

THE LANCET Microbe

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed.
We post it as supplied by the authors.

Supplement to: Li Y, Rivers J, Mathis S, et al. Genomic cluster formation among invasive group A streptococcal infections in the USA: a whole-genome sequencing and population-based surveillance study. *Lancet Microbe* 2024. [https://doi.org/10.1016/S2666-5247\(24\)00169-1](https://doi.org/10.1016/S2666-5247(24)00169-1)

Contents

1. Supplementary Methods - Page 1
1.1 Study Design and iGAS Cases and Isolates - Page 2
1.2 Procedures - Page 2
1.3 Statistical Analysis - Page 5
1.3.1 Cluster Formation Incidence and Hazard - Page 5
1.3.2 Univariable Analysis - Page 7
1.3.3 Multivariable Analysis - Page 7
1.3.4 Machine Learning Models - Page 9
1.4 Grouping of <i>emm</i> types - Page 13
1.5 Proportional hazard- Page 13
1.5 Genomic Clusters and Transmission - Page 13

2. Supplementary Tables

Table S1: Features and Cluster Formation among 2764 At-Risk iGAS Cases at 3-SNV Threshold - Page 16
Table S2: Features and Cluster Formation among 4348 At-Risk iGAS Cases at 1-SNV Threshold - Page 19
Table S3: Features and Cluster Formation among 2239 At-Risk iGAS Cases at 5-SNV Threshold - Page 22
Table S4: Multivariable Cox Regression Analysis of Variables Associated with Cluster Formation at 1-SNV Threshold - Page 25
Table S5: Multivariable Cox Regression Analysis of Variables Associated with Cluster Formation at 5-SNV Threshold - Page 27
Table S6: Initial Cox Regression Model for Variables Associated with Cluster Formation - Page 29
Table S7: Mixed Effects Cox Regression Analysis of Variables Associated with Cluster Formation - Page 31
Table S8: Performance Comparison among 13 Machine Learning Models - Page 33
Table S9: Predictive Performance for Cluster Formation Based on 1-SNV and 5-SNV Threshold - Page 34
Table S10: Predictive Performance for Binary Outcome (Cluster Formation Yes/No) at 30-, 90-, 180-Day, and 360-Day After Detection - Page 35
Table S11: Within-Cluster Genomic Differences by SNV- and Time-Cutoff - Page 36

3. Supplementary Figures

Figure S1: Relationship between isolate genomic relatedness and case epidemiological relatedness in time - Page 37
Figure S2: Relationship between isolate genomic relatedness and case epidemiological relatedness in location - Page 38
Figure S3: Relationship between isolate genomic relatedness and case epidemiological relatedness in population - Page 39
Figure S4: Evaluation of SNV Distances after removing SNVs within 1000bp - Page 40
Figure S5: Evaluation of SNV Distances without removing SNVs within 1000bp - Page 41
Figure S6: Assessment of the Schoenfeld residuals in the multivariable Cox regression model - Page 42
Figure S7: Functional Form of Age as a Continuous Variable with Binary Cluster Formation Outcome - Page 43
Figure S8: Variable Importance Plot of the Random Survival Forest Model - Page 44
Figure S9: Distribution of the Predicted Risk Score - Page 45
Figure S10: Predictive Performance (C-index) Evaluated by Leave-One-Out Cross-Validation - Page 46
Figure S11: Distribution of Risk Scores in the derivation cohort - Page 47
Figure S12: At-Risk Cases with Observed Cluster Formation - Page 48
Figure S13: Cumulative Incidence Curves in Leave-One-Out Cross-Validation - Page 49
Figure S14: Distribution of Risk Scores in the Validation Cohort - Page 50

4. References - Page 51

SUPPLEMENTARY METHODS

Study design and iGAS cases and isolates

In this whole-genome sequencing (WGS) and population-based surveillance study, we identified iGAS cases and isolates from Active Bacterial Core surveillance (ABCs), a population- and laboratory-based surveillance system of invasive bacterial infections covering approximately 34 million persons in 10 U.S. states¹. The iGAS case definition was an illness with isolation of GAS from a normally sterile site or isolation of GAS from a wound culture that was accompanied by necrotizing fasciitis or streptococcal toxic shock syndrome¹. Necrotizing fasciitis and streptococcal toxic shock syndrome each account for approximately 5% of all iGAS cases and the relative proportion has not changed substantially over time¹. WGS data enabled confirmation or exclusion of cases from epidemiologically defined outbreaks², assisted identification of community outbreaks^{3,4}, and provided detailed descriptions of cluster characteristics^{5,6}.

ABCs routinely contacted and audited clinical laboratories serving the surveillance population to ensure all iGAS cases were captured. Case information on basic demographics, underlying medical conditions, and risk factors for infection was collected. All available iGAS isolates were sent to CDC for WGS-based laboratory characterization. Cases and isolates were collected from residents of the 10 ABCs sites (California, Colorado, Connecticut, Georgia, Maryland, Minnesota, New Mexico, New York, Oregon, Tennessee) between January 1, 2015, and December 31, 2019. Residents of all ages residing in the 10 ABCs sites with iGAS infection that fit the cases definition were selected. Any GAS infections reported in the surveillance that did not fit the case definition were excluded. We also excluded cases in which whole genome assemblies of case isolate contained <1.5M total bases or >150 contigs as a sequencing

25 quality control measure. The quality filtering removed 249 isolates from a total of 9948 isolates with
26 assemblies. The remaining 9699 assemblies showed high quality metrics of assembly as indicated by
27 N50 (median: 151 Kb, Inter quartile range (IQR), 119-184 Kb), number of contigs (median 38, IQR: 32-
28 46), total assembled bases (median: 1.8Mb IQR: 1.77-1.84 Mb), and length of the longest contig
29 (median: 307 Kb, IQR: 243-425 Kb), indicating stringent filtering. In ABCs, each individual can
30 contribute only one isolate the same infection but can contribute more than one isolates from different
31 infections, defined as infections separated by at least 3 months. Within the at-risk cases, we had no
32 evidence of two isolates from the same individual.

33

34 **Procedures**

35 WGS-based strain characterization was performed at CDC's *Streptococcus* Laboratory as previously
36 described ⁷. Briefly, colony-purified iGAS isolates were cultured in Todd Hewitt Broth supplemented
37 with Yeast Extract (THY) medium (CDC Division of Scientific Resources, Atlanta, GA, USA) for DNA
38 isolation. DNA was extracted with the QIAamp DNA Mini Kit (QIAGEN, Germantown, Maryland,
39 USA) and prepared for sequencing on an Illumina MiSeq (Illumina, San Diego, CA, USA) platform
40 using the NEBNext Ultra kit (New England Biolabs, Ipswich, MA, UAS) and sample-specific
41 barcoding. Genomes were sequenced on the MiSeq platform with 250bp paired end reads. Quality of
42 reads was examined with FastQC version 0.11.5 and cutadapt version 1.8.3 was used for trimming
43 adapter sequences and low-quality bases (parameters: -q 20; --minimum-length 50). WGS reads were
44 analyzed by a validated bioinformatics pipeline to generate strain features (Table S1) and whole genome
45 assembly⁷. Briefly, GAS whole genome assembly was generated from WGS reads using
46 VelvetOptimiser version 2.2.6. *emm* types were determined by identifying an exact match of the 180-bp
47 sequence that defines *emm* types in the genome assembly based on the *emm* typing sequence database
48 maintain at the CDC (ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/tsemm/; last accessed March 24,

2024). Multilocus sequence type (MLST) was determined using the WGS reads and the SRST2 software⁸ based on the *Streptococcus pyogenes* multilocus sequence typing database (<https://pubmlst.org/organisms/streptococcus-pyogenes/>). Antimicrobial resistance (AMR) was determined by mapping WGS reads to previously described AMR databases⁷ to identify the presence of AMR genes. Sequence targets were identified for detection of the presence/absence of virulence related genes, including 11 pyrogenic exotoxin genes (*speA*, *speC*, *speG*, *speH*, *speI*, *speJ*, *speK*, *speL*, *speM*, *ssa*, and *smeZ*), Fibronectin-binding protein type A (FBAA) gene *fbaA*, Fibronectin-binding protein F2 (PRTF2) gene *prtF*, *Alp* gene family adhesin (R28), Fibronectin-binding protein I (SFB1) gene *sfbl*, Serum opacity factor (SOF) gene *sof*, hyaluronic acid synthetase (CAPSULE) gene *hasA*, Active NADase (NADASE_D330G), Pnga 3 - Clade 3 up-regulated promoter of the *nga* operon (PNGA3), *rocA* null mutations (ROCA), Virulence associated DNase (SDA1), and Streptococcal inhibitor of complement (SIC). The positivity of a virulence gene was defined as the presence of allele-specific sequence target in the WGS reads.

