterogeneity* level in the gatherings of family with some members being carriers of the COVID-19 virus. Heterogeneity causes the prediction of COVID-19 incidence rates to be haphazard, if not uncanny.

COVID-19 was first noticed in Wuhan Province in the central part of China. Those with COVID-19 symptoms leaving Wuhan to participate in family gatherings in Gansu Province are recognized as the primary cases (Fan et al., 2020). Those at the family gatherings in Gansu Province who were infected by the primary cases are recognized as secondary cases. Is there a significant difference in the virality among the secondary cases in comparison to the primary cases? This research question is answered in the second part of this article. For this purpose, new concepts of infectivity that restricted IB modeling processes are defined, and they are utilized to derive analytic expressions with an intention to better predict the future COVID-19 infection rates and cases. The data are borrowed from Fan et al., 2020.

Uneven COVID-19 Jump Start Proportion

To ease the comprehension of the derivation of an infectivity rate restricted version of the inverse binomial (IB) which adds on under and equal variance properties in addition to preserving already existing over variance properties, we start with the COVID-19 scenarios. Suppose there is an unknown risk level, $0 < \theta < 1$ for a person to contract COVID-19 at a family gathering. Also, assume that there exists an unknown restriction level, $\tau > 1$ on infectivity due to persons who may have a strong immunity and/or have undertaken strict preventive measures

such as physical social distancing, wearing face coverings, washing hands with soap/sanitizer frequently, etc. Consequently, the original risk level, θ changes to a new risk level $\frac{\theta}{(1-\theta)^{(\tau-1)}}$, where $\tau > 1$. Here, the non-negative parameter τ is to indicate the heterogeneous resistance level to COVID-19's infectivity potential.

When $\tau > 1$, it is indicative of heterogeneously resistant to the COVID-19's virus among the participants. Notice that the new risk level validates the requirement that $0 < \frac{\theta}{(1-\theta)^{(\tau-1)}} < 1$,

which results in a restricted infection rate: $0 < \theta < \frac{1}{\tau} < 1$ with $\tau \geq 1$. In the case of $\tau > 1$, any

participant's chance of being safe from contracting COVID-19 is $0 < 1 - \frac{\theta}{(1-\theta)^{(\tau-1)}} < 1$ after the

family gathering, while her/his chance of contracting COVID-19 is $0 < \frac{\theta}{(1-\theta)^{(\tau-1)}} < 1$ after the

family gathering as it happened in Gansu Province, China. Various above-mentioned preventive measures might not offer absolute safeguards from the COVID-19 infection potential all the time. Consequently, the odds of contracting COVID-19 exist whether the participants are homogeneously immune (that is, $\tau = 1$) or heterogeneously resistant (that is, $\tau > 1$) to contract the

COVID-19 virus. Hence, the odds for being safe after the family gathering is $\{\frac{(1-\theta)^{(\tau-1)}}{\theta} - 1\}$

among the heterogeneously resistant participants as they identified by the infection rate restricted inverse binomial (IRRIB) model.

In the case of $\tau = 1$, as the infectivity is *unrestricted* in a sense: $0 < \theta < 1$, any participant's chance of being *safe* from contracting COVID-19 is $0 < 1 - \theta < 1$ after the family gathering, while her/his chance of contracting COVID-19 is $0 < \theta < 1$ after the family gathering as it happened in

Gansu Province, China. Hence, the odds for being safe after the family union is $\frac{(1-\theta)}{\theta}$ among the homogeneously immune participants as they identified by the IB model. The traditional IB model is an extreme special case of the IRRIB model when $\tau = 1$, because the denominator $(1-\theta)^{(\tau-1)} = 1$ as a baseline value. In other words, when the family gathering consists of attendees having homogeneous immunity to COVID-19 (that is, $\tau = 1$), using the traditional IB model is meaningful.

Otherwise (that is, with a family gathering in which the attendees have heterogeneous resistance to COVID-19), the involvement of $\tau > 1$ is a necessity in modeling to predict future COVID-19 cases. Hence, the modeling approach based on IRRIB is versatile to predict incidence of COVID-19 in all three (equal, under, and over variance) scenarios mentioned above.

We now discuss the odds of contracting COVID-19 among attendees having homogeneous resistance to COVID-19 versus heterogeneous resistance to COVID-19 in the gathering. Amid homogeneous resistance to COVID-19 family members in the gathering, the factor $\frac{(1-\theta)}{\theta}$ is recognized as the odds for not contracting COVID-19, and it is a safe situation. The odds for a safe situation in the IRRIB process (synonymous to a situation in which the family gathering involves attendees having heterogeneous resistance to COVID-19) against contracting COVID-19 is

$$Odds_{\tau \neq 1}^{safe} = \frac{\Pr(Y=0)}{\Pr(Y \geq 1)} = \left[\left\{ 1 - \frac{\theta_{\tau \neq 1}}{(1-\theta_{\tau \neq 1})^{(\tau-1)}} \right\}^{-m\tau} - 1 \right]^{-1}; 0 < \theta < \frac{1}{\tau} < 1 \quad (1)$$

in comparison to the $Odds_{\tau=1}^{safe} = [(1-\theta_{\tau=1})^{-m} - 1]^{-1}; 0 < \theta < 1$ for safe in the IB process (synonymous to a situation in which the family gathering involves attendees having homogeneous resistance to COVID-19). The odds ratio in family gatherings in which the

attendees having heterogeneous resistance to COVID-19 against having homogeneous resistance to COVID-19 is

$$OR_{\frac{\tau \neq 1}{\tau = 1}} = \frac{Odds_{\tau \neq 1}}{Odds_{\tau = 1}} = \frac{[(1 - \theta_{\tau \neq 1})^{-m} - 1]}{[1 - \frac{\theta_{\tau \neq 1}}{(1 - \theta_{\tau \neq 1})^{(\tau-1)}}]^{-m\tau} - 1}; 0 < \theta < \frac{1}{\tau} < 1. \quad (2)$$

Jump Rate-Incidence of COVID-19 Cases

We now move on to discuss the jump rate in the incidence of COVID-19 cases.

Furthermore, let us assume that there are $m \geq 1$ family gatherings with a group of participants with heterogeneous resistance to COVID-19, $\tau \geq 1$ at the family gathering. Let Y be a random number of COVID-19 cases emerging out of the m family gatherings. Notice that the probability

for a single new COVID-19 case to arise in any gathering is $\Pr(Y = 1) = \frac{\theta}{(1 - \theta)^{(\tau-1)}}$

in comparison to the probability of an infection free situation, $\Pr(Y = 0) = 1 - \frac{\theta}{(1 - \theta)^{(\tau-1)}}$.

