scientific reports

Check for updates

OPEN

Latent class analysis identifies risk groups to model the expected benefits of SARS-CoV-2 interventions among university students

Callum R. K. Arnold[1,2]✉, Nita Bharti[1,2], Cara Exten[3], Meg Small[4,5], Sreenidhi Srinivasan[2,6], Suresh V. Kuchipudi[7], Vivek Kapur[2,6,8] & Matthew J. Ferrari[1,2]

Non-pharmaceutical public health measures (PHMs) were central to pre-vaccination efforts to reduce Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) exposure risk; heterogeneity in adherence placed bounds on their potential effectiveness, and correlation in their adoption makes assessing the impact attributable to an individual PHM difficult. During the Fall 2020 semester, we used a longitudinal cohort design in a university student population to conduct a behavioral survey of intention to adhere to PHMs, paired with an IgG serosurvey to quantify SARS-CoV-2 exposure at the end of the semester. Using latent class analysis on behavioral survey responses, we identified three distinct groups among the 673 students with IgG samples: 256 (38.04%) students were in the most adherent group, intending to follow all guidelines, 306 (46.21%) in the moderately-adherent group, and 111 (15.75%) in the least-adherent group, rarely intending to follow any measure, with adherence negatively correlated with seropositivity of 25.4%, 32.2% and 37.7%, respectively. Moving all individuals in an SIR model into the most adherent group resulted in a 77–96% reduction in seroprevalence, dependent on assumed assortativity. The potential impact of increasing PHM adherence was limited by the substantial exposure risk in the large proportion of students already following all PHMs.

Keywords Latent class analysis, SIR model, Approximate Bayesian Computation, Behavioral survey, IgG serosurvey

Within epidemiology, the importance of heterogeneity, whether that host, population, statistical, or environmental, has long been recognized[1–5]. For example, when designing targeted interventions, it is crucial to understand and account for differences that may exist within populations[6–8]. These differences can present in a variety of forms: heterogeneity in susceptibility, transmission, response to guidance, and treatment effects etc.; all of which affect the dynamics of an infectious disease[1,2,6,9–14]. While heterogeneity may exist on a continuous spectrum, it can be difficult to incorporate into analysis and interpretation, so individuals are often placed in discrete groups according to a characteristic that aims to represent the true differences[15–19]. When examining optimal influenza vaccination policy in the United Kingdom, Baguelin et al.[20] classified individuals within one of seven age groups. Explicitly accounting for, and grouping, individuals by whether they inject drugs can help target interventions to reduce human immunodeficiency virus (HIV) and Hepatitis C Virus incidence[21]. Similarly, epidemiological models have demonstrated the potential for HIV pre-exposure prophylaxis to reduce

[1]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA. [2]Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA 16802, USA. [3]Ross and Carole Nese College of Nursing, Pennsylvania State University, University Park, PA 16802, USA. [4]College of Health and Human Development, Pennsylvania State University, University Park, PA 16802, USA. [5]Social Science Research Institute, Pennsylvania State University, University Park, PA 16802, USA. [6]Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA. [7]Department of Infectious Diseases and Microbiology, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [8]Department of Animal Science, Pennsylvania State University, University Park, PA 16802, USA. ✉email: contact@callumarnold.com

racial disparities in HIV incidence[22]. Therefore, heterogeneity can be used to inform more complete theories of change, increasing intervention effectiveness[23].

When discretizing a population for the purposes of inclusion within a mechanistic model, three properties need to be defined: (1) the number of groups, (2) the size of the groups, and (3) the differences between the groups. Typically, as seen in the examples above, demographic data is used e.g., age, sex, race, ethnicity, socio-economic status, etc., often in conjunction with the contact patterns and rates[7,9,15,17,20,22,24]. There are several reasons for this: the data is widely available, and therefore can be applied almost universally; it is easily understandable; and there are clear demarcations of the groups, addressing properties 1) and 2). However, epidemiological models often aim to assess the effects of heterogeneity with respect to infection, e.g., "how does an individual's risk tolerance affect their risk of infection for influenza?". When addressing questions such as these, demographic data does not necessarily provide a direct link between the discretization method and the heterogeneous nature of the exposure and outcome, particularly if behavioral mechanisms are a potential driver. Instead, it relies on assumptions and proxy measures e.g., an individual's age approximates their contact rates, which in turn approximates their risk of transmission. This paper demonstrates an alternative approach to discretizing populations for use within mechanistic models, highlighting the benefits of an interdisciplinary approach to characterize heterogeneity in a manner more closely related to the risk of infection.

In early 2020, shortly after the World Health Organization (WHO) declared the SARS-CoV-2 outbreak a public health emergency of international concern[25], universities across the United States began to close their campuses and accommodations, shifting to remote instruction[26,27]. By Fall 2020, academic institutions transitioned to a hybrid working environment (in-person and online), requiring students to return to campuses[28–30]. In a prior paper[31] we documented the results of a large prospective serosurvey conducted in State College, home to The Pennsylvania State University (PSU) University Park (UP) campus. We examined the effect of 35,000 returning students (representing a nearly 20% increase in the county population[32]) on the community infection rates, testing serum for the presence of anti-Spike Receptor Binding Domain (S/RBD) IgG, indicating prior exposure[33]. Despite widespread concern that campus re-openings would lead to substantial increases in surrounding community infections[28,34,35], very little sustained transmission was observed between the two geographically coincident populations[31].

Given the high infection rate observed among the student body (30.4% seroprevalence), coupled with the substantial heterogeneity in infection rates between the two populations, we hypothesized that there may be further variation in exposure within the student body, resulting from behavioral heterogeneity. Despite extensive messaging campaigns conducted by the University[36], it is unlikely that all students equally adhered to public health guidance regarding SARS-CoV-2 transmission prevention. We use students' responses to the behavioral survey to determine and classify individuals based on their intention to adhere to public health measures (PHMs). We then show that these latent classes are correlated with SARS-CoV-2 seroprevalence. Finally, we parameterize a mechanistic model of disease transmission within and between these groups, and explore the impact of public health guidance campaigns, such as those conducted at PSU[36]. We show that interventions designed to increase student compliance with PHMs would likely reduce overall transmission, but the relatively high initial compliance limits the scope for improvement via PHM adherence alone.