Pair-wise single-nucleotide variant (SNV) distances among all iGAS isolates belonging to the same *emm* type were calculated using the MUMmer package (version 3.9.0) as previously described⁹. SNVs within 1-kb of pair-wise alignment were removed to reduce the potential confounding effect of recombination on SNV distance. We verified this method in a subset of isolates by comparing the SNV distances generated in this study to those generated by the Gubbins software¹⁰ for the same isolates (Figure S4-S5). Among 1000 randomly selected pair-wise distances, SNV distances after removing SNVs within 1-kb showed an excellent correlation with distances generated by the Gubbins software ($R^2=0.982$). In contrast, distances without removing SNVs within 1-kb showed a much weaker correlation with distances generated by the Gubbins software ($R^2=0.363$).

Two iGAS case were defined as genomically clustered with each other if the case isolates differed by ≤ 3 SNVs per 1.5Mb of the two aligned genomes (2×10^{-6} SNV/site). Two case isolates of different *emm* types were considered as two different GAS strains and un-clustered with each other given the small SNV threshold to define a cluster. The SNV cutoff value was selected based on analysis of the relationship between isolate genomic relatedness and case epidemiological relatedness (time, location, and patient population) of two iGAS cases in the surveillance (Figure S1-S3). Compared with two randomly selected iGAS cases in the surveillance by permutation, two iGAS cases with isolates differed from each other by ≤ 7 SNVs were more likely to have occurred within a 90-day period (Figure S1), more likely to be in the same census tract (a geographic region defined for the purpose of taking a census in US with a population ranging from 8,000 to 1,200 people) (Figure S2), and more likely to be residents of the same long-term care facility (Figure S3). In contrast, two iGAS cases with isolates differed by ≥ 8 SNVs did not show substantially elevated level of relatedness in time, location, and patient population compared with two randomly selected iGAS cases. Further, within the range of 0-7 SNV threshold values, there was a clear decrease in the level of epidemiological relatedness as the genomic relatedness decreased (higher SNV distance). We therefore choose an intermediate level of SNV threshold (3 SNVs) to represent genomically clustered iGAS cases with intermediately level of elevated epidemiological relatedness for our main analysis. We further evaluated SNV values toward the higher end (5 SNVs) and lower end (1 SNV) of the threshold range, to represent genomically clustered iGAS cases with lower and higher levels of elevated epidemiological relatedness, respectively, in additional analyses. A breakdown table is used to illustrate the within-cluster SNV differences according to a range of thresholds for SNV and time cutoff (Table S11).

An iGAS case was defined at risk of cluster formation at the time of detection (culture positive date) if the case did not cluster with any prior cases in the surveillance (Figure 1). The outcome of interest was

cluster formation event, defined as the detection of a subsequent iGAS case in the surveillance that was clustered with the at-risk case. When an at-risk case (case 1) became clustered with a subsequent case (case 2) to result in a cluster formation event, the latter (case 2) was by definition *not* at-risk because it was already clustered with a prior case (case 1) on detection date. It is possible that a GAS lineage begins in one population and subsequently spreads to other populations. Therefore, iGAS cases were allowed to form clusters with any other cases in the surveillance areas to accommodate potential transmission between different populations and geographic regions. We identified at-risk cases based on a conserved approach that required at least 1 year trace back time to allow adequate comparisons with prior cases. For example, the at-risk status of an iGAS case with detection on January 1, 2015 was unevaluable because there were no prior cases in the dataset to compare with. Therefore, at-risk cases analyzed in this study included none of the 2015 iGAS cases. Similarly, we identified cluster formation events based on a conserved approach that required at least 1 year trace forward time to allow adequate observation of event occurrence. At-risk cases analyzed in this study included none of the 2019 iGAS cases. As a result, all at-risk cases had detection dates within 2016-2018 (Figure 1).

Statistical analysis

Cluster formation incidence and hazard

We calculated the proportion of at-risk cases that had cluster formation events among all at-risk cases as a descriptive statistic. The 95% confidence intervals (CIs) were estimated using the Wilson score interval method.

The Kaplan–Meier estimator implemented in the R “survival” package was used to estimate the cumulative incidence (risk) of cluster formation as a function of time. In the analysis of cluster

121 formation, it is important to consider right-censoring, where the trace forward period ends before the
122 cluster formation event occurs. Time-to-cluster formation analysis can appropriately process censored
123 data, ensuring that the information they provide is fully utilized. For predictive modeling, time-to-event
124 models can predict not just if an event will occur, but the risk of event as a function of time. This
125 temporal information is valuable for planning interventions, follow-up schedules, and managing
126 resources. The survival probability, which directly reflects risk, can be derived from the hazard function.
127 The survival function $S(t)$ is often calculated using the cumulative hazard $H(t)$ through the formula
128 $S(t)=\exp(-H(t))$. Risk is interpreted as the probability of a cluster formation event occurring within a
129 specified time period. It's a cumulative measure that integrates the hazard rate over time to provide the
130 likelihood of the event occurring by a certain point in time.

131

132 The cumulative incidence is the number of cluster formation events that occur during a given time
133 period divided by the number at-risk at the beginning of the period, which is the survival (being cluster
134 formation free) subtracted from one. The Kaplan-Meier estimator equation is: $S(t) = \text{Product}[(n_i - d_i) /$
135 $n_i]$ for all times t_i less than or equal to t . Here, $S(t)$ is the probability of free of cluster formation event
136 until time t , n_i is the number of subjects at risk just before time t_i , and d_i is the number of at-risk cases
137 who had a cluster formation event at time t_i . The product is calculated over all distinct times t_i where an
138 event occurs up to time t . To calculate the 95% confidence interval of the Kaplan-Meier estimator, we
139 used Greenwood's formula¹¹ for the estimator's variance $V(S(t))$, then apply the log(-log) transformation
140 for the survival probabilities $S(t)$. The formula for the variance is: $V(S(t)) = S(t)^2 * \text{Sum}[(d_i / (n_i * (n_i -$
141 $d_i)))]$ for all $t_i \leq t$, Where d_i is the number of events at time t_i , and n_i is the number of subjects at risk
142 just prior to t_i . The 95% confidence interval for $\log(-\log(S(t)))$ can be calculated using the normal
143 distribution, typically with a Z value of 1.96 for the 95% confidence level. Finally, back-transform these
144 intervals to the original scale to get the 95% CI for the Kaplan-Meier estimate.

145

146 The hazard function for cluster formation event was estimated using a nonparametric method
147 implemented in the R “bshazard” package ¹². This method estimates the hazard function without
148 assuming a specific statistical form (e.g. exponential, Weibull) for the survival times using smoothed
149 estimates based on B-splines. Point-wise confidence band for the smooth estimate is computed for the
150 linear predictor with variance matrix, which is then transformed to the 95% confidence interval at the
151 hazard scale¹².

152

153 **Univariable analysis**

154 In univariable analysis, the Log-rank test was used to compare the survival distributions of two groups
155 based on the Kaplan-Meier estimator. For each of the 72 binary case features shown in Table S1, we
156 compared the feature-positive at-risk cases with the feature-negative at-risk cases using the Log-rank
157 test. It tested the null hypothesis that there is no difference between the two populations in the
158 probability of an event at any time point. The p-value in the Log-rank test is calculated by comparing the
159 observed number of events in each group to what would be expected if there were no difference between
160 the groups. The test statistic follows a chi-square distribution with one degree of freedom under the null
161 hypothesis. The calculation involves summing the differences between observed and expected events
162 across all time points, squaring this sum, and then dividing by the variance. This final value is then
163 referenced against the chi-square distribution to obtain the p-value. In univariable analysis, p-values
164 were corrected for multiple testing by using the Benjamini-Hochberg procedure to control the false
165 discovery rate at 0.05 and only corrected p-values are reported.