The jump rate (Shanmugam and Radhakrishnan, 2011 for its definition and details) from a COVID-19 free situation to the pandemic is

$$jumpRate_{(\theta, \tau)} = \frac{\Pr(Y = 1)}{\Pr(Y = 0)} = \left(\frac{(1 - \theta)^{(\tau-1)}}{\theta} - 1 \right)^{-1} \quad (3)$$

if the family gatherings involve attendees having heterogeneous resistance to COVID-19. When the gathering consists of attendees having homogeneously immunity to COVID-19 (i.e., $\tau = 1$), the jump rate from COVID-19 free status to contracting COVID-19 is simply

$jumpRate_{(\theta, \tau=1)} = \frac{\theta}{(1 - \theta)}$ and it pertains to the ideal, traditional IB model scenario for any

communicable disease but not necessarily the highly infectious and treacherous COVID-19

scenario. Now, consider a domino effect of this jump rate, especially in the COVID-19 situation.

Their recursive probabilities are connected to the jump rate as follows.

$$\Pr(Y = y) = \left(\frac{m\tau + y - 1}{y} \right) \left\{ \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\} \Pr(Y = y-1); y = 0, 1, 2, \dots; 0 < \theta < \frac{1}{\tau} < 1. \quad (4)$$

That is,

$$\Pr(Y = y) \propto \frac{\Gamma(m\tau + y)}{y! \Gamma(m\tau)} \left\{ \frac{\text{jumpRate}_{(\theta, \tau)}}{1 + \text{jumpRate}_{(\theta, \tau)}} \right\}^y; y = 0, 1, 2, \dots; 0 < \theta < \frac{1}{\tau} < 1 \quad (5)$$

with an appropriate normalizer

$$N(\theta, \tau) = \left\{ 1 - \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\}^{m\tau} = \left\{ 1 + \text{jumpRate}_{(\theta, \tau)} \right\}^{-m\tau} \quad (6)$$

(its dynamic in S1 Figure 1). The dynamic nature of

$$\frac{(1-\theta)^{(\tau-1)}}{\theta} = 1 + \frac{1}{\text{jumpRate}_{(\theta, \tau)}}$$

(7)

is noticed in S2 Fig 2. By imposing $\tau = 1$ or $m = 1$, the traditional IB model which is often

employed to deal with a group homogeneously immune to COVID-19 after the family gathering.

On the contrary, together when $m\tau = 1$, they reduce to what we wish to call the infection rate

restricted geometric (IRRG) process. That is,

$$\Pr(Y = y | m\tau = 1) = \left\{ 1 - \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\} \frac{1}{y!} \left\{ \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\}^y; y = 0, 1, 2, \dots; 0 < \theta < \frac{1}{\tau} < 1. \quad (8)$$

S1 Figure 1. Probability of safe, $\left\{ 1 - \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\}$

S2 Figure 2. Risk for COVID-19, $\frac{\theta}{(1-\theta)^{(\tau-1)}}$

However, it is easy to notice that the expected value, $E(Y)$ and variance, $Var(Y)$ of the IRRIB model are, respectively,

$$E(Y) = \sum_{y=0}^{\infty} yN(\theta, \tau) \Pr(Y = y) = m\tau \left\{ 1 + \frac{1}{jumpRate_{(\theta, \tau)}} \right\} \quad (9)$$

and

$$Var(Y) = \left\{ 1 + \frac{1}{jumpRate_{(\theta, \tau)}} \right\} E(Y). \quad (10)$$

Explanations for the sake of predicting future COVID-19 cases are given in Appendix I.

Prediction of COVID-19 Cases After Family Gatherings

To illustrate the main concepts and analytic expressions in Section 2, let us consider the COVID-19 data connecting Gansu Province (32°31'N–42°57'N, 92°13'E–108°46'E), in Northwest China and Wuhan (30.5928° N, 114.3055° E) in Central China. COVID-19 was first noticed on December 31, 2019 in Wuhan, China. Wuhan is connected to Gansu by travel options including airplanes, railroads, interstate buses, and private cars. The COVID-19 virus originated in Wuhan.

The primary and secondary COVID-19 cases are defined as follows. The primary cases refer to those who traveled from Wuhan to Gansu. The secondary COVID-19 cases refer to those who never left Gansu. The secondary COVID-19 cases were the outcome of family gatherings in which the primary and secondary COVID-19 cases mingled together. In other words, the secondary cases might have been infected by the primary cases. It is assumed (Fan et al., 2020) that the COVID-19 virus does not mutate to reduce its virulence in transmission. One then wonders whether the secondary cases exhibit different characteristics from that of primary cases? This question is the research aim of this article.

We want to capture and compare the virality of the COVID-19 incidence rates/cases in the primary versus secondary groups. When we plot them in the same graph, we noticed that there are two clusters in the data (S3 Figure 3). The Table 1 (with gatherings of ten families) contains lesser primary cases, while Table 2 (with gatherings of seven families) contains larger primary cases of COVID-19.

S3 Figure 3. Primary vs. Secondary COVID-19 Cases

Comparing the sample expected value and variance of the primary and secondary cases in Tables 1-2, there exists over (under) variance. Figures 3 is to notice that there are two clusters in the union data. The over (under) variance is synonymous with the null hypothesis statement

$H_o: \tau=1$ (with the alternative statement $H_1: \tau \neq 1$). For both the data, the test score, $\frac{\bar{y}^2}{|s_y^2 - \bar{y}|}$ was

computed and displayed in the Table 1 and Table 2. Utilizing all the derived expressions, the values of $\hat{\theta}_{mle, H_o}$ under

$$H_o: \tau=1, E\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} | H_o\right\}, \sqrt{Var\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} | H_o\right\}}, \quad (11)$$

correlations under $H_o: \tau=1$ in a situation in which the family union consists of attendees having heterogeneous resistance to COVID-19 are calculated and displayed in the Table 1 and Table 2.

The values of $\hat{\tau}_{mle, H_1}$ under $H_1: \tau \neq 1$, $\hat{\theta}_{mle, H_1}$ under

$$H_1: \tau \neq 1, E\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} | H_1\right\}, \sqrt{Var\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} | H_1\right\}}, \quad (12)$$

and correlations under $H_1: \tau \neq 1$ in a situation in which the family union consists of attendees having heterogeneous resistance to COVID-19 are computed and displayed in Tables 1 and 2.

The p-value and the statistical power are calculated and displayed in Tables 1 and 2. A comparison of them reveals that the p-values are significant for the primary cases and secondary cases in Table 1 and in Table 2. With a selection of $\tau = \hat{\tau}_1$ under the alternative hypothesis, the statistical power is calculated and displayed in Tables 1 and 2. The power is excellent for the primary and secondary cases in Table 1 and in Table 2. Next, we express final comments and conclusions with respect to predicting implications of our model and analytic results for this study concerning COVID-19.