Methods

Design, setting, and participants

This research was conducted with PSU Institutional Review Board approval and in accordance with the Declaration of Helsinki, and informed consent was obtained for all participants. The student population has been described in detail previously[31], but in brief, students were eligible for the student cohort if they were: ≥ 18 years old; fluent in English; capable of providing their own consent; residing in Centre County at the time of recruitment (October 2020) with the intention to stay through April 2021; and officially enrolled as PSU UP students for the Fall 2020 term. Upon enrollment, students completed a behavioral survey in REDCap[37] to assess adherence and attitudes towards public health guidance, such as attendance at gatherings, travel patterns, and nonpharmaceutical interventions. Shortly after, they were scheduled for a clinic visit where blood samples were collected. Students were recruited via word-of-mouth and cold-emails.

Outcomes

The primary outcome was the presence of S/RBD IgG antibodies, measured using an indirect isotype-specific (IgG) screening ELISA developed at PSU[38]. An optical density (absorbance at 450 nm) higher than six standard deviations above the mean of 100 pre-SARS-CoV-2 samples collected in November 2019, determined a threshold value of 0.169 for a positive result. Comparison against virus neutralization assays and RT-PCR returned sensitivities of 98% and 90%, and specificities of 96% and 100%, respectively[39]. Further details in the Supplement of the previous paper[31].

Statistical methods

To identify behavioral risk classes, we fit a range of latent class analysis (LCA) models (two to seven class models) to the student's behavioral survey responses, using the poLCA package[40] in the R programming language, version 4.3.3 (2024-02-29)[41]. We considered their answers regarding the frequency with which they intended to engage in the following behaviors to be a priori indicators of behavioral risk tolerance: wash hands with soap and water for at least 20s; wear a mask in public; avoid touching their face with unwashed hands; cover cough and sneeze; stay home when ill; seek medical attention when experiencing symptoms and call in advance; stay at least 6 feet (about 2 arms lengths) from other people when outside of their home; and, stay out of crowded places and avoid mass gatherings of more than 25 people. The behavioral survey collected responses on the Likert scale of: Never, Rarely, Sometimes, Most of the time, and Always. For all PHMs, Always and Most of the time accounted

for > 80% of responses (with the exception of intention to stay out of crowded places and avoid mass gatherings, where Always and Most of the time accounted for 78.8% of responses). To reduce the parameter space of the LCA and minimize overfitting, the behavioral responses were recoded as Always and Not Always. Measures of SARS-CoV-2 exposure e.g., IgG status, were not included in the LCA model fitting, as they reflect the outcome of interest. We focused on responses regarding intention to follow behaviors because this information can be feasibly collected during a public health campaign for a novel or emerging outbreak; it has also been shown that intentions are well-correlated with actual behaviors for coronavirus disease 2019 (COVID-19) public health guidelines, as well as actions that have short-term benefits[42,43]. We examined the latent class models using Bayesian Information Criterion, which is a commonly recommended as part of LCA model evaluation[44,45], to select the model that represented the best balance between parsimony and maximal likelihood fit.

Using the best-fit LCA model, we performed multivariate logistic regression of modal class assignment against IgG seropositivity to assess the association between the latent classes and infection. This "three-step" approach is recommended over the "one-step" LCA model fit that includes the outcome of interest as a covariate in the LCA model[45,46]. The following variables were determined a priori to be potential risk factors for exposure[31]: close proximity (6 feet or less) to an individual who tested positive for SARS-CoV-2; close proximity to an individual showing key COVID-19 symptoms (fever, cough, shortness of breath); lives in University housing; ate in a restaurant in the past 7 days; ate in a dining hall in the past 7 days; only ate in their room/apartment in the past 7 days; travelled in the 3 months prior to returning to campus; and travelled since returning to campus for the Fall term. Variables relating to attending gatherings were not included in the logistic regression due to overlap with intention variables of the initial LCA fit. Missing variables were deemed "Missing At Random" and imputed using the mice package[47], as described in the supplement of the previous paper[31]. To examine the effect of modal class assignment, we computed the GLM results of 100 simulations where class assignment was drawn according to each individual's probability of class membership, resulting from the LCA fitting process. The percentage of simulations where class membership p-values ≤ 0.05 were computed, and the mice package was used to produce pooled odds ratios and associated 95% confidence intervals.

We parameterized a deterministic compartmental Susceptible-Infected-Recovered (SIR) model using Approximate Bayesian Computation (ABC) against the seroprevalence within each latent class. The recovery rate was set to 8 days (Supplemental Table 1). Diagonal values of the transmission matrix were constrained such that $\beta HH \leq \beta MM \leq \beta LL$ (H represents high-adherence to public health guidelines, and M and L represent medium- and low-adherence, respectively), with the following parameters fit: the transmission matrix diagonals, a scaling factor for the off-diagonal values (ϕ), and a scaling factor for the whole transmission matrix (ρ). The off-diagonal values are equal to a within-group value (diagonal) multiplied by a scaling factor (ϕ). This scaling factor can either multiply the within-group beta value of the source group (e.g., $\beta HL = \phi \cdot \beta LL$; Eq. 1A), or the recipient group (e.g., $\beta LH = \phi \cdot \beta LL$; Eq. 1B), each with a different interpretation.