166

167 **Multivariable analysis**

168

169 In multivariable analysis, Cox regression was used to simultaneously assess the effect of multiple at-risk
170 case features on cluster formation risk after adjusting for potential confounding effects among the
171 features. The equation for the Cox model is $h(t)=h_0(t)\exp(\beta_1X_1+\beta_2X_2+\dots+\beta_pX_p)$, where: $h(t)$ is the
172 hazard at time t ; $h_0(t)$ is the baseline hazard, representing the hazard for an individual with baseline
173 values (typically zeroes) of the covariates; $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients estimated from the data,
174 which represent the log hazard ratio associated with a one-unit increase in the respective covariate X_1 ,
175 $X_2, \dots, X_p$; and $X_1, X_2, \dots, X_p$ are the covariates (predictor variables) for each individual.

176

177 Initially, all features with p-value <0.2 in the univariable analysis were considered for selection in the
178 multivariable Cox model as predictor variables. To address potential multicollinearity, we inspected the
179 correlation between pair-wise predictor variables and those strongly correlated with *emm* type (Cramer's
180 $V > 0.8$)¹³ were excluded. The remaining predictor variables did not show strong correlation with each
181 other and were all incorporated into initial full model (Table S6). Stepwise selection¹⁴ was then
182 performed using the Akaike Information Criteria to select the most relevant variables that resulted in the
183 best fit of model and these selected variables were used in the final multivariable Cox model. The
184 proportional hazards assumption of the Cox regression models was evaluated using the Schoenfeld test
185 based on weighted residuals¹⁵. Based on the fitted final Cox model, the confounder-adjusted survival
186 curves and 95% confidence intervals were calculated using the Corrected Group Prognosis method¹⁶
187 implemented in the R 'adjustedCurves' package.

188

189 In addition, we conducted the multivariable analysis using a mixed-effects Cox model additionally
190 incorporating long-term care facility identification number (LTCF ID) and county as random effects,
191 accounting for clustering of LTCF cases in the same facility and clustering of people experiencing

homeless (PEH) and Persons Who Inject Drugs (PWID) cases in the same county, respectively (Table S7).

Machine learning models

Next, we utilized machine learning models as a predictive tool for the risk of cluster formation. Machine learning models generally do not require the proportional hazards assumption that is central to the Cox model. This allows them to model survival outcomes even when the hazards are not proportional over time, offering greater flexibility in capturing the underlying risk structure. Machine learning models can also automatically detect and model non-linear relationships and interactions between variables without requiring them to be explicitly defined in the model.

We divided all at-risk cases into 2 cohorts (derivation and validation) of equal size by random selection. Machine learning model comparison, model tuning, model training, and leave-one-out cross validation (LOOCV) were all conducted based on the derivation cohort data and utilized no information from the validation cohort.

A total of 13 machine learning model were included in the model comparison analysis, including the Cox proportional hazard model as the baseline model for comparison. For all models, the task was to predict the risk of cluster formation over time using at-risk case features in the deviation cohorts. The performance metrics used for comparison was the concordance index (c-Index) ¹⁷, which is a generalization of the area under the receiver operating characteristic (ROC) curve with $c=0.5$ for no discrimination power and $c=1$ for a perfect prediction. For each model, the derivation cohort data was randomly divided into 10 equal sized parts (folds) to be used in a 10-fold cross validation. For each

iteration, one of the 10 folds is used as the test set, and the remaining 9 folds are combined to form the training set. The model is trained on the 9 folds of the training set and then evaluated on the test fold, measuring the C-index. This process is repeated 10 times, with each of the 10 folds serving as the test set exactly once. The results from the 10 iterations are used to calculate the mean C-index and its 95% confidence intervals. The model with the highest mean C-index (random survival forest¹⁸) was selected for further analysis.

In the comparison analysis using the training dataset this study, several machine learning algorithms showed higher predictive performance (C-index) compared to the Cox model, although the improvement was modest (Table S8). The random survival forest model not only showed the highest mean C-index but also showed the highest lower bound of the 95% CI among all machine learning models.

Tuning of random survival forest model hyper-parameters was performed using the “tune” function in the “randomForestSRC” R package based on subsampling of the derivation cohort data. Then the model was then trained using the entire derivation cohort data and the tuned hyper-parameters (nodesize=20, mtry=20).

Model tuning and training using the derivation cohort served purpose to identify the best fitted variables and model structure. The trained model can be subsequently validated using the validation cohort. However, only using significant features identified from the univariable analysis of all at-risk cases in model training would be unsuitable because information from the validation cohort also contributed to the identification of these features, thus leading to data leakage. In building the machine-learning model, no information in the validation cohort was used in model training, including selecting important features.

240

241 The output of the random survival forests machine-learning model are matrices containing the
242 cumulative hazard functions (CHF) and survival function. Each row corresponds to an individual's
243 ensemble CHF or survival function evaluated at each time point. The output also provides a vector of
244 mortality values representing estimated risk for each individual calibrated to the scale of the number of
245 events¹⁸ and the mortality values were used as the risk scores. The model is set up to classify an at-risk
246 iGAS case into one of the three risk groups (low, moderate, and high), indicating different levels of
247 cluster formation risk after case detection date, based on the predicted risk score. The predicted risk
248 score is generated by the model using only at-risk case information that were available on the detection
249 date and no subsequent information.

250

251

252 Based on the risk score distribution of the random survival forests model fitted to the derivation cohort
253 data (n=1382), we selected risk score thresholds to define three risk groups (Figure S9). An at-risk case
254 with risk score below the 25th percentile of the derivation cohort risk scores (≤ 12.8) was defined as “low
255 risk” case while a case with risk score above the 75th percentiles (> 34.5) was defined as “high risk”
256 case. A case with risk score between the 25th and the 75th percentiles ($(12.8-34.5]$) was defined as
257 “moderate risk” case.

258

259 The predictive performance of the trained model was validated using the validation cohort dataset
260 (n=1382). For each at-risk case in the validation cohort, cumulative hazard function (CHF), survival
261 function, and risk score were predicted using the trained model and case features. C-index was used as
262 the predictive performance metrics. The 95% confidence interval of C-index was estimated with the use
263 of the nonparametric percentile bootstrap method with 1000 repetitions. The validation cohort were

264 stratified into three risk groups based on the pre-defined risk score threshold (Figures S9, S14).
265 Observed cluster formation risk for each risk group was calculated and compared with the other groups.
266
267 The AUC (Area Under the Curve) of ROC (Receiver Operating Characteristic) curve was used as an
268 alternative predictive performance metrics based on the risk scores and binary outcomes (cluster
269 formation yes/no) at 30-, 90-, 180-day, and 360-day after detection (Table S10). The AUC appeared to
270 peak for the 90-day outcome and had lower values for prediction cluster formation in longer periods.
271
272 The derivation and validation cohorts were both derived from the surveillance population in ABCs.
273 Therefore, the validation represents an internal validation and caution should be taken when generalizing
274 findings to populations outside ABCs.
275
276 In addition, we carried out an exhaustive Leave-One-Out Cross-Validation (LOOCV) for the random
277 survival forests model using the derivation cohort dataset. In LOOCV, the model is trained and validated
278 across as many iterations as there are data rows in the dataset. In each iteration of LOOCV, a single data
279 point is selected as the test set, and the remaining data points are used as the training set. This process is
280 repeated 1382 times, with each of the 1382 observations serving as the test set once. The model output
281 from all iterations was combined and compared with the true event data to calculate the performance
282 metrics C-index. The derivation cohort were stratified into three risk groups based on the pre-defined
283 risk score threshold (Figures S9, S11). Observed cluster formation risk for each risk group was
284 calculated and compared with the other groups.
285
286 All analyses were performed using the R software version 3.4.3 ¹⁹. In univariable analysis, P values
287 were corrected for multiple testing by using the Benjamini-Hochberg procedure to control the false

288 discovery rate at 0.05 and only the corrected p-values were reported. All p values were two-sided and p
289 values less than 0.05 were considered statistically significant.