Cluster 1: Small Primary Cases

Here, the number of primary COVID-19 cases is in a single digit (lesser) as in Table 1. In this case, the number of gatherings is also less, the data have *over* variance than the expected value in both primary as well as secondary cases, the estimate of contracting COVID-19 in primary and secondary type is 0.78 and 0.14 respectively, the prediction of a future number of primary and secondary COVID-19 cases will be 5 and 20 respectively when the family gathering consists of attendees having *homogeneous resistance to COVID-19*. The p-value for the suitability of the IB process for COVID-19 data in Table 1 are 0.0001 in the case of primary as well as secondary infection. The *odds* for contracting COVID-19 are 0.27 and 6.053 as primary and secondary, respectively if the attendees are homogeneously immune to the COVID-19 virus. The estimate of contracting COVID-19 (in family gatherings in which the attendees have heterogeneous resistance to COVID-19) is 0.37 among primary cases and 0.99 among secondary cases. The (statistical) power of accepting the research hypothesis $H_1: \tau = \hat{\tau}_{mle}$ is 0.76 for the primary and 0.99 for the secondary when $\hat{\tau}_{mle}$ is the true value.

Cluster 2: Large Primary Cases

Here, the number of primary COVID-19 cases is in double digits (larger) as in Table 2. In this case, the number of gatherings is also large, but the data have under variance than the expected value in both primary as well as secondary cases. The risk of contracting the COVID-19 virus in primary and secondary type is 0.25 and 0.28, respectively, the prediction of a future number of primary and secondary COVID-19 cases will be 40 and 56 respectively when the family gathering consists of attendees having homogeneous resistance to COVID-19. The odds of contracting COVID-19 in primary and secondary type is 2.53 and 2.143, respectively, the prediction of a future number of primary and secondary COVID-19 cases will be 40 and 56 respectively when the family gathering consists of attendees having homogeneous resistance to COVID-19. The prediction of a future number of primary and secondary COVID-19 cases will be 54 and 78 respectively when the family gathering consists of attendees having homogeneous resistance to COVID-19. The p-value for the suitability of IB model processes for the COVID-19 data in Table 2 are 0.006 in the case of primary and 0.0000058 in the case of secondary infection.

The odds of contracting COVID-19 are 2.91 and 2.57 among the primary and secondary cases, respectively if the family gathering consists of attendees having homogeneous resistance to COVID-19. The odds for contracting COVID-19 are 2.53 and 2.14 among the primary and secondary cases, respectively if the family gathering consists of attendees having heterogeneous resistance to COVID-19. The estimate of contracting COVID-19 (in family gatherings in which the attendees have heterogeneous resistance to COVID-19) is 0.984 among primary cases and 0.72 among secondary cases. The (statistical) power of accepting the research hypothesis

$H_1: \tau = \hat{\tau}_{me}$ are 0.984 for the primary cases and 0.998 for the secondary cases when $\hat{\tau}_{me}$ is the

true value. The odds for contracting COVID-19 are 100.16 and 140.09 among the primary and secondary cases, respectively in the family gathering whose attendees have heterogeneous resistance to COVID-19.

Discussion

The pandemic called COVID-19 is a challenge not only to healthcare professionals, policy makers, epidemiologists and biostatisticians who try to model data and make successful predictions but also to the general public who faithfully practice what was suggested to them such as social distancing, face covering, and utilizing good and frequent sanitizing practices. Despite all these precautionary efforts, many families desire to get together for occasions and events. Some participants in such gatherings may have originated from a place like Wuhan, China in where the COVID-19 virus was spreading and they are labelled as primary cases (Y_p).

The other participants at a gathering occurring at a location like Gansu Province, China where the COVID-19 virus had not yet appeared could get infected by the primary cases from Wuhan, and they are labelled as secondary cases (Y_s) in the epidemiologic data collection process. An unknown in such gatherings is whether the participants are homogeneously immune to the infectivity or heterogeneously risky to contract the virus. Amid less clarity, an issue of interest might be how best one can predict the number of COVID-19 cases after the family gathering. In other words, a research goal for the analysts of infectious diseases is to address the similarities versus differences between the primary and secondary groups in the family gathering. The research goal appears simple and easy on the surface but is actually very complicated and challenging, as pointed out in a recent article by Ioannidis et al. (2020).

We learned the following from the statistical analysis of COVID-19 data in Table 1 in this article. The spreading of the COVID-19 virus is expedited by family gatherings with attendees having heterogeneous resistance to COVID-19; the change is significantly different from its earlier risk of contracting COVID-19. In a macro sense, the number of secondary COVID-19 cases would have been much less if the number of primary COVID-19 cases was smaller in the first place. When these family gatherings contribute to massive infection incidents, many communities, not to mention nations as a whole, do not have enough resources to treat a deluge of patients. When citizens are locked down in homes without working, the nations' productivity reduces to near zero levels, and the global economy suffers consequently. The absence of vaccinations to prevent the spread of COVID-19 makes the scenario bleaker, although several countries have faithfully committed to social distancing, wearing face coverings and other mitigation regimens. It seems that we all have a long way to go to reach the day in which the virus of COVID-19 is totally controlled and eventually eradicated. To attain this optimum level, professionals need to do more research work with pertinent data on COVID-19.

The advantages of this new approach include the model's ability to estimate the community's health system memory for future policy development, as such policies might reduce the COVID-19 viral spread in an effort to control the pandemic. In our approach, as demonstrated, the current hazard level to become infected with COVID-19 and the odds of contracting COVID-19 among the primary in comparison to the secondary groups are estimable and interpretable. In essence, family gatherings, especially with more vulnerable family members who are aged, have chronic diseases, or have issues of reduced immunity, need to be highly scrutinized by the family members before engaging in such events. Computer mediated communication such as tele/video conferencing should be "pushed" and possibly subsidized by

each governmental level to encourage family interaction but while utilizing safe venues. What was omitted in the itemized reasons in Ioannidis et al. (2020) for not successfully predicting COVID-19 cases is the role of an appropriate underlying model for the data. A model is an abstraction of reality. To rectify this situation, this article has constructed and illustrated a restricted infection rate inverse binomial-based approach to better predict future COVID-19 cases after a family gathering or social event.

Declaration of Interests

The authors have no competing or conflict of interests.

Funding

The authors received no specific funding for this work.

Acknowledgments

The authors thank the Texas State University and School of Community and Rural Health at The University of Texas Health Science Center at Tyler for the conducive educational and research environments for this research work.