$$\rho\begin{pmatrix}\beta_{HH} & \beta_{HM} & \beta_{HL}\\ \beta_{MH} & \beta_{MM} & \beta_{ML}\\ \beta_{LH} & \beta_{LM} & \beta_{LL}\end{pmatrix} \rightarrow \rho\begin{pmatrix}\beta_{HH} & \phi\beta_{MM} & \boldsymbol{\phi\beta_{LL}}\\ \phi\beta_{HH} & \beta_{MM} & \boldsymbol{\phi\beta_{LL}}\\ \phi\beta_{HH} & \phi\beta_{MM} & \boldsymbol{\beta_{LL}}\end{pmatrix} \text{ mixing structure } \boldsymbol{A}$$
$$\rightarrow \rho\begin{pmatrix}\beta_{HH} & \phi\beta_{HH} & \phi\beta_{HH}\\ \phi\beta_{MM} & \beta_{MM} & \phi\beta_{MM}\\ \boldsymbol{\phi\beta_{LL}} & \boldsymbol{\phi\beta_{LL}} & \boldsymbol{\beta_{LL}}\end{pmatrix} \text{ mixing structure } \boldsymbol{B} \quad (1)$$

The former assumes that between-group transmission is dominated by the transmissibility of the source individuals, implying that adherence to the PHMs primarily prevents onwards transmission, rather than protecting against infection. The latter assumes that between-group transmission is dominated by the susceptibility of the recipient individuals, implying that adherence to the PHMs primarily prevents infection, rather than protecting against onwards transmission. A range of between-group scaling values (ϕ) were simulated to perform sensitivity analysis for the degree of assortativity. Results are only shown for matrix structure $\boldsymbol{A}$, but alternative assumptions about between-group mixing can be found in the supplement (Supplemental Figs. 1–4). To examine the effect of an intervention to increase PHM adherence, we redistributed a proportion of low- and medium adherence individuals to the high adherence latent class, i.e., a fully effective intervention is equivalent to a single-group SIR model of high adherent individuals. Model fitting and simulation was conducted using the Julia programming language, version 1.10.5 [48].

Results

Demographics

Full details can be found in the prior paper[31], but briefly: 1410 returning students were recruited, 725 were enrolled, and 684 students completed clinic visits for serum collection between 26 October and 21 December 2020. Of these, 673 students also completed the behavioral survey between 23 October and 8 December 2020. The median age of the participants was 20 years (IQR: 19–21), 64.5% identified as female and 34.6% as male, and 81.9% identified as white. A large proportion (30.4%) were positive for IgG antibodies, and 93.5% (100) of the 107 students with a prior positive test reported testing positive only after their return to campus.

LCA fitting

Of the 673 participants, most students intended to always mask (81.0%), always cover their coughs/ sneezes (81.9%), and always stay home when ill (78.2%) (Table 1). Two of the least common intentions were social distancing by maintaining a distance of at least 6 feet from others outside of their home, avoiding crowded places

Intention to always	Always	Not always
Avoid face-touching with unwashed hands	293 (43.54%)	380 (56.46%)
Cover cough and sneeze	551 (81.87%)	122 (18.13%)
Seek medical attention when have symptoms and call in advance	480 (71.32%)	193 (28.68%)
Stay at least 6 feet (about 2 arms lengths) from other people when outside of home.	292 (43.39%)	381 (56.61%)
Stay home when ill	526 (78.16%)	147 (21.84%)
Stay out of crowded places and avoid mass gatherings > 25 people	357 (53.05%)	316 (46.95%)
Tested for COVID-19 twice or more	544 (80.83%)	129 (19.17%)
Wash hands often with soap and water for at least 20 s.	434 (64.49%)	239 (35.51%)
Wear a face cover (mask) in public	545 (80.98%)	128 (19.02%)

Table 1. Participants' intention to always or not always follow 8 public health measures.

Classes	Log likelihood	Akaike information criterion	Bayesian information criterion
2	− 2895.40	5828.81	5914.53
3	− 2715.67	5489.35	5620.19
4	− 2673.50	5425.00	5600.96
5	− 2658.46	5414.93	5636.00
6	− 2647.01	5412.03	5678.22
7	− 2636.05	5410.10	5721.41

Table 2. Log likelihood, AIC, and BIC of two to seven class LCA model fits.

Classes	AIC (Mean)	BIC (Mean)	AIC (Median)	BIC (Median)
2	794.33	839.44	794.18	839.29
3	794.29	843.92	794.23	843.86
4	797.52	851.66	797.50	851.64
5	799.69	858.34	799.70	858.35
6	796.91	860.08	796.84	860.00
7	794.68	862.36	794.67	862.35

Table 3. Mean and median AIC and BIC of multiply-imputed logistic regressions for two to seven class LCA models against IgG serostatus.

and mass gatherings > 25 people (43.4% and 53.1% respectively), and avoiding face-touching with unwashed hands (43.5%).

The four- and the three-class LCA models had the lowest BIC respectively (Table 2). Examining the four-class model, there was minimal difference in the classification of individuals, relative to the three-class model. In the four-class model, the middle class (of the three-class model) was split into two groups with qualitatively similar class-conditional item response probabilities i.e., conditional on class membership, the probability of responding "Always" to a given question, except for hand washing and avoiding face-touching with unwashed hands (Supplemental Table 2).

We fit a logistic regression model to predict binary IgG serostatus that included inferred class membership, in addition to other predictor variables we previously identified in[31]. The mean and median BIC and AIC indicated similar predictive ability of the three- and four-class LCA models (Table 3). Given these factors, the three-class model was selected for use in simulation for parsimony, requiring fewer assumptions and parameters to fit.