290

291 **Grouping of *emm* types**

292 *emm* types 1, 89, 12, 28, 4, 3, 11, 77, 6 and 92 are the 10 most common *emm* types among the at-risk
293 isolates, ranking from high to low. These 10 *emm* types were treated as individual types while all other
294 *emm* types were grouped together as “Other than the 10 most common types listed above” (“Other”). In
295 univariable analysis, each case feature (including isolate *emm* type) is a binary variable that was either
296 positive (present) or negative (absent) in an iGAS case. In multivariable analysis *emm* type was
297 incorporated as a multi-level categorical variable. *emm*89 was selected as the reference group in
298 multivariable regression because it is a common *emm* type and showed an intermediate level cluster
299 formation event among at-risk cases (19%).

300

301 **Proportional hazard**

302

303 While an evolving hazard rate was observed (Figure 2C), it was not necessarily inconsistent with the
304 assumption of proportional hazards. The Cox proportional hazards model requires that the ratio of the
305 hazard rates between any two individuals or groups must remain constant over time but does not require
306 the hazard rate to be constant over time for each individual or group. The proportional hazards
307 assumption was evaluated using the Schoenfeld test and no evidence of violation of the proportional-
308 hazards assumption was observed (Figure S6; $p > 0.05$ for all covariates).

309

310 **Genomic clusters and transmission**

311

312 Direct transmission of iGAS is a rare occurrence²⁰, although transmission to health care workers^{21,22} or
313 family members^{23,24} have been documented. GAS is most commonly spread from people with less
314 severe GAS diseases (e.g., pharyngitis and impetigo) or asymptomatic carriers to susceptible individuals
315 through direct person-to-person transmission, and iGAS infection primarily arises as an infrequent but
316 severe sequel of non-invasive skin/soft tissue infection or pharyngeal infection. Therefore, it is important
317 to note that transmission of a GAS lineage in a population resulting in two iGAS infections that form a
318 genomic cluster is not necessarily direct transmission of an iGAS infection. It is possible that an
319 epidemiological cluster is caused by genetically distinct GAS strains. For example, Alcolea-Medina et
320 al. reported that *emm12* and *emm1* were the predominant strains in an ongoing GAS outbreak in London,
321 United Kingdom, in December 2022²⁵. Distinct iGAS genomic clusters that overlapped in time and
322 occurred in geographically proximal locations have also been documented²⁶. The presence or absence of
323 epidemiologically overlapping iGAS genomic clusters could help us better understand the transmission
324 dynamics during an iGAS outbreak. For instance, if only genomically closely related isolates are found
325 in an outbreak, it could suggest closely connected transmission or a common source. In contrast, a lack
326 of genomic clustering in an outbreak could indicate multiple sources or independent introductions of
327 different strains.

328

329 The iGAS genomic clustering in this study was identified based on bacterial genomic distance.
330 However, the clustering process was clearly not driven by bacterial genomic factors alone. Predictors of
331 new clustering such as LTCF, PEH, and PWID and seasonality are inherently epidemiological factors.
332 The greatest potential benefit of predicting cluster formation risk is to inform prevention strategies and
333 priorities. By understanding which iGAS cases are more likely to result in clustered iGAS infections
334 before the onset of the clusters, public health resources can be allocated more effectively. This includes

335 targeting surveillance efforts, better infection control practices in healthcare settings, and developing
336 educational campaigns to areas with the highest risk cases.

339 **Table S1. Features and cluster formation among 2764 at-risk iGAS cases at 3-SNV threshold**

Feature ^a	At-risk cases (N=2764) ^b	Cluster formation events among at-risk cases No. (%) ^c	95% Confidence Interval (CI)	Log-rank test P value ^d
STRAIN				
<i>emm</i> type				
1	398(14.4%)	130(32.7%)	28.2% to 37.4%	<0.0001
89	358(13.0%)	68(19.0%)	15.3% to 23.4%	0.0528
12	274(9.9%)	69(25.2%)	20.4% to 30.6%	0.8098
28	242(8.8%)	32(13.2%)	9.5% to 18.1%	0.0004
4	195(7.1%)	27(13.8%)	9.7% to 19.4%	0.0031
11	125(4.5%)	31(24.8%)	18.1% to 33.0%	0.8098
77	125(4.5%)	28(22.4%)	16.0% to 30.5%	0.8098
3	116(4.2%)	42(36.2%)	28.0% to 45.3%	0.0019
6	85(3.1%)	31(36.5%)	27.0% to 47.1%	0.0099
92	60(2.2%)	26(43.3%)	31.6% to 55.9%	0.0002
Other than the 10 most common types listed above	786(28.4%)	172(21.9%)	19.1% to 24.9%	0.3272
<i>emm</i> Pattern^e				
A-C	884(32.0%)	274(31.0%)	28.0% to 34.1%	<0.0001
D	166(6.0%)	43(25.9%)	19.8% to 33.1%	0.6128
E	1664(60.2%)	337(20.3%)	18.4% to 22.3%	<0.0001
Antimicrobial Susceptibility				
Erythromycin Non-susceptible	543(19.6%)	131(24.1%)	20.7% to 27.9%	0.8098
Clindamycin Non-susceptible	502(18.2%)	127(25.3%)	21.7% to 29.3%	0.4613
Tetracycline Non-susceptible	585(21.2%)	153(26.2%)	22.8% to 29.9%	0.179
Levofloxacin Non-susceptible	56(2.0%)	13(23.2%)	14.1% to 35.8%	0.9918
Pyrogenic Exotoxin Genes				
<i>speA</i>	549(19.9%)	177(32.2%)	28.5% to 36.3%	<0.0001
<i>speC</i>	1568(56.7%)	350(22.3%)	20.3% to 24.4%	0.0944
<i>speG</i>	2407(87.1%)	593(24.6%)	23.0% to 26.4%	0.0099
<i>speH</i>	612(22.1%)	152(24.8%)	21.6% to 28.4%	0.67
<i>speI</i>	551(19.9%)	139(25.2%)	21.8% to 29.0%	0.566
<i>speJ</i>	859(31.1%)	229(26.7%)	23.8% to 29.7%	0.032
<i>speK</i>	383(13.9%)	100(26.1%)	22.0% to 30.7%	0.3578
<i>speL</i>	182(6.6%)	49(26.9%)	21.0% to 33.8%	0.4543
<i>speM</i>	179(6.5%)	47(26.3%)	20.4% to 33.2%	0.6055
<i>ssa</i>	458(16.6%)	85(18.6%)	15.3% to 22.4%	0.0154

<i>smeZ</i>	2440(88.3%)	584(23.9%)	22.3% to 25.7%	0.6966
Surface Protein gene				
Fibronectin-binding protein type A (FBAA) <i>fbaA</i>	2233(80.8%)	509(22.8%)	21.1% to 24.6%	0.0514
Fibronectin-binding protein F2 (PRTF2) <i>prtF</i>	1736(62.8%)	383(22.1%)	20.2% to 24.1%	0.0202
<i>Alp</i> family adhesin (R28)	427(15.4%)	65(15.2%)	12.1% to 18.9%	0.0001
Fibronectin-binding protein I (SFB1) <i>sfbl</i>	1954(70.7%)	422(21.6%)	19.8% to 23.5%	0.0001
Serum opacity factor (SOF) <i>sof</i>	1739(62.9%)	357(20.5%)	18.7% to 22.5%	<0.0001
Other Virulence Factors				
<i>hasA</i> hyaluronic acid synthetase (CAPSULE)	1544(55.9%)	442(28.6%)	26.4% to 30.9%	<0.0001
Active NADase (NADASE_D330G)	2223(80.4%)	531(23.9%)	22.2% to 25.7%	0.8835
Pnga 3 - Clade 3 up-regulated promoter of the <i>nga</i> operon (PNGA3)	1559(56.4%)	371(23.8%)	21.7% to 26.0%	0.936
<i>rocA</i> null mutations (ROCA)	118(4.3%)	42(35.6%)	27.5% to 44.6%	0.0031
Virulence associated DNase (SDA1)	703(25.4%)	199(28.3%)	25.1% to 31.7%	0.0068
Streptococcal inhibitor of complement (SIC)	646(23.4%)	197(30.5%)	27.1% to 34.2%	<0.0001
PATIENT				
Male Sex	1497(54.2%)	351(23.4%)	21.4% to 25.7%	0.8098
Female Sex	1267(45.8%)	305(24.1%)	21.8% to 26.5%	0.8098
Age, years (Median, interquartile range)	56, 37 to 53	NA	NA	NA
Age Group, years				
<18	221(8.0%)	54(24.4%)	19.2% to 30.5%	0.8098
18-34	386(14.0%)	97(25.1%)	21.1% to 29.7%	0.6457
35-49	489(17.7%)	130(26.6%)	22.9% to 30.7%	0.218
50-64	692(25.0%)	167(24.1%)	21.1% to 27.5%	0.9022
65-74	455(16.5%)	92(20.2%)	16.8% to 24.2%	0.1449
>=75	521(18.8%)	116(22.3%)	18.9% to 26.0%	0.5582
Race				
White	1859(67.3%)	447(24.0%)	22.2% to 26.0%	0.8098
Nonwhite	667(24.1%)	150(22.5%)	19.5% to 25.8%	0.6275
Unknown	238(8.6%)	59(24.8%)	19.7% to 30.6%	0.8098
Syndrome				
Cellulitis	975(35.3%)	252(25.8%)	23.2% to 28.7%	0.1286
Bacteremia without focus	594(21.5%)	130(21.9%)	18.7% to 25.4%	0.4262
Pneumonia	378(13.7%)	86(22.8%)	18.8% to 27.2%	0.8098
Streptococcal Toxic Shock	72(2.6%)	18(25.0%)	16.4% to 36.1%	0.9022