References

1. Barmalzan G, Saboori H, and Kosari SA. Modified negative binomial distribution: properties, over variance and under variance. Journal of Statistical Theory and Applications. 2020; 18(4): 343–350.
2. Blumenfeld D. Operations research calculations handbook. Boca Raton: CRC Press; 2010.

3. Box GEP. Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN, editors. Robustness in Statistics. Academic Press; 1976. pp. 201-36.
4. Conway RW and Maxwell WL. A queuing model with state dependent service rates. Journal of Industrial Engineering. 1962;12: 132–136.
5. Fan J, Liu X, Pan W, Douglas MW, Bao S. Epidemiology of Coronavirus Disease in Gansu Province, China. Emerging Infectious Diseases. 2020; 26(6): 1257-1265. Available from: www.cdc.gov/eid
6. Field EH. All models are wrong, but some are useful. Seismological Research Letters. 2015; 86 (2A): 291–293. doi: 10.1785/02201401213.
7. Ioannidis PA, Cripps S, and Tanner MA. Predicting for COVID-19 has failed. International Journal of Predicting. 2020; doi: 10.1016/j.ijpredict.2020.08.004.
8. Khokhlov V. Conditional value-at-risk for elliptical distributions. Evropský časopis Ekonomiky a Managementu. 2016;2(6): 70-79.
9. Neyman J. Optimal asymptotic tests of composite hypotheses. In: Probability and Statistics: The Harold Cramer Volume. New York: Wiley, 1959.
10. Rajan C and Shanmugam R. Discrete distributions in engineering and the applied sciences synthesis lectures on mathematics and statistics. 2020;12: 1-227, Morgan & Claypool Press, 82 Wintersport Ln, Williston, VT 05495, USA.
11. Shanmugam R. Restricted prevalence rates of COVID-19's infectivity, hospitalization, recovery, mortality in the USA and their implications. Journal of Healthcare Informatics Research. 2020 a doi:10.1007/s41666-020-00078-0.

12. Shanmugam R. Probabilistic patterns among coronavirus confirmed, cured and deaths in thirty-two of India's states/territories. International Journal of Ecological Economics and Statistics. 2020 b; Forthcoming.
13. Shanmugam R and Chattamvelli R. Statistics for scientists and engineers. Hoboken: John Wiley Inter-Science Publication; 2015.
14. Shanmugam R and Radhakrishnan R. Infection jump rate reveals over/under variance in count data. International Journal of Data Analysis, and Information Systems. 2011; 3(1): 1-8.
15. Stuart A. and Ord K. Kendall's advanced theory of statistics. Volume 1. London: Oxford University Press; 2015.
16. Tin A. Modeling zero-inflated count data with under variance and over variance. SAS Global Forum, Statistics and Data Analysis 2008; paper 372-402.

S1 Appendix

Table 1. The family union in which the primary cases were in single digit (with over variance)

Family Union	Y_p = Primary COVID-19 Cases	Y_s = Secondary COVID-19 Cases
1	1	9
2	1	26
3	1	27
4	4	14
5	4	46
6	6	10
7	6	11
8	6	12
9	6	13
n= 10	6	23
$\bar{y} = \sum_{i=1}^n y_i / n$	4.1	19.1
$s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$	5.21 Over Variance	134.76 Over Variance
$m = \# \text{ unions}$	41	191
Correlation (Y_p, Y_s)	-0.28	
$\hat{\tau}_{mle}$ (if homogeneously immune to COVID-19)	1	1
$\hat{\theta}_{mle}$ (if homogeneously immune to COVID-19)	0.78	0.14

$\hat{\tau}_{mle}$ (if <i>heterogeneously resistance to COVID-19</i>)	2.27	8.05
$\hat{\theta}_{mle}$ (if <i>heterogeneously resistance to COVID-19</i>)	0.55	0.87
$Odds_{\tau=1}^{safe}$ (if <i>homogeneously immune to COVID-19</i>)	0.27	6.05
$Odds_{\tau \neq 1}^{safe}$ (if <i>heterogeneously resistance to COVID-19</i>)	0.37	0.99
$OR_{\frac{\tau \neq 1}{\tau = 1}}$	1.36	0.16
$E\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_o\right\}$ (if <i>homogeneously immune to COVID-19</i>)	4.56	19.11
$\sqrt{Var\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_o\right\}}$ (if <i>homogeneously immune to COVID-19</i>)	3.59	2.71
$E\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_1\right\}$ (family union consists of attendees having <i>heterogeneous resistance to COVID-19</i>)	10.29	149.54
$\sqrt{Var\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_1\right\}}$ (family union consists of attendees having <i>heterogeneous resistance to COVID-19</i>)	13.09	1081.33
Test score $\frac{\bar{y}^2}{ s_y^2 - \bar{y} }$	15.12	3.15
p Value	0.99	1.85043E-22
Power	0.76	0.99

S2 Appendix

Table 2. The family union in which the primary cases were in double digit (with under variance)

Family Union	Y_p = Primary COVID-19 Cases	Y_s = Secondary COVID-19 Cases
11	29	39
12	29	47
13	36	48
14	36	49
15	36	50
16	36	51
17 (n=7)	36	52
$\bar{y} = \sum_{i=1}^n y_i / n$	34	48
$s_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)$	11.66 Under Variance	18.66 Under Variance
$m = \# \text{ unions}$	238	336
$\text{Corr} (Y_p, Y_s)$	0.79	0.79
$\hat{\tau}_{mle}$ (if homogeneously immune to COVID-19)	1	1
$\hat{\theta}_{mle}$ (if homogeneously immune to COVID-19)	0.25	0.28
$\hat{\tau}_{mle}$ (if heterogeneously resistance to COVID-19)	1.34	1.38
$\hat{\theta}_{mle}$ (if heterogeneously resistance to COVID-19)	0.25	0.28

$Odds_{\tau=1}^{safe}$ (if <i>homogeneously immune to COVID-19</i>)	1.34	1.38
$Odds_{\tau \neq 1}^{safe}$ (if <i>heterogeneously resistance to COVID-19</i>)	2.91	2.57
$OR_{\frac{\tau \neq 1}{\tau=1}}$	2.53	2.14
$E\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_o\right\}$ (if <i>homogeneously immune to COVID-19</i>)	0.87	0.83
$\sqrt{Var\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_o\right\}}$ (if <i>homogeneously immune to COVID-19</i>)	39.69	56.02
$E\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_1\right\}$ (family union consists of attendees having <i>heterogeneous resistance to COVID-19</i>)	10.15	15.70
$\sqrt{Var\left\{\frac{\bar{y}^2}{ s_y^2 - \bar{y} } \middle H_1\right\}}$ (family union consists of attendees having <i>heterogeneous resistance to COVID-19</i>)	53.30	77.80
Test score $\frac{\bar{y}^2}{ s_y^2 - \bar{y} }$	15.08	24.77
p Value	0.01	0.01
Power	0.98	0.99

420

421

422

S3 Appendix

A few explanations are worthwhile here for the sake of predicting the future COVID-19 cases. First, a non-intuitive explanation is that when the $jumpRate_{(\theta, \tau)}$ is smaller, the expected number of future COVID-19 cases is larger and that is why COVID-19 is mysterious. Next, let us now look at other intuitive explanations. When the number m of family unions is greater, the expected number of future COVID-19 cases is greater. In other words, the $jumpRate_{(\theta, \tau)}$ of COVID-19 cases is a factor not only of the predicted number but also as the heterogeneity (that is, variance) level of the COVID-19 incidences. When the $jumpRate_{(\theta, \tau)}$ is larger, the expected number of future COVID-19 cases is closer to its variance, illustrating an equal variance scenario.