In the three-class model, approximately 15.75% of individuals were members of the group that rarely intended to always follow the PHMs, 38.04% intended to always follow all guidelines, and the remaining 46.21% mostly intended to mask, test, and manage symptoms, but not distance or avoid crowds (Table 4). We have labelled the three classes as "Low", "High" and "Medium Adherence" groups, respectively, for ease of interpretation. Examining the class-conditional item response probabilities, the Medium Adherence class had a probability of 0.88 of always wearing a mask in public, but a probability of only 0.19 of social distancing when outside of their homes, for example. Calculating the class-specific seroprevalence, the Low Adherence group had the highest infection rates (37.7%, 95% Binomial CI: 28.5–47.7%), the medium adherence the next highest (32.2%, 95% Binomial CI: 27.0-37.7%), and the most adherent group experienced the lowest infection rates (25.4%, 95% Binomial CI: 20.2–31.1%). Incorporating latent class membership into the imputed GLM model described in our previous paper (30) retained the relationship between adherence and infection. Relative to the least adherent group, the Medium Adherence group experienced a non-significant reduction in infection risk (aOR, 95% CI:

Measure Intention to Always:	Low Adherence	Medium Adherence	High Adherence
Wash my hands often with soap and water for at least 20 seconds.	0.04	0.57	0.96
Wear a face cover (mask) in public	0.13	0.88	0.99
Avoid face-touching with unwashed hands	0.00	0.21	0.86
Cover cough and sneeze	0.22	0.86	1.00
Stay home when ill	0.07	0.83	1.00
Seek medical attention when have symptoms and call in advance	0.03	0.70	0.98
Stay at least 6 feet (about 2 arms lengths) from other people when outside of my home.	0.00	0.19	0.87
Stay out of crowded places and avoid mass gatherings > 25 people	0.03	0.39	0.88
Tested for COVID-19 twice or more	0.76	0.82	0.81
Group Size	15.75%	46.21%	38.04%
Seroprevalence	37.70%	32.20%	25.40%

Table 4. Class-conditional item response probabilities shown in the main body of the table for a three class LCA model, with footers indicating the size of the respective classes, and the class-specific seroprevalence.

Covariate (response)/reference levels	aOR (multiple imputation)
Close proximity to known COVID-19 positive individual (yes)/no	3.41 (2.29–5.08, $p < 0.001$)
Close proximity to individual showing COVID-19 symptoms (yes)/no	0.86 (0.58–1.29, $p = 0.474$)
Lives in University housing (yes)/no	0.90 (0.55–1.47, $p = 0.685$)
Latent Class (medium adherence)/low adherence	0.73 (0.45–1.18, $p = 0.203$)
Latent Class (high adherence)/low adherence	0.59 (0.36–0.98, $p = 0.043$)
Travelled in the 3 months prior to campus arrival (yes)/no	1.12 (0.76–1.63, $p = 0.57$)
Travelled since campus arrival (yes)/no	0.87 (0.6–1.25, $p = 0.447$)
Ate in a dining hall in the past 7 days (yes)/no	1.32 (0.76–2.29, $p = 0.332$)
Ate in a restaurant in the past 7 days (yes)/no	1.14 (0.8–1.64, $p = 0.465$)
Only ate in their room in the past 7 days (yes)/no	0.87 (0.59–1.29, $p = 0.499$)

Table 5. Adjusted odds ratio (aOR) for risk factors of infection among the returning PSU UP student cohort.

0.73, 0.45–1.18), and the most adherent group a significant reduction (aOR, 95% CI: 0.59, 0.36–0.98) (Table 5). When class assignment was determined probabilistically, similar relationships were observed in the pooled results of logistic regression without confounders, relative to the least adherent group: Medium Adherence group (OR, 95% CI: 0.77, 0.47–1.28); and the High Adherence group (OR, 95% CI: 0.60, 0.36–0.99). 66% of the High Adherence group simulations resulted in a p-value ≤ 0.05.

Compartmental model

The ABC distance distributions indicated that low-moderate levels of between-group mixing better fit the data (Fig. 1). After model parameterization (Supplemental Tables 1,4–6), we examined the effect of increasing adherence to public health guidance. Moving all individuals into the High Adherence class resulted in a 77–96% reduction in final size; when low-moderate between-group mixing is simulated, a fully effective intervention results in approximately 96% (95th percentiles: 88–99%) reduction in final seroprevalence, and when between-group mixing is as likely as within-group mixing, a 89% (95th percentiles: 34–99%) reduction is observed (Fig. 2).