Syndrome (STSS)				
Necrotizing fasciitis	142(5.1%)	33(23.2%)	17.1% to 30.8%	0.8739
Residence				
Private	2433(88.0%)	543(22.3%)	20.7% to 24.0%	<0.0001
Long term care facility	136(4.9%)	43(31.6%)	24.4% to 39.8%	0.0442
Homeless (PEH)	141(5.1%)	54(38.3%)	30.7% to 46.5%	<0.0001
Comorbidities				
Diabetes	802(29.0%)	189(23.6%)	20.8% to 26.6%	0.8952
Obesity	537(19.4%)	114(21.2%)	18.0% to 24.9%	0.2364
Chronic skin disease	318(11.5%)	71(22.3%)	18.1% to 27.2%	0.6917
Alcohol abuse	229(8.3%)	70(30.6%)	25.0% to 36.8%	0.0225
Cirrhosis	139(5.0%)	47(33.8%)	26.5% to 42.0%	0.0075
People who inject drugs (PWID)	204(7.4%)	86(42.2%)	35.6% to 49.0%	<0.0001
HIV	56(2.0%)	17(30.4%)	19.9% to 43.3%	0.4978
ENVIROMENT				
Detection Season				
Fall-Winter (October-March)	1565(56.6%)	410(26.2%)	24.1% to 28.4%	0.0012
Spring-Summer (April-September)	1199(43.4%)	246(20.5%)	18.3% to 22.9%	0.0012
Detection Region ^f				
Northeast (CT, NY, MD)	638(23.1%)	151(23.7%)	20.5% to 27.1%	0.9918
Southeast (GA, TN)	626(22.6%)	140(22.4%)	19.3% to 25.8%	0.4566
Midwest (MN)	459(16.6%)	119(25.9%)	22.1% to 30.1%	0.3976
Southwest (NM)	260(9.4%)	78(30.0%)	24.8% to 35.8%	0.0356
West (CA, CO, OR)	781(28.3%)	168(21.5%)	18.8% to 24.5%	0.2364

- a. Each feature is either positive (present) or negative (absent) in an iGAS case.
- b. Number of at-risk cases that were feature positive.
- c. Cluster formation events among feature-positive cases
- d. Log-rank test to test the hypothesis that two groups (feature positive vs. feature negative) have the same survival functions. P-values were corrected for multiple testing by using the Benjamini-Hochberg procedure to control the false discovery rate at 0.05.
- e. *emm* pattern refers to a system used to categorize Group A Streptococcus (GAS) strains based on the composition of *emm*-family genes along with their respective peptidoglycan-spanning (PG) domain subfamily types²⁷
- f. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota, NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

Table S2. Features and cluster formation among 4348 at-risk iGAS cases at 1-SNV threshold

Feature ^a	At-risk cases ^b	Cluster formation events, No. (%) ^c	95% Confidence Interval (CI)	Log-rank test P value ^d
STRAIN				
<i>emm</i> type				
1	676	130(19.2%)	28.2% to 37.4%	0.0014
89	458	56(12.2%)	15.3% to 23.4%	0.2442
12	351	38(10.8%)	20.4% to 30.6%	0.0926
28	286	14(4.9%)	9.5% to 18.1%	<0.0001
4	228	19(8.3%)	9.7% to 19.4%	0.0246
11	193	26(13.5%)	18.1% to 33%	0.7194
77	161	23(14.3%)	16% to 30.5%	0.9185
3	228	38(16.7%)	28% to 45.3%	0.5587
6	136	15(11%)	27% to 47.1%	0.3691
92	222	58(26.1%)	31.6% to 55.9%	<0.0001
Other than the 10 most common types listed above	1409	220(15.6%)	19.1% to 24.9%	0.3142
<i>emm</i> Pattern^e				
A-C	1403	222(15.8%)	28% to 34.1%	0.2442
D	328	66(20.1%)	19.8% to 33.1%	0.0138
E	2557	347(13.6%)	18.4% to 22.3%	0.0536
Antimicrobial Susceptibility				
Erythromycin Non-susceptible	898	146(16.3%)	20.7% to 27.9%	0.2442
Clindamycin Non-susceptible	853	144(16.9%)	21.7% to 29.3%	0.1043
Tetracycline Non-susceptible	1000	182(18.2%)	22.8% to 29.9%	0.0018
Levofloxacin Non-susceptible	70	11(15.7%)	14.1% to 35.8%	0.8815
Pyrogenic Exotoxin Genes				
<i>speA</i>	907	157(17.3%)	28.5% to 36.3%	0.0401
<i>speC</i>	2354	294(12.5%)	20.3% to 24.4%	0.0001
<i>speG</i>	3898	589(15.1%)	23% to 26.4%	0.0401
<i>speH</i>	1027	147(14.3%)	21.6% to 28.4%	0.8435
<i>speI</i>	942	139(14.8%)	21.8% to 29%	0.9185
<i>speJ</i>	1454	233(16%)	23.8% to 29.7%	0.1504
<i>speK</i>	585	70(12%)	22% to 30.7%	0.1153
<i>speL</i>	304	50(16.4%)	21% to 33.8%	0.4984
<i>speM</i>	298	47(15.8%)	20.4% to 33.2%	0.7021
<i>ssa</i>	639	75(11.7%)	15.3% to 22.4%	0.0669
<i>smeZ</i>	3723	540(14.5%)	22.3% to 25.7%	0.6243
Surface Protein gene				
Fibronectin-binding protein type A (FBAA) <i>fbaA</i>	3566	543(15.2%)	21.1% to 24.6%	0.0651

Fibronectin-binding protein F2 (PRTF2) <i>prtF</i>	2681	369(13.8%)	20.2% to 24.1%	0.1021
<i>Alp</i> family adhesin (R28)	503	33(6.6%)	12.1% to 18.9%	<0.0001
Fibronectin-binding protein I (SFB1) <i>sfb1</i>	2861	359(12.5%)	19.8% to 23.5%	<0.0001
Serum opacity factor (SOF) <i>sof</i>	2710	367(13.5%)	18.7% to 22.5%	0.0317
Other Virulence Factors				
<i>hasA</i> hyaluronic acid synthetase (CAPSULE)	2813	487(17.3%)	26.4% to 30.9%	<0.0001
Active NADase (NADASE_D330G)	3341	465(13.9%)	22.2% to 25.7%	0.0401
Pnga 3 - Clade 3 up-regulated promoter of the <i>nga</i> operon (PNGA3)	2213	305(13.8%)	21.7% to 26%	0.1910
<i>rocA</i> null mutations (ROCA)	231	38(16.5%)	27.5% to 44.6%	0.5971
Virulence associated DNase (SDA1)	1063	167(15.7%)	25.1% to 31.7%	0.4345
Streptococcal inhibitor of complement (SIC)	996	168(16.9%)	27.1% to 34.2%	0.0687
PATIENT				
Male Sex	2416	350(14.5%)	21.4% to 25.7%	0.8435
Age Group, years				
<18	309	46(14.9%)	19.2% to 30.5%	0.9185
18-34	613	106(17.3%)	21.1% to 29.7%	0.1043
35-49	801	125(15.6%)	22.9% to 30.7%	0.5971
50-64	1138	165(14.5%)	21.1% to 27.5%	0.9185
65-74	701	84(12%)	16.8% to 24.2%	0.0838
>=75	786	111(14.1%)	18.9% to 26%	0.7819
Race				
White	2909	442(15.2%)	22.2% to 26%	0.2687
Nonwhite	1031	132(12.8%)	19.5% to 25.8%	0.1264
Unknown	408	63(15.4%)	19.7% to 30.6%	0.7634
Syndrome				
Cellulitis	1578	239(15.1%)	23.2% to 28.7%	0.6239
Bacteremia without focus	952	141(14.8%)	18.7% to 25.4%	0.9185
Pneumonia	608	89(14.6%)	18.8% to 27.2%	0.9657
Streptococcal Toxic Shock Syndrome (STSS)	96	13(13.5%)	16.4% to 36.1%	0.8435
Necrotizing fasciitis	217	27(12.4%)	17.1% to 30.8%	0.4874
Residence				
Private	3677	485(13.2%)	20.7% to 24%	<0.0001
Long term care facility	251	57(22.7%)	24.4% to 39.8%	0.0004
Homeless (PEH)	336	85(25.3%)	30.7% to 46.5%	<0.0001
Comorbidities				