When the *heterogeneity resistant level*, τ , to COVID-19 increases more and more from its baseline value, $\tau = 1$ (which is indicative of best homogeneously immune level with respect to COVID-19), the variance, $Var(Y)$ changes in proportion to the expected number, $E(Y)$ of COVID-19 cases. The proportionality, $\frac{(1-\theta)^{(\tau-1)}}{\theta}$ reduces as the *heterogeneously resistant level*, τ increases.

When $\tau = 1$, the future expected value and the variance do also change to

$$E(Y) = \frac{m(1-\theta)}{\theta} = mOdds_{IB}^{safe} \quad (13)$$

and

$$Var(Y) = \frac{m(1-\theta)}{\theta^2} = \frac{E(Y)}{\theta} = E(Y)\{1 + Odds_{IB}^{safe}\} \quad (14)$$

With homogeneously immune participants in the family union, which is indicated by the traditional IB process, where $Odds_{IB}^{safe} = \frac{(1-\theta)}{\theta}$. The predictor must realize that the number m of unions could also increase proportionally the number of COVID-19 cases and consequently might inflate the heterogeneity (that is, more variance). Unlike the traditional IB situation which exhibits only *over* variance, the new IRRIB model in this article portrays additionally *equal* and *under* variance. That is, in the case of *under variance*, note that $(1-\theta)^{(\tau-1)} < \theta$ or equivalently $\{\tau > 1 + Odds_{IB}^{safe}\}$ if *attendees are heterogeneously* resistant to COVID-19 virus.

This is obtained after taking the logarithm on both sides and applying the Taylor's series expansion:

$$\ln \theta = \ln \{1 - (1 - \theta)\} \approx -(1 - \theta); 0 < \theta < 1. \quad (14).$$

Notice that the odds of being safe from COVID-19 among those with homogeneously immune is: $Odds_{IB}^{safe} = \frac{(1-\theta)}{\theta}$. Otherwise (that is, when $\tau < 1 + Odds_{IB}^{safe}$), the IRRIB process portrays only *over variance*. When $\tau = 1 + Odds_{IB}^{safe}$, the IRRIB process echoes *equality of expected value and variance*, countering that the equality of expected value and variance is unique to the Poisson process (Stuart and Ord, 2015). The odds for a *safe* situation in the family union involving attendees having *heterogeneous resistance to COVID-19* against contracting COVID-19 is:

$$Odds_{\tau \neq 1}^{safe} = \frac{\Pr(Y = 0)}{\Pr(Y \geq 1)} = \left| \frac{(1 - \theta_{\tau \neq 1})^{(\tau-1)}}{\theta} - 1 \right|; 0 < \theta < \frac{1}{\tau} < 1 \quad (15)$$

in comparison to the

$$Odds_{\tau=1}^{safe} = [(1 - \theta_{\tau=1})^{-m} - 1]^{-1}; 0 < \theta < 1 \quad (16)$$

for *safe* in the family union involving attendees who are *homogeneously immune to COVID-19*.

The other statistical properties of the IRRIB process are the following. The probability

generating function (that is, $E(z^y)$) is

$$E(z^y) = \sum_{y=0}^{\infty} z^y \Pr(y) = \frac{\theta}{(1-\theta)^{(\tau-1)}(1-z) + z}. \quad (17)$$

The moment generating function (that is, $E(e^{ty}) = \sum_{y=0}^{\infty} e^{ty} \Pr(y)$) is easily seen to be

$$\frac{\theta}{(1-\theta)^{(\tau-1)}(1-e^t) + e^t}. \quad (18)$$

The skewness is a measure of asymmetry. The skewness of IRRIB process is

$$Skewness = \frac{2(1-\theta)^{(\tau-1)} - \theta}{\sqrt{m\tau(1-\theta)^{(\tau-1)}\{(1-\theta)^{(\tau-1)} - \theta\}}}. \quad (19)$$

The kurtosis is a measure of the tail thickness. The excessive kurtosis in IRRIB process is

$$ExcessKurtosis = \frac{6}{m\tau} + \frac{\theta^2}{\{(1-\theta)^{(\tau-1)} - \theta\}}. \quad (20)$$

The Fisher's expected information, $I(\theta, \tau, m)$ quantifies what a randomly selected sample informs about the *infectivity/risk parameter*, θ . That is,

$$I(\theta, \tau, m) = -E\{\partial_{\theta}^2 \ln \Pr(y)\} = \frac{m\tau(1-\theta)^{3(1-\tau)}}{\theta^2\{(1-\theta)^{(1-\tau)} - \theta\}}, \quad (21)$$

where the notation ∂_{θ}^2 is the second derivative with respect to the infectivity parameter, θ . The survival function of the IRRIB process is then

$$survivalFunction_{IRRIB} = \Pr(Y \geq p) = P(F_{(2p, 2m\tau)} \leq \frac{m\tau\{(1-\theta)^{(\tau-1)} - \theta\}}{\theta}); 0 < \theta < \frac{1}{\tau}. \quad (22)$$

The *hazard rate* is recognized as a force of mortality.