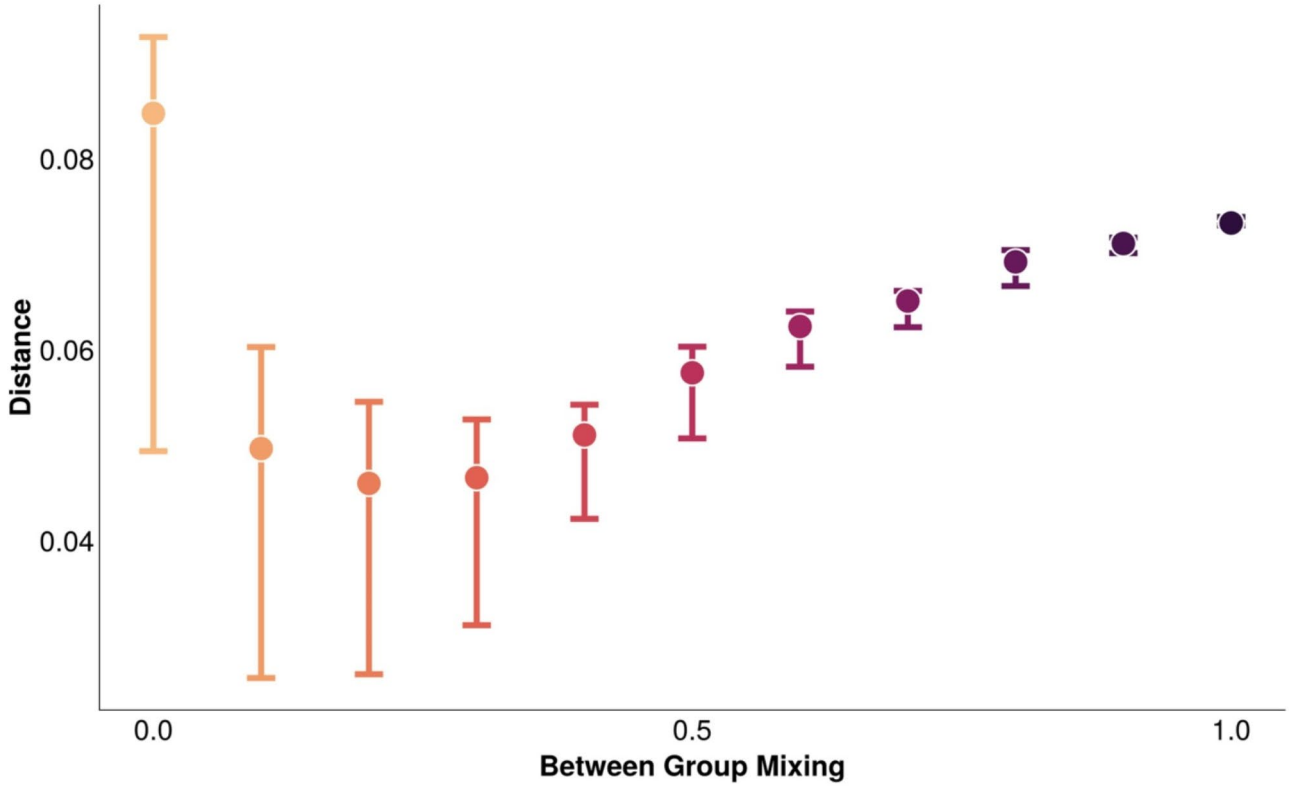

Fig. 1. Distribution of the distance from the ABC fits, with the minimum and maximum distances illustrated by the whiskers, and the median distance by the point. Between-group mixing of 1.0 equates to between-group mixing as likely as within-group mixing.

Discussion

In this interdisciplinary analysis, we collected behavioral data from surveys and integrated it with serosurveillance results. This approach allowed us to use LCA to categorize a population's transmission potential with measures related to risk tolerance and behavior. The LCA model was fit without inclusion of infection status data, but class membership was correlated with IgG seroprevalence. The classes that were the most adherent to PHMs experienced the lowest infection rates, and the least adherent exhibited the highest seroprevalence. As the logistic model cannot account for indirect effects resulting from between-class interactions, we parameterized a dynamical SIR model to explore the effect of interventions of varying degrees of effectiveness.

Although a four-class LCA model was a marginally better fit for the data, there were not substantial differences in class assignment relative to the three-class LCA model. The three-class model was selected for use in simulation for parsimony, requiring fewer assumptions and parameters to fit. Upon parametrizing the compartmental model, smaller ABC distance values were observed for low to moderate levels of between-group mixing, implying some degree of assortativity in our population, though the exact nature cannot be determined from our data. Examining the three classes, 38% of individuals already intended to always follow all PHMs. As a result, only 62% of the study population could have their risk reduced with respect to the PHMs surveyed. Further, the infection rates observed in the High Adherence group indicates that even a perfectly effective intervention aimed at increasing adherence to non-pharmaceutical PHMs (i.e., after the intervention, all individuals always followed every measure) would not eliminate transmission in a population, an observation that aligns with prior COVID-19 research[49–52]. The extent to which the infection in the High Adherence group is a result of mixing with lower adherence classes cannot be explicitly described, but the sensitivity analysis allows for an exploration of the effect and ABC fits suggest low-moderate levels of between-group mixing occurred. Varying the structure of the transmission matrix yielded very similar quantitative and qualitative results (Supplemental Figs. 1–4).

Examining the impact of increasing adherence to PHMs (modeled as increasing the proportion of the population in the High Adherence class), a fully effective intervention saw between a 77–96% reduction in the final size of the simulation outbreak. We note that the effect at a fully effective intervention is conceptually analogous to the population attributable fraction (PAF) proposed in[53]; though rather than quantifying the impact of removing one risk group, as in[53], we consider the impact of all individuals moving to the low-risk group. Each set of simulations for a given degree of assortativity has a different associated set of parameter values for the transmission matrix. The difference in the magnitude of the achievable reduction at a given level of intervention for the different assortativity levels is attributed to the difference in the corresponding fitted parameters (Supplemental Tables 4–6). With higher levels of between-group mixing, the initial SIR parameterization generally results in lower transmission parameters for the High-High adherence interactions,