Diabetes	1243	163(13.1%)	20.8% to 26.6%	0.1457
Obesity	804	96(11.9%)	18% to 24.9%	0.0575
Chronic skin disease	514	70(13.6%)	18.1% to 27.2%	0.66
Alcohol abuse	410	70(17.1%)	25% to 36.8%	0.2303
Cirrhosis	275	49(17.8%)	26.5% to 42%	0.2134
People who inject drugs (PWID)	424	106(25%)	35.6% to 49%	<0.0001
HIV	99	12(12.1%)	19.9% to 43.3%	0.6111
ENVIROMENT				
Detection Season				
Fall-Winter (October-March)	2449	422(17.2%)	24.1% to 28.4%	<0.0001
Spring-Summer (April-September)	1899	215(11.3%)	18.3% to 22.9%	<0.0001
Detection Region^f				
Northeast (CT, NY, MD)	973	138(14.2%)	20.5% to 27.1%	0.7838
Southeast (GA, TN)	849	102(12%)	19.3% to 25.8%	0.0401
Midwest (MN)	704	109(15.5%)	22.1% to 30.1%	0.5971
Southwest (NM)	456	87(19.1%)	24.8% to 35.8%	0.0233
West (CA, CO, OR)	1366	201(14.7%)	18.8% to 24.5%	0.9185

- a. Each feature is either positive (present) or negative (absent) in an iGAS case.
- b. Number of at-risk cases that were feature positive.
- c. Cluster formation events among feature-positive cases
- d. Log-rank test to test the hypothesis that two groups (feature positive vs. feature negative) have the same survival functions. P-values were corrected for multiple testing by using the Benjamini-Hochberg procedure to control the false discovery rate at 0.05.
- e. *emm* pattern refers to a system used to categorize Group A Streptococcus (GAS) strains based on the composition of *emm*-family genes along with their respective peptidoglycan-spanning (PG) domain subfamily types²⁷
- f. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota, NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

Table S3. Features and cluster formation among 2239 at-risk iGAS cases at 5-SNV threshold

Feature ^a	At-risk cases ^b	Cluster formation events, No. (%) ^c	95% Confidence Interval (CI)	Log-rank test P value ^d
STRAIN				
<i>emm</i> type				
1	292	104(35.6%)	30.3% to 41.3%	0.0013
89	313	67(21.4%)	17.2% to 26.3%	0.0822
12	222	63(28.4%)	22.9% to 34.6%	0.7393
28	213	39(18.3%)	13.7% to 24%	0.0208
4	172	31(18%)	13% to 24.4%	0.028
11	103	31(30.1%)	22.1% to 39.5%	0.6462
77	103	29(28.2%)	20.4% to 37.5%	0.8264
3	76	32(42.1%)	31.6% to 53.3%	0.0037
6	68	28(41.2%)	30.3% to 53%	0.0089
92	32	15(46.9%)	30.9% to 63.6%	0.0104
Other than the 10 most common types listed above	645	156(24.2%)	21% to 27.6%	0.2537
<i>emm</i> Pattern^e				
A-C	669	229(34.2%)	30.7% to 37.9%	<0.0001
D	135	34(25.2%)	18.6% to 33.1%	0.8698
E	1386	330(23.8%)	21.6% to 26.1%	0.0013
Antimicrobial Susceptibility				
Erythromycin Non-susceptible	435	120(27.6%)	23.6% to 32%	0.7415
Clindamycin Non-susceptible	399	115(28.8%)	24.6% to 33.5%	0.4752
Tetracycline Non-susceptible	465	129(27.7%)	23.9% to 32%	0.6762
Levofloxacin Non-susceptible	45	14(31.1%)	19.5% to 45.7%	0.484
Pyrogenic Exotoxin Genes				
<i>speA</i>	412	148(35.9%)	31.4% to 40.7%	<0.0001
<i>speC</i>	1284	323(25.2%)	22.9% to 27.6%	0.1691
<i>speG</i>	1920	528(27.5%)	25.5% to 29.5%	0.0412
<i>speH</i>	486	138(28.4%)	24.6% to 32.6%	0.5304
<i>speI</i>	431	120(27.8%)	23.8% to 32.3%	0.6762
<i>speJ</i>	662	189(28.5%)	25.2% to 32.1%	0.2811
<i>speK</i>	305	97(31.8%)	26.8% to 37.2%	0.0635
<i>speL</i>	140	35(25%)	18.6% to 32.8%	0.8264
<i>speM</i>	137	33(24.1%)	17.7% to 31.9%	0.6762
<i>ssa</i>	382	94(24.6%)	20.6% to 29.2%	0.484
<i>smeZ</i>	1973	522(26.5%)	24.6% to 28.4%	0.8351
Surface Protein gene				
Fibronectin-binding protein type A (FBAA) <i>fbaA</i>	1822	467(25.6%)	23.7% to 27.7%	0.0966

Fibronectin-binding protein F2 (PRTF2) <i>prtF</i>	1423	359(25.2%)	23% to 27.5%	0.1409
<i>Alp</i> family adhesin (R28)	372	75(20.2%)	16.4% to 24.5%	0.0093
Fibronectin-binding protein I (SFB1) <i>sfb1</i>	1623	399(24.6%)	22.6% to 26.7%	0.0023
Serum opacity factor (SOF) <i>sof</i>	1446	348(24.1%)	21.9% to 26.3%	0.0017
Other Virulence Factors				
<i>hasA</i> hyaluronic acid synthetase (CAPSULE)	1176	360(30.6%)	28% to 33.3%	<0.0001
Active NADase (NADASE_D330G)	1819	488(26.8%)	24.8% to 28.9%	0.7963
Pnga 3 - Clade 3 up-regulated promoter of the <i>nga</i> operon (PNGA3)	1268	339(26.7%)	24.4% to 29.2%	0.985
<i>rocA</i> null mutations (ROCA)	78	33(42.3%)	32% to 53.4%	0.0036
Virulence associated DNase (SDA1)	539	165(30.6%)	26.9% to 34.6%	0.0716
Streptococcal inhibitor of complement (SIC)	492	165(33.5%)	29.5% to 37.8%	0.0013
PATIENT				
Male Sex	1220	329(27%)	24.6% to 29.5%	0.8351
Age Group, years				
<18	188	46(24.5%)	18.9% to 31.1%	0.6762
18-34	323	93(28.8%)	24.1% to 34%	0.5316
35-49	396	121(30.6%)	26.2% to 35.3%	0.1303
50-64	546	137(25.1%)	21.6% to 28.9%	0.4752
65-74	373	92(24.7%)	20.6% to 29.3%	0.6462
>=75	413	106(25.7%)	21.7% to 30.1%	0.8008
Race				
White	1503	417(27.7%)	25.5% to 30.1%	0.2148
Nonwhite	544	136(25%)	21.5% to 28.8%	0.5855
Unknown	192	42(21.9%)	16.6% to 28.2%	0.2811
Syndrome				
Cellulitis	791	226(28.6%)	25.5% to 31.8%	0.2540
Bacteremia without focus	478	114(23.8%)	20.2% to 27.9%	0.2811
Pneumonia	307	80(26.1%)	21.5% to 31.2%	0.8896
Streptococcal Toxic Shock Syndrome (STSS)	60	13(21.7%)	13.1% to 33.6%	0.5316
Necrotizing fasciitis	116	31(26.7%)	19.5% to 35.4%	0.9125
Residence				
Private	1991	509(25.6%)	23.7% to 27.5%	0.0021
Long term care facility	101	33(32.7%)	24.3% to 42.3%	0.1725
Homeless (PEH)	101	40(39.6%)	30.6% to 49.4%	0.0020
Comorbidities				