478 The hazard rate, $h(y)$ of the IRRIB process is

$$479 \quad hazardRate = \frac{\Gamma(m\tau + y) \left\{ \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\}^y \left\{ 1 - \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\}^{m\tau}}{y! \Gamma(m\tau) P(F_{(2[y+1], 2m\tau)df} \leq \frac{m\tau \{(1-\theta)^{(\tau-1)} - \theta\}}{\theta})}; 0 < \theta < \frac{1}{\tau} < 1. \quad (23)$$

480 The Markov chain is a way of capturing the *memory* in the chance mechanism. It is known
481 (Stuart and Ord, 2015) that the traditional geometric process maintains no memory. Its expected
482 value is that $\Pr(Y \geq q | Y \geq p) = \Pr(Y \geq q)$. With $\tau = 1$, we note a memory relation that

$$483 \quad \Pr(Y \geq q | Y \geq p) = \frac{\Pr(Y \geq p + q)}{\Pr(Y \geq p)} = \frac{\theta^{p+q}}{\theta^p} = \theta^q = \Pr(Y \geq q); 0 < \theta < 1 \quad (24)$$

484 confirming that the traditional geometric process does not hold any memory. Does the IRRIB
485 process deviates from such lack of memory? To answer this question, we first notice
486 $\Pr(Y \geq q | Y \geq p) \neq \Pr(Y \geq q)$. It suggests that there is a memory in the chance-oriented health
487 mechanism which is portrayed by IRRIB process. The memory level is

$$488 \quad \begin{aligned} & healthSystem'sMemory \\ &= \frac{P(F_{(2[p+q], 2m\tau)df} \leq \frac{m\tau \{(1-\theta)^{(\tau-1)} - \theta\}}{\theta})}{P(F_{(2p, 2m\tau)df} \leq \frac{m\tau \{(1-\theta)^{(\tau-1)} - \theta\}}{\theta})} - P(F_{(2q, 2m\tau)df} \leq \frac{m\tau \{(1-\theta)^{(\tau-1)} - \theta\}}{\theta}); 0 < \theta < \frac{1}{\tau} < 1. \end{aligned} \quad (25)$$

489 When $\tau = 1$, corresponding to the family union in which the attendees are homogeneously
490 immune to COVID-19, the survival function, hazard rate, and the memory of the traditional IB
491 process become, respectively,

$$492 \quad survivalFunction = S(p, \theta, \tau = 1) = P(F_{(2p, 2m)df} \leq \frac{m(1-\theta)}{\theta}); 0 < \theta < 1, \quad (26)$$

493 and

$$heathSystem'sMemory = \frac{P(F_{(2[p+q], 2m)df} \leq \frac{m(1-\theta)}{\theta})}{P(F_{(2p, 2m)df} \leq \frac{m(1-\theta)}{\theta})} - P(F_{(2q, 2m)df} \leq \frac{m(1-\theta)}{\theta}); 0 < \theta < 1$$

Next, we discuss the estimation of the parameters of the IRRIB process and obtain parameter estimates for IB process as particular cases.

Consider a random sample $\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3, \dots, \mathcal{Y}_n$ of size n from the IRRIB process. We resort to the maximum likelihood estimate (MLE) of the parameters because of its virtue. That is, the maximum likelihood estimate of a function of the parameters is the function of their MLEs. In the case of homogeneously immune to COVID-19, the MLE of the risk to contract COVID-19 is

$$\hat{\theta}_{\tau=1, mle} = \left(\frac{\bar{y}}{s_y^2}\right), \text{ if } s_y^2 > \bar{y} \text{ and } \hat{\theta}_{\tau=1, mle} = \left(\frac{s_y^2}{\bar{y} + s_y^2}\right), \text{ if } s_y^2 > \bar{y}. \quad (28)$$

In the case of heterogeneously resisting to COVID-19, the MLEs are

$$\hat{\theta}_{\tau \neq 1, mle} = \left(\frac{s_y^2}{s_y^2 + \bar{y}}\right) \quad \text{and consequently,} \quad \hat{\tau}_{mle, \tau \neq 1} \approx \frac{\left(\frac{s_y^2}{\bar{y}}\right)}{\hat{\theta}}. \quad (29)$$

To perform the test of hypothesis on the significance of the MLE, we could consider Neyman's $C(\alpha)$ methodology. Regression based test was developed by Neyman (1959) to perform a composite hypothesis in dealing with a non-Gaussian random sample and it is recognized as $\mathcal{Q}(\alpha)$ test. An alternative but easier approach to test the null hypothesis $H_0: \tau=1$ against an alternative hypothesis $H_1: \tau \neq 1$ is developed below. The null hypothesis is rejected when the test statistic,

$$z = \left(\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \right) \sqrt{n-1} \quad (30)$$

is significantly larger and accept the alternative hypothesis. In other words, the probability for the null hypothesis $H_o: \tau=1$ to be the true statement is

$$p - value = \Phi \left[\left(\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} - E \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} \right) \left(\sqrt{\text{var} \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1}} \right)^{-1} \right], \quad (31)$$

where $\Phi(\cdot)$ is the Gaussian cumulative function,

$$E \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} \quad \text{and} \quad \text{Var} \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} \quad (32)$$

are the expected value and variance under the null hypothesis. The null hypothesis $H_o: \tau=1$ is synonymous with the statement that the random sample is drawn from the IB process (which is characterized by $s_y^2 > \bar{y}$). Consequently, the significance level for the null hypothesis $H_o: \tau=1$ is

$$p - value \approx \Phi \left[\left(\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} - E \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} \right) / \sqrt{\text{Var} \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1}} \right]. \quad (33)$$

The (statistical) power of accepting an alternative value $H_a: \tau=\tau^*$ is $power \approx 1 - \Phi(\tau^*)$, with $\alpha = \Pr(\text{Reject } H_o: \tau=1) = 0.05$ and

$$\tau^* = \frac{1.645 \sqrt{\text{var} \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} + E \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1} - E \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_1 \right\}_{\tau \neq 1}}}{\sqrt{\text{var} \left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o \right\}_{\tau=1}}}. \quad (34)$$

The hazard rate to contract COVID-19 in a situation in which the family union consists of attendees having *heterogeneous resistance to COVID-19* is

$$h(r|\theta, \tau) = 1 - \frac{S(r+1, \theta_{\tau \neq 1}, \tau)}{S(r, \theta_{\tau \neq 1}, \tau)} = 1 - \frac{P(F_{(2[r+1], 2m\tau)df} \leq \frac{m\tau\{(1-\theta_{\tau \neq 1})^{(\tau-1)} - \theta_{\tau \neq 1}\}}{\theta_{\tau \neq 1}})}{P(F_{(2r, 2m\tau)df} \leq \frac{m\tau\{(1-\theta_{\tau \neq 1})^{(\tau-1)} - \theta_{\tau \neq 1}\}}{\theta_{\tau \neq 1}})}; 0 < \theta_{\tau \neq 1} < \frac{1}{\tau} < 1.$$

(35)

Initially (that is, $r = 1$), the hazard rate is

$$h(1|\theta_{\tau \neq 1}, \tau) = 1 - \frac{P(F_{(4, 2m\tau)df} \leq \frac{m\tau\{(1-\theta_{\tau \neq 1})^{(\tau-1)} - \theta_{\tau \neq 1}\}}{\theta_{\tau \neq 1}})}{P(F_{(2, 2m\tau)df} \leq \frac{m\tau\{(1-\theta_{\tau \neq 1})^{(\tau-1)} - \theta_{\tau \neq 1}\}}{\theta_{\tau \neq 1}})}; 0 < \theta < \frac{1}{\tau} < 1 \quad (36)$$

in the family union's attendees having *heterogeneous resistance to COVID-19* in comparison to

$$h(1|\theta_{\tau=1}, \tau = 1) = 1 - \frac{P(F_{(4, 2m)df} \leq \frac{m(1-\theta_{\tau=1})}{\theta_{\tau=1}})}{P(F_{(2, 2m)df} \leq \frac{m(1-\theta_{\tau=1})}{\theta_{\tau=1}})}; 0 < \theta < 1 \quad (37)$$

in a situation in which the family union consists of attendees having *homogeneously immune to COVID-19*. The reduction in the hazards due to having attendees with *heterogeneously resistant to COVID-19* in the family union is $Reduction = h(1|\theta_{\tau=1}, \tau = 1) - h(1|\theta_{\tau \neq 1}, \tau)$.