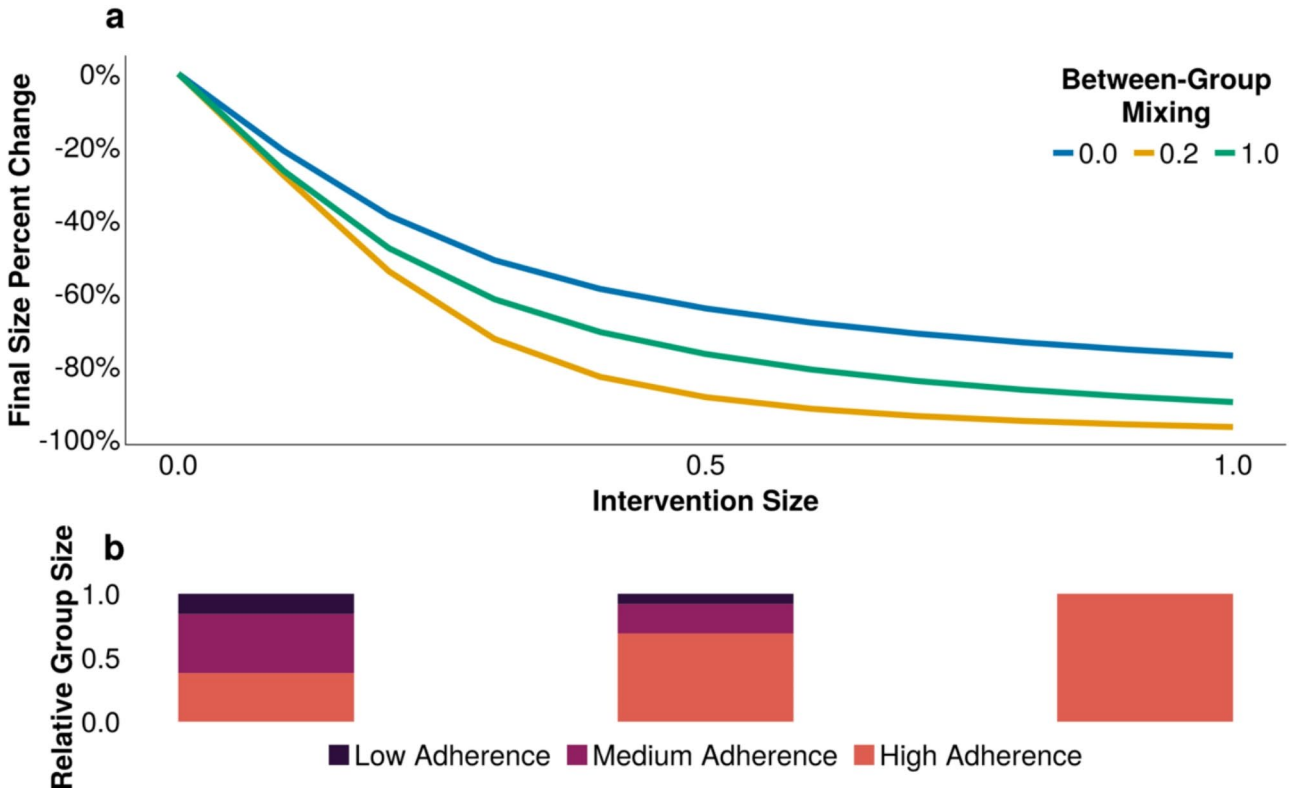

Fig. 2. (**A**) The reduction in final infection size across a range of intervention effectiveness (1.0 is a fully effective intervention), accounting for a range of assortativity. Between-group mixing of 1.0 equates to between-group mixing as likely as within-group mixing. Each line represents the mean reduction (of 200 simulations) resulting from an intervention, and is associated with a β matrix specific to the degree of between-group mixing illustrated; (**B**) The relative distribution of group sizes at three levels of intervention effectiveness (0.0, 0.2, 1.0).

as more infections in the High Adherence group originate from interactions with Low and Medium Adherence individuals. Increasing adherence, therefore, results in a greater reduction of the overall transmission rate than in simulations with less assortativity.

Limitations and strengths

The student population was recruited using convenience sampling, and therefore may not be representative of the wider population. Those participating may have been more cognizant and willing to follow public health guidelines. Similarly, because of the University's extensive messaging campaigns and efforts to increase access to non-pharmaceutical measures[36], such as lateral flow and polymerase-chain reaction diagnostic tests, the students likely had higher adherence rates than would be observed in other populations. However, these limitations are not inherent to the modeling approach laid out, and efforts to minimize them would likely result in stronger associations and conclusions due to larger differences in the latent behavioral classes and resulting group infection rates.

It is well known that classification methods, like LCA, can lead to the "naming fallacy"[44], whereby groups are assigned and then specific causal meaning is given to each cluster, affecting subsequent analyses and interpretation of results. In this paper, this effect is reduced by virtue of the analysis plan being pre-determined, and the relationship with the outcome showing a positive association with the classes in the mechanistically plausible direction (i.e., increasing adherence to PHMs results in reduced infection rates). Our decision to conduct the simulation analysis with the three-class model was, in part, to avoid the potential bias that would arise from naming or assigning an order to the two intermediate risk groups.

Despite these limitations, this work presents a novel application of a multidisciplinary technique, outlining how alternate data sources can guide future model parameterization and be incorporated into traditional epidemiological analysis, particularly within demographically homogeneous populations where there is expected or observed heterogeneity in transmission dynamics. This is particularly important in the design of interventions that aim to target individual behaviors, allowing the categorization of populations into dynamically-relevant risk groups and aiding in the efficient use of resources through targeted actions. Future research should consider including perceived agency and efficacy for PHM adherence.

Data availability

The datasets generated during and/or analyzed in the primary stages are not publicly available as they contain personally identifiable information, but are available from the corresponding author on reasonable request. All simulation code is readily available at https://github.com/arnold-c/The-Maximal-Expected-Benefit-of-SARScov2-intervention.

Data access, responsibility, and analysis

Callum Arnold and Dr. Matthew J. Ferrari had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Callum Arnold and Dr. Matthew J. Ferrari (Department of Biology, Pennsylvania State University) conducted the data analysis.

Received: 4 November 2024; Accepted: 19 March 2025
Published online: 02 April 2025