Diabetes	649	166(25.6%)	22.4% to 29.1%	0.6462
Obesity	443	106(23.9%)	20.2% to 28.1%	0.3262
Chronic skin disease	256	66(25.8%)	20.8% to 31.5%	0.8351
Alcohol abuse	177	55(31.1%)	24.7% to 38.2%	0.2811
Cirrhosis	103	35(34%)	25.6% to 43.6%	0.1409
People who inject drugs (PWID)	164	74(45.1%)	37.7% to 52.8%	<0.0001
HIV	48	21(43.8%)	30.7% to 57.7%	0.0682
ENVIROMENT				
Detection Season				
Fall-Winter (October-March)	1261	369(29.3%)	26.8% to 31.8%	0.0037
Spring-Summer (April-September)	978	226(23.1%)	20.6% to 25.9%	0.0037
Detection Region^f				
Northeast (CT, NY, MD)	557	148(26.6%)	23.1% to 30.4%	0.9524
Southeast (GA, TN)	520	145(27.9%)	24.2% to 31.9%	0.6762
Midwest (MN)	343	98(28.6%)	24% to 33.6%	0.5918
Southwest (NM)	200	60(30%)	24.1% to 36.7%	0.4752
West (CA, CO, OR)	619	144(23.3%)	20.1% to 26.8%	0.1243

368

- 369 a. Each feature is either positive (present) or negative (absent) in an iGAS case.
- 370 b. Number of at-risk cases that were feature positive.
- 371 c. Cluster formation events among feature-positive cases
- 372 d. Log-rank test to test the hypothesis that two groups (feature positive vs. feature negative) have
- 373 the same survival functions. P-values were corrected for multiple testing by using the Benjamini-
- 374 Hochberg procedure to control the false discovery rate at 0.05.
- 375 e. *emm* pattern refers to a system used to categorize Group A Streptococcus (GAS) strains based on
- 376 the composition of *emm*-family genes along with their respective peptidoglycan-spanning (PG)
- 377 domain subfamily types²⁷
- 378 f. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota,
- 379 NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

380
381
382

Table S4. Multivariable Cox regression analysis of variables associated with cluster formation at 1-SNV threshold.

Variable	Adjusted HR	95% CI	P-value
<i>emm</i> type			
1	1.69	1.2-2.39	0.003
89	Reference	-	-
12	0.97	0.64-1.48	0.903
28	0.47	0.26-0.84	0.011
4	0.8	0.47-1.35	0.399
11	1.19	0.74-1.91	0.474
77	1.16	0.71-1.9	0.555
3	1.49	0.97-2.29	0.072
6	1.07	0.6-1.91	0.817
92	2.26	1.52-3.35	<0.001
Other than the 10 most common types listed above	1.23	0.91-1.67	0.186
Exotoxin C			
Negative	Reference	-	-
Positive	0.82	0.67-1.01	0.065
Residence			
Private	Reference	-	-
Homeless (PEH)	1.55	1.18-2.04	0.002
Long term care facility (LTCF)	2.07	1.56-2.76	<0.001
Other than listed above	0.95	0.51-1.78	0.868
Alcohol abuse ^a	1.07	0.83-1.4	0.593
Cirrhosis ^a	1.11	0.82-1.51	0.481
Injection drug (PWID) ^a	1.73	1.36-2.21	<0.001
Detection Season			
Spring-Summer (Q2-Q3)	Reference	-	-
Fall-Winter (Q4-Q1)	1.59	1.34-1.87	<0.001
Detection Region ^b			
Northeast (CT, NY, MD)	Reference	-	-
Midwest (MN)	1.28	0.99-1.65	0.061
Southeast (GA, TN)	0.87	0.67-1.13	0.288
Southwest (NM)	1.22	0.92-1.62	0.16
West (CA, CO, OR)	0.95	0.76-1.19	0.654

a. Referent group consists of at-risk iGAS cases that did not present with this comorbidity.

383

384

- b.** CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota,
385 NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

386
387
388

Table S5. Multivariable Cox regression analysis of variables associated with cluster formation at 5-SNV threshold.

Variable	Adjusted HR	95% CI	P-value
emm type			
1	2.14	1.52-3.01	<0.001
89	Reference	-	-
12	1.48	1.04-2.1	0.027
28	0.88	0.59-1.31	0.519
4	0.8	0.52-1.23	0.314
11	1.36	0.88-2.1	0.16
77	1.48	0.95-2.3	0.085
3	2.59	1.68-4	<0.001
6	2.38	1.53-3.72	<0.001
92	1.99	1.11-3.57	0.022
Other than the 10 most common types listed above	1.13	0.84-1.51	0.409
Exotoxin C			
Negative	Reference	-	-
Positive	1.1	0.9-1.36	0.349
Residence			
Private	Reference	-	-
Homeless (PEH)	1.28	0.87-1.89	0.208
Long term care facility (LTCF)	1.69	1.18-2.42	0.004
Other than listed above	1.18	0.68-2.06	0.555
Alcohol abuse ^a	1.12	0.83-1.5	0.448
Cirrhosis ^a	1.2	0.83-1.72	0.331
Injection drug (PWID) ^a	2.14	1.59-2.87	<0.001
Detection Season			
Spring-Summer (Q2-Q3)	Reference	-	-
Fall-Winter (Q4-Q1)	1.32	1.12-1.57	0.001
Detection Region ^b			
Northeast (CT, NY, MD)	Reference	-	-
Midwest (MN)	1.19	0.92-1.54	0.184
Southeast (GA, TN)	1.04	0.83-1.31	0.725
Southwest (NM)	1.08	0.8-1.47	0.611
West (CA, CO, OR)	0.89	0.7-1.12	0.316

a. Referent group consists of at-risk iGAS cases that did not present with this comorbidity.

389

390
391

- b. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota,
NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

392
393

Table S6. Initial Cox regression model for variables associated with cluster formation.

Variable	Adjusted HR	95% CI	P-value
<i>emm</i> type			
1	2.44	1.75-3.4	<0.001
89	Reference	-	-
12	1.46	1.04-2.05	0.029
28	0.72	0.47-1.1	0.131
4	0.69	0.44-1.09	0.111
3	2.86	1.91-4.29	<0.001
11	1.28	0.83-1.97	0.264
77	1.33	0.85-2.09	0.21
6	2.23	1.39-3.58	0.001
92	2.34	1.52-3.62	<0.001
Other than the 10 most common types listed above	1.15	0.86-1.53	0.338
Exotoxin C			
Negative	Reference	-	-
Positive	1.18	0.97-1.45	0.104
Age Group, years			
<18	1.1	0.77-1.58	0.591
18-34	Reference	-	-
35-49	1.06	0.81-1.4	0.664
50-64	1.04	0.79-1.36	0.785
65-74	0.88	0.64-1.2	0.413
>=75	0.95	0.7-1.29	0.756
Syndrome			
Cellulitis	1.13	0.91-1.4	0.283
Bacteremia without focus	Reference	-	-
Pneumonia	0.82	0.62-1.09	0.169
Streptococcal Toxic Shock Syndrome (STSS)	0.98	0.59-1.61	0.935
Necrotizing fasciitis	0.81	0.55-1.2	0.299
Other than listed above	0.98	0.77-1.25	0.886
Residence			
Private	Reference	-	-
Homeless (PEH)	1.39	0.99-1.95	0.054
Long term care facility (LTCF)	1.91	1.36-2.66	<0.001

Other than listed above	1.46	0.88-2.43	0.14
Alcohol abuse ^a	1.24	0.95-1.62	0.118
Cirrhosis ^a	1.33	0.97-1.83	0.075
Injection drug (PWID) ^a	2.05	1.53-2.74	<0.001
Detection Season			
Spring-Summer (Q2-Q3)	Reference	-	-
Fall-Winter (Q4-Q1)	1.36	1.16-1.6	<0.001
Detection Region ^b			
Northeast (CT, NY, MD)	Reference	-	-
Midwest (MN)	1.21	0.95-1.54	0.129
Southeast (GA, TN)	0.9	0.72-1.14	0.398
Southwest (NM)	1.2	0.91-1.59	0.201
West (CA, CO, OR)	0.89	0.71-1.11	0.309

a. Referent group consists of at-risk iGAS cases that did not present with this comorbidity.

b. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota, NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

Table S7. Mixed effects Cox regression analysis of variables associated with cluster formation.