We next consider a popular concept called *Tail Value at Risk* (TVaR) in the business world (Khokhlov, 2016 for details) and it is useful in the context of contracting COVID-19. That is,

$$TVaR_{\tau} = E[Y|Y \geq 1, \theta_{\tau \neq 1}, \tau] \approx 1 + \frac{m\tau\left\{\frac{(1-\theta_{\tau \neq 1})^{(\tau-1)}}{\theta_{\tau \neq 1}} - 1\right\}}{\theta_{\tau \neq 1}P(F_{(2, 2m\tau)df} \leq m\tau\left\{\frac{(1-\theta_{\tau \neq 1})^{(\tau-1)}}{\theta_{\tau \neq 1}} - 1\right\})}; 0 < \theta < \frac{1}{\tau} < 1 \quad (38)$$

in the family union with attendees having *heterogeneous resistance to COVID-19* in comparison
to

$$TVaR_{\tau=1} = E[Y|Y \geq 1, \theta_{\tau=1}, \tau = 1] \approx 1 + \frac{m(1 - \theta_{\tau=1})}{\theta_{\tau=1} P(F_{(2, 2m)df} \leq \frac{m\{(1 - \theta_{\tau=1})\}}{\theta_{\tau=1}})}; 0 < \theta < 1 \quad (39)$$

in the family union consisting of attendees having *homogeneous resistance to COVID-19*.

S4 Appendix

Mathematical Derivations

A. Derivation of survival function

$$\begin{aligned} survivalFunction_{IRRI\bar{B}} &= S(p, \theta, \tau) = \Pr(Y \geq p) = \sum_{i=p}^{\infty} \frac{\Gamma(m\tau + i)}{i! \Gamma(m\tau)} \left\{ \frac{\theta}{(1 - \theta)^{(\tau-1)}} \right\}^i \left\{ 1 - \frac{\theta}{(1 - \theta)^{(\tau-1)}} \right\}^{m\tau} \\ &= P(F_{(2p, 2m\tau)df} \leq \frac{m\tau\{(1 - \theta)^{(\tau-1)} - \theta\}}{\theta}); 0 < \theta < \frac{1}{\tau} < 1. \end{aligned} \quad (40)$$

B. Derivation of hazard rate

$$\begin{aligned} hazardRate &= h(y) = \frac{\Pr(y)}{S(y+1)} \\ &= \frac{\Gamma(m\tau + y) \left\{ \frac{\theta}{(1 - \theta)^{(\tau-1)}} \right\}^y \left\{ 1 - \frac{\theta}{(1 - \theta)^{(\tau-1)}} \right\}^{m\tau}}{y! \Gamma(m\tau) P(F_{(2[y+1], 2m\tau)df} \leq \frac{m\tau\{(1 - \theta)^{(\tau-1)} - \theta\}}{\theta})}; 0 < \theta < \frac{1}{\tau} < 1. \end{aligned} \quad (41)$$

C. Markov chain

$$\Pr(Y \geq q | Y \geq p) = \frac{\Pr(Y \geq p+q)}{\Pr(Y \geq p)} = \frac{P(F_{(2[p+q], 2)df} \leq \frac{\{(1 - \theta)^{(\tau-1)} - \theta\}}{\theta})}{P(F_{(2p, 2)df} \leq \frac{\{(1 - \theta)^{(\tau-1)} - \theta\}}{\theta})} \neq P(F_{(2q, 2)df} \leq \frac{\{(1 - \theta)^{(\tau-1)} - \theta\}}{\theta})$$

554 (42)

555 D. Derivation of memory

$$556 \quad \begin{aligned} & healthSystem'sMemory = \Pr(Y \geq q | Y \geq p) - \Pr(Y \geq q) \\ & = \frac{P(F_{(2[p+q], 2m\tau)df} \leq \frac{m\tau\{(1-\theta)^{(\tau-1)} - \theta\}}{\theta})}{P(F_{(2p, 2m\tau)df} \leq \frac{m\tau\{(1-\theta)^{(\tau-1)} - \theta\}}{\theta})} - P(F_{(2q, 2m\tau)df} \leq \frac{m\tau\{(1-\theta)^{(\tau-1)} - \theta\}}{\theta}); 0 < \theta < \frac{1}{\tau} < 1. \end{aligned}$$

557 (43)

558 E. Derivation of memory in union with homogeneous attendees

$$559 \quad \begin{aligned} & heathSystem'sMemory = \Pr(Y \geq q | Y \geq p) - \Pr(Y \geq q) \\ & = \frac{P(F_{(2[p+q], 2m)df} \leq \frac{m(1-\theta)}{\theta})}{P(F_{(2p, 2m)df} \leq \frac{m(1-\theta)}{\theta})} - P(F_{(2q, 2m)df} \leq \frac{m(1-\theta)}{\theta}); 0 < \theta < 1 \end{aligned} \quad (44)$$

560 F. Derivation of maximum likelihood estimators for the parameters

561 The log-likelihood function is

$$562 \quad \begin{aligned} & \ln L(y_1, y_2, y_3, \dots, y_n, m, \tau, \theta) \\ & = \sum_{i=1}^n \{ \ln \Gamma(m\tau + y_i) - \ln y_i \} + n\bar{y} \{ \ln \theta - (\tau - 1) \ln(1 - \theta) \} + n \{ m\tau \ln(1 - \frac{\theta}{(1-\theta)^{(\tau-1)}}) - \ln \Gamma(m\tau) \}. \end{aligned}$$

563 (45)

564 Remember that the parameter m is assumed to be known. To get the MLE for the risk parameter, θ

565 and the restriction parameter, τ , their score functions $\partial_{\theta} \ln L$ and $\partial_{\tau} \ln L$ are equated to zero