References

1. Fletcher, J. What is heterogeneity and is it important?? *BMJ* **334**, 94–96. https://doi.org/10.1136/bmj.39057.406644.68 (2007).
2. Nold, A. Heterogeneity in disease-transmission modeling. *Math. Biosci.* **52**, 227–240. https://doi.org/10.1016/0025-5564(80)90069-3 (1980).
3. Trauer, J. M. et al. The importance of heterogeneity to the epidemiology of tuberculosis. *Clin. Infect. Dis.* **69**, 159–166. https://doi.org/10.1093/cid/ciy938 (2019).
4. Zhang, Y., Britton, T. & Zhou, X. Monitoring real-time transmission heterogeneity from incidence data. *PLoS Comput. Biol.* **18**, e1010078. https://doi.org/10.1371/journal.pcbi.1010078 (2022).
5. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. https://doi.org/10.1038/nature04153 (2005).
6. Woolhouse, M. E. J. et al. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci. U S A* **94**, 338–342 (1997).
7. Wang, H., Ghosh, A., Ding, J., Sarkar, R. & Gao, J. Heterogeneous interventions reduce the spread of COVID-19 in simulations on real mobility data. *Sci. Rep.* **11**, 7809. https://doi.org/10.1038/s41598-021-87034-z (2021).
8. McDonald, S. A., Devleesschauwer, B. & Wallinga, J. The impact of Individual-Level heterogeneity on estimated infectious disease burden: A simulation study. *Popul. Health Met.* **14**, 47. https://doi.org/10.1186/s12963-016-0116-y (2016).
9. Sevelius, J. M., Patouhas, E., Keatley, J. G. & Johnson, M. O. Barriers and facilitators to engagement and retention in care among transgender women living with human immunodeficiency virus. *Ann. Behav. Med.* **47**, 5–16. https://doi.org/10.1007/s12160-013-9565-8 (2014).
10. Tuschhoff, B. M. & Kennedy, D. A. Detecting and quantifying heterogeneity in susceptibility using contact tracing data. https://doi.org/10.1101/2023.10.04.560944 (2023).
11. Delaney, K. P. Strategies adopted by gay, bisexual, and other men who have sex with men to prevent Monkeypox virus Transmission—United States, August 2022. *MMWR Morb Mortal. Wkly. Rep.* **71**. https://doi.org/10.15585/mmwr.mm7135e1 (2022).
12. Anderson, A. T L et al. Quantifying Individual-Level heterogeneity in infectiousness and susceptibility through household studies. *Medrxiv* **2022** https://doi.org/10.1101/2022.12.02.22281853 (2022).
13. MacDonald, K. S. et al. Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection. *J. Infect. Dis.* **181**, 1581–1589. https://doi.org/10.1086/315472 (2000).
14. Elie, B., Selinger, C. & Alizon, S. The source of individual heterogeneity shapes infectious disease outbreaks. *Proc. R. Soc. B: Biol.l Sci.* **289** 20220232. https://doi.org/10.1098/rspb.2022.0232 (2022).
15. Mossong, J. et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74. https://doi.org/10.1371/journal.pmed.0050074 (2008).
16. Klepac, P., Kissler, S. & Gog, J. Contagion! The bbc four pandemic—the model behind the documentary. *Epidemics* **24**, 49–59. https://doi.org/10.1016/j.epidem.2018.03.003 (2018).
17. Davies, N. G. et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* **26**, 1205–1211. https://doi.org/10.1038/s41591-020-0962-9 (2020).
18. Hay, J., Routledge, I., & Takahashi, S. Serodynamics a review of methods for epidemiological inference using serological data. https://doi.org/10.31219/osf.io/kqdsn (2024).
19. Yang, B. et al. Life course exposures continually shape antibody profiles and risk of seroconversion to influenza. *PLoS Pathog.* **16**, e1008635. https://doi.org/10.1371/journal.ppat.1008635 (2020).
20. Baguelin, M. et al. Assessing optimal target populations for influenza vaccination programmes: An evidence synthesis and modelling study. *PLoS Med.* **10**, e1001527. https://doi.org/10.1371/journal.pmed.1001527 (2013).
21. Levitt, A., Mermin, J., Jones, C. M., See, I. & Butler, J. C. Infectious diseases and injection drug use: Public health burden and response. *J. Infect. Dis.* **222**, S213–S217. https://doi.org/10.1093/infdis/jiaa432 (2020).
22. Jenness, S. M. et al. Addressing gaps in HIV preexposure prophylaxis care to reduce Racial disparities in HIV incidence in the united States. *Am. J. Epidemiol.* **188**, 743–752. https://doi.org/10.1093/aje/kwy230 (2019).
23. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980. https://doi.org/10.1038/s41562-021-01143-3 (2021).
24. Fox, S. J. et al. Disproportionate impacts of COVID-19 in a large US City. *PLoS Comput. Biol.* **19**, e1011149. https://doi.org/10.1371/journal.pcbi.1011149 (2023).
25. World Health Organization. Statement on the Second Meeting of the International Health Regulations. Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV) n.d. (2005). https://www.who.int/news/item/30-01-2020-statement-on-the-secondmeeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-theoutbreak-of-novel-coronavirus-(2019-ncov) (Accessed 19 Jan 2023).
26. Map Coronavirus and School Closures in 2019–2020. Education Week n.d.
27. Collegian, J. C. & | T D. TIMELINE | Penn State's Journey through the Coronavirus Pandemic (2021). https://www.psucollegian.com/news/coronavirus/timeline-penn-state-s-journey-through-thecoronavirus-pandemic/article_36086cac-a890-11eb-8dd3-27bfc231fbd1.html (Accessed 7 Feb 2024).
28. Adams, G. B., Shannon, J. & Shannon, S. Return to university campuses associated with 9% increase in new COVID-19 case rate.*Medrix* https://doi.org/10.1101/2020.10.13.20212183 (2020)
29. Hadden, J. What the Top 25 Colleges and Universities in the US Have Said about Their Plans to Reopen in Fall 2020, from Postponing the Semester to Offering More Remote Coursework n.d. (2024). https://www.businessinsider.com/how-major-us-colleges-plan-reopen-for-fall-2020-semester2020-5