566 simultaneously and solved. For this purpose, we need to solve the score function $\frac{\theta}{(1-\theta)^{(\tau-1)}} = \frac{\bar{y}}{s_y^2}$

567 we take its logarithm and rewrite as below. That gives

$$568 \quad \ln[1 - (1 - \theta)] - (\tau - 1) \ln(1 - \theta) = \ln\left(\frac{\bar{y}}{s_y^2}\right) \quad (46)$$

569 Applying an approximation $-\ln(1-u) \approx u, 0 < u < 1$ on both sides and simplifying, we note that

$$570 \quad -(1-\theta) + (\tau-1)\theta \approx \ln\left(\frac{\bar{y}}{s_y^2}\right), \text{ which is equivalent to } \tau = \frac{\{1 + \ln(\frac{\bar{y}}{s_y^2})\}}{\theta}. \quad (47)$$

571 In the case of homogeneously immune to COVID-19, the MLE of the risk to contract COVID-19

572 is $\hat{\theta}_{\tau=1, mle} = (\frac{\bar{y}}{s_y^2})$, if $s_y^2 > \bar{y}$ and $\hat{\theta}_{\tau=1, mle} = (\frac{s_y^2}{\bar{y} + s_y^2})$, if $s_y^2 < \bar{y}$. In the case of heterogeneously

573 resisting to COVID-19, the MLEs are

$$574 \quad \hat{\theta}_{\tau \neq 1, mle} = (\frac{s_y^2}{s_y^2 + \bar{y}}) \text{ and consequently, } \hat{\tau}_{mle, \tau \neq 1} \approx \frac{(\frac{s_y^2}{\bar{y}})}{\hat{\theta}}. \quad (48)$$

575 **G. Derivation of general and special expressions**

576 We derive their expressions in a general framework and then deduct separately expressions
577 under the null and alternative hypothesis. Recall what is already known that the sample expected
578 value and variance are stochastically independent when the data are drawn from the Gaussian
579 population (Stuart and Ord, 2015) but not necessarily so from the IRRIB process. Note that

$$580 \quad E\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|}\right\} = \frac{E(\bar{y}^2)}{E(|s_y^2 - \bar{y}|)} \left\{1 + \frac{Var(|s_y^2 - \bar{y}|)}{E(|s_y^2 - \bar{y}|)}\right\} - \rho_{\bar{y}^2, |s_y^2 - \bar{y}|} \frac{\sqrt{Var(\bar{y}^2)Var(|s_y^2 - \bar{y}|)}}{\{E(|s_y^2 - \bar{y}|)\}^2}$$

581 (49)

582 and

$$583 \quad Var\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|}\right\} = \left\{\frac{E(\bar{y}^2)}{E(|s_y^2 - \bar{y}|)}\right\}^2 \left\{\frac{Var(\bar{y}^2)}{[E(\bar{y}^2)]^2} + \frac{Var(|s_y^2 - \bar{y}|)}{[E(|s_y^2 - \bar{y}|)]^2}\right\} - 2\rho_{\bar{y}^2, |s_y^2 - \bar{y}|} \frac{\sqrt{Var(\bar{y}^2)Var(|s_y^2 - \bar{y}|)}}{\{E(|s_y^2 - \bar{y}|)\} \{E(\bar{y}^2)\}}.$$

584 (50)

585 (Blumenfeld, D, 2010) for details for the expectation and variance of the correlated expressions
586 in the ratio). Note that

$$E(\bar{y}^2) \approx \frac{m\tau}{n} \left\{ \frac{(1-\theta)^{(\tau-1)}}{\theta} \right\}^2; 0 < \theta < \frac{1}{\tau} < 1 \quad (51)$$

$$Var(\bar{y}^2) \approx \frac{(m\tau)^3}{n} \left\{ \frac{(1-\theta)^{(\tau-1)}}{\theta} \right\}^4; 0 < \theta < \frac{1}{\tau} < 1, \quad (52)$$

$$E(|s_y^2 - \bar{y}|) \approx (1 - \frac{1}{n}) m\tau \left\{ \frac{(1-\theta)^{(\tau-1)}}{\theta} \right\}^2; 0 < \theta < \frac{1}{\tau} < 1 \quad (53)$$

and

$$Var(|s_y^2 - \bar{y}|) \approx \frac{m\tau}{n} \left\{ \frac{(1-\theta)^{(\tau-1)}}{\theta} \right\}^2; 0 < \theta < \frac{1}{\tau} < 1. \quad (54)$$

Following an approach in Shanmugam (2020), the correlation, $\rho_{\bar{y}^2, |s_y^2 - \bar{y}|}$ between $\bar{y}^2$ and $|s_y^2 - \bar{y}|$ is

$$\rho_{\bar{y}^2, |s_y^2 - \bar{y}|} = \frac{E(|s_y^2 - \bar{y}|)}{E(|s_y^2 - \bar{y}|) + E(\bar{y}^2)} \approx \frac{\frac{\theta}{(1-\theta)^{(\tau-1)}} - (1 - \frac{1}{n})}{\frac{\theta}{(1-\theta)^{(\tau-1)}} - (1 - \frac{1}{n}) + m\tau}; 0 < \theta < \frac{1}{\tau} < 1 \quad (55)$$

and hence, from (14), we note that

$$E\left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \right\} \approx \frac{m\tau}{(n-1)} \left\{ 1 - \frac{\rho_{\bar{y}^2, |s_y^2 - \bar{y}|} \theta}{(1-\theta)^{(\tau-1)}} \right\}; 0 < \theta < \frac{1}{\tau} < 1 \quad (56)$$

and

$$Var\left\{ \frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \right\} \approx \frac{m\tau}{(n-1)} \left\{ \frac{\theta}{(1-\theta)^{(\tau-1)}} \right\} \{1 - \rho_{\bar{y}^2, |s_y^2 - \bar{y}|}\}; 0 < \theta < \frac{1}{\tau} < 1. \quad (57)$$

To calculate the p-value, we need specific expressions for

$$E\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o\right\}_{\tau=1} \quad (58)$$

$$\text{and} \quad Var\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_o\right\}_{\tau=1} . \quad (59)$$

To compute the statistical power of accepting the alternative hypothesis $H_1: \tau \neq 1$ when it is true,

we need expressions for

$$E\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_1\right\}_{\tau \neq 1} \quad \text{and} \quad Var\left\{\frac{\bar{y}^2}{|s_y^2 - \bar{y}|} \middle| H_1\right\}_{\tau \neq 1} . \quad (60)$$

bioRxiv preprint doi: <https://doi.org/10.1101/2020.12.21.423761>; this version posted December 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1

bioRxiv preprint doi: <https://doi.org/10.1101/2020.12.21.423761>; this version posted December 21, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3