30. Times, T. N. Y. Tracking the Coronavirus at U.S. Colleges and Universities - The New York Times 2020. (2021). https://www.nytimes.com/interactive/2020/us/covid-college-cases-tracker.html (Accessed 30 Jan)
31. Arnold, C. R. K. et al. A longitudinal study of the impact of university student return to campus on the SARS-CoV-2 Seroprevalence among the community members. *Sci. Rep.* **12**, 8586. https://doi.org/10.1038/s41598-022-12499-5 (2022).
32. United States Census Bureau. U.S. Census Bureau QuickFacts: Centre County, Pennsylvania 2019. (Accessed 30 January 2021). https://www.census.gov/quickfacts/centrecountypennsylvania
33. Long, Q-X. et al. Antibody responses to SARSCoV-2 in patients with COVID-19. *Nat. Med.* **26**, 845–848. https://doi.org/10.1038/s41591-0200897-1 (2020).
34. Lopman, B. et al. A modeling study to inform screening and testing interventions for the control of SARS-CoV-2 on university campuses. *Sci. Rep.* **11**, 5900. https://doi.org/10.1038/s41598-021-85252-z (2021).
35. Benneyan, J. et al. COVID-19 risk uncertainty under university reopening scenarios: Model-based analysis. *JMIR Public. Health Surveillance* **7**, e24292. https://doi.org/10.2196/24292 (2021).
36. Pennsylvania State University. Mask Up or Pack Up 2021. Accessed 6 Feb, (2021). https://virusinfo.psu.edu/mask-upor-pack-up/
37. Harris, P. A. et al. The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inf.* **95**, 103208. https://doi.org/10.1016/j.jbi.2019.103208 (2019).
38. Gontu, A. et al. Quantitative Estimation of IgM and IgG antibodies against SARS-CoV-2. Protocolsio 2020. https://doi.org/10.17504/protocols.io.bivgke3w
39. Gontu, A. et al. Limited window for donation of convalescent plasma with high live-virus neutralizing antibody titers for COVID-19 immunotherapy. *Commun. Biol.* **4**, 1–9. https://doi.org/10.1038/s42003-02101813-y (2021).
40. Linzer, D. A., Lewis, J. B. & poLCA An R package for polytomous variable latent class analysis. *J. Stat. Soft.* **42**, 1–29. https://doi.org/10.18637/jss.v042.i10 (2011).
41. R Core Team. R: A Language and Environment for Statistical Computing 2021. https://www.rproject.org/
42. Conner, M., Wilding, S. & Norman, P. Does intention strength moderate the intention–Health behavior relationship for Covid-19 protection behaviors? *Ann. Behav. Med.* **58**, 92–99. https://doi.org/10.1093/abm/kaad062 (2024).
43. McDonald, J., McDonald, P., Hughes, C. & Albarracín, D. Recalling and intending to enact health recommendations: Optimal number of prescribed behaviors in multibehavior messages. *Clin. Psychol. Sci.* **5**, 858–865. https://doi.org/10.1177/2167702617704453 (2017).
44. Weller, B. E., Bowen, N. K. & Faubert, S. J. Latent class analysis: A guide to best practice. *J. Black Psychol.* **46**, 287–311. https://doi.org/10.1177/0095798420930932 (2020).
45. Nylund-Gibson, K. & Choi, A. Y. Ten frequently asked questions about latent class analysis. *Transl. Issues Psychol. Sci.* **4**, 440. https://doi.org/10.1037/tps0000176 (2018).
46. Bolck, A., Croon, M. & Hagenaars, J. Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Anal.* **12**, 3–27. https://doi.org/10.1093/pan/mph001 (2004).
47. van Buuren, S., & Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
48. Bezanson, J., Edelman, A., Karpinski, S., Shah, V. B. & Julia A fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98. https://doi.org/10.1137/141000671 (2017).
49. Flaxman, S. et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261. https://doi.org/10.1038/s41586-020-2405-7 (2020).
50. Banholzer, N. et al. Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. *PLOS ONE.* **16**, e252827. https://doi.org/10.1371/journal.pone.0252827 (2021).
51. Brauner, J. M. et al. Inferring the effectiveness of government interventions against COVID-19. *Science* **371**, eabd9338. https://doi.org/10.1126/science.abd9338 (2021).
52. Ge, Y. et al. Untangling the changing impact of non-pharmaceutical interventions and vaccination on European COVID-19 trajectories. *Nat. Commun.* **13**, 3106. https://doi.org/10.1038/s41467-022-30897-1 (2022).
53. Brooks-Pollock, E. & Danon, L. Defining the population attributable fraction for infectious. Diseases *Int. J. Epidemiol.* ;**46**:976–982. https://doi.org/10.1093/ije/dyx055 (2017).

Acknowledgements

Florian Krammer, Mount Sinai, USA for generously providing the transfection plasmid pCAGGS- RBD. Scott E. Lindner, Allen M. Minns, Randall Rossi produced and purified RBD. The D4A Research Group: Dee Bagshaw, Clinical & Translational Science Institute, Cyndi Flanagan, Clinical Research Center and the Clinical & Translational Science Institute; Thomas Gates, Social Science Research Institute; Margeaux Gray, Dept. of Biobehavioral Health; Stephanie Lanza, Dept. of Biobehavioral Health and Prevention Research Center; James Marden, Dept. of Biology and Huck Institutes of the Life Sciences; Susan McHale, Dept. of Human Development and Family Studies and the Social Science Research Institute; Glenda Palmer, Social Science Research Institute; Connie J. Rogers, Dept. of Nutritional Sciences; Rachel Smith, Dept. of Communication Arts and Sciences and Huck Institutes of the Life Sciences; and Charima Young, Penn State Office of Government and Community Relations.. The authors thank the following for their assistance in the lab: Sophie Rodriguez, Natalie Rydzak, Liz D. Cambron, Elizabeth M. Schwartz, Devin F. Morrison, Julia Fecko, Brian Dawson, Sean Gullette, Sara Neering, Mark Signs, Nigel Deighton, Janhayi Damani, Mario Novelo, Diego Hernandez, Ester Oh, Chauncy Hinshaw, B. Joanne Power, James McGee, Riëtte van Biljon, Andrew Stephenson, Alexis Pino, Nick Heller, Rose Ni, Eleanor Jenkins, Julia Yu, Mackenzie Doyle, Alana Stracuzzi, Brielle Bellow, Abriana Cain, Jaime Farrell, Megan Kostek, Amelia Zazzera, Sara Ann Malinchak, Alex Small, Sam DeMatte, Elizabeth Morrow, Ty Somberger, Haylea Debolt, Kyle Albert, Corey Price, Nazmiye Celik.

Author contributions

Conceptualization: C.A., M.J.F. Data curation: C.A., M.J.F. Formal analysis: C.A., M.J.F. Funding acquisition: M.J.F. Investigation: N.B., C.E., M.S., S.S., S.K., V.S. Methodology: C.A., N.B., M.J.F. Project administration: M.J.F. Software: C.A., M.J.F. Supervision: M.J.F. Validation: C.A., M.J.F. Visualization: C.A., M.J.F. Writing—original draft: C.A. Writing—review and editing: all authors.

Funding

This work was supported by funding from the Office of the Provost and the Clinical and Translational Science Institute, Huck Life Sciences Institute, and Social Science Research Institutes at the Pennsylvania State University. The project described was supported by the National Center for Advancing Translational Sciences, National

Institutes of Health, through Grant UL1 TR002014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding sources had no role in the collection, analysis, interpretation, or writing of the report.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-95164-x.

Correspondence and requests for materials should be addressed to C.R.K.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.