Variable ^a	Adjusted HR	95% CI	P-value
<i>emm</i> type			
1	2.38	1.71-3.31	<0.001
89	Reference	-	-
12	1.47	1.05-2.06	0.025
28	0.71	0.46-1.09	0.114
4	0.71	0.45-1.12	0.143
3	2.74	1.83-4.09	<0.001
11	1.28	0.83-1.98	0.257
77	1.35	0.86-2.11	0.188
6	2.23	1.39-3.59	0.001
92	2.29	1.49-3.52	<0.001
Other than the 10 most common types listed above	1.15	0.87-1.54	0.329
Exotoxin C			
Negative	Reference	-	-
Positive	1.18	0.96-1.45	0.107
Residence			
Private	Reference	-	-
Homeless (PEH)	1.43	1.02-2.01	0.036
Long term care facility (LTCF)	1.85	1.34-2.56	<0.001
Other than listed above	1.46	0.88-2.42	0.140
Alcohol abuse ^b	1.26	0.97-1.64	0.089
Cirrhosis ^b	1.35	0.99-1.85	0.060
Injection drug (PWID) ^b	2.07	1.58-2.72	<0.001
Detection Season			
Spring-Summer (Q2-Q3)	Reference	-	-
Fall-Winter (Q4-Q1)	1.34	1.14-1.57	<0.001
Detection Region ^c			
Northeast (CT, NY, MD)	Reference	-	-
Midwest (MN)	1.22	0.96-1.55	0.112
Southeast (GA, TN)	0.91	0.72-1.14	0.402
Southwest (NM)	1.18	0.89-1.56	0.245
West (CA, CO, OR)	0.89	0.71-1.11	0.294

- a. Variables listed in the table were incorporated in the mixed effects Cox model as fixed effects. LTCF ID and county as incorporated as random effects, accounting for clustering of LTCF cases in the same facility and clustering of PEH and PWID cases in the same county, respectively.
- b. Referent group consists of at-risk iGAS cases that did not present with this comorbidity.
- c. CA: California, CO: Colorado, CT: Connecticut, GA: Georgia, MD: Maryland, MN: Minnesota, NM: New Mexico, NY: New York, OR: Oregon, TN: Tennessee

407 **Table S8. Performance comparison among 13 machine learning model**

Algorithm	Description	Mean C-index	95% CI	Ref
surv.rfsrc	Survival Random Forest SRC	0.6071	0.5811-0.6330	28
surv.cforest	Survival Conditional Random Forest	0.6046	0.5783-0.6309	29
surv.glmnet	Generalized linear models with elastic net regularization	0.6027	0.5646-0.6408	30
surv.coxboost	Survival Cox Model with Likelihood Based Boosting Learner	0.6025	0.5624-0.6425	31
surv.aorsf	Accelerated Oblique Random Survival Forest	0.6007	0.5717-0.6298	32
surv.coxph	Cox Proportional Hazards Model	0.5993	0.5637-0.6348	33
surv.gbm	Survival Gradient Boosting Machine Model	0.5950	0.5610-0.6289	34
surv.glmboost	Boosted Generalized Linear Survival Model	0.5921	0.5578-0.6264	35
surv.ranger	Ranger Survival Learner	0.5850	0.5587-0.6114	36
surv.ctree	Survival Conditional Inference Tree	0.5731	0.5435-0.6027	29
surv.rpart	Recursive Partitioning and Regression Trees	0.5698	0.5537-0.5859	37
surv.cv_coxboost	Survival Cox Model with Cross-Validation Likelihood Based Boosting	0.5689	0.5528-0.5851	31
surv.xgboost	Extreme Gradient Boosting regression	0.5641	0.5384-0.5899	38

408

409

410 **Table S9** Predictive performance for cluster formation based on 1-SNV and 5-SNV threshold.

SNV threshold	C-index (Validation cohort)	95% CI
1-SNV	0.6325	0.6061-0.6631
5-SNV	0.6056	0.5740-0.6387

411

412 **Table S10** Predictive performance for binary outcome (cluster formation yes/no) at 30-, 90-, 180-day,
413 and 360-day after detection

End Point	AUC (Validation cohort)	95% CI
30-Day	0.6813	0.6223-0.7404
90-Day	0.6856	0.6452-0.7261
180-Day	0.6557	0.6172-0.6942
360-Day	0.6467	0.6111-0.6822

414

415

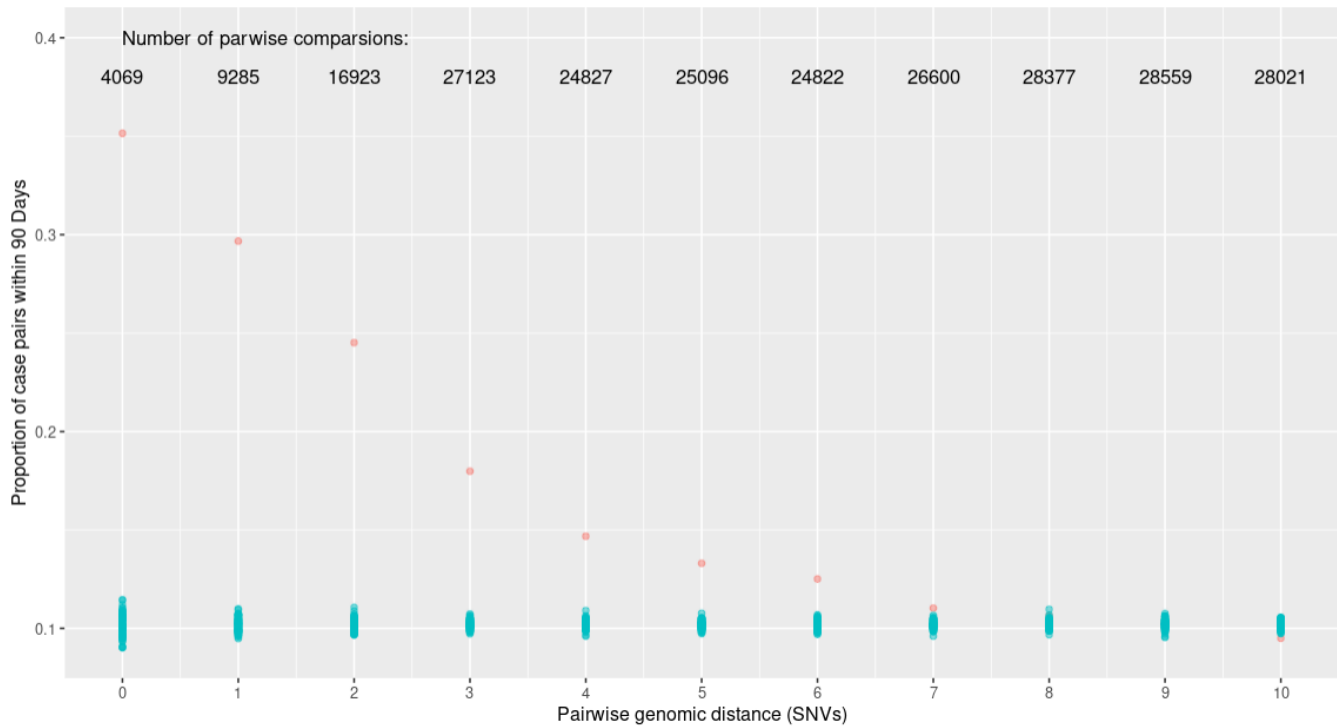
416 **Table S11. Within-cluster genomic differences by SNV- and time-cutoff**

SNV cutoff	Time cutoff	Mean pairwise SNV within cluster	95% CI
≤ 1	≤ 0.5 year	0.67	0.66-0.68
≤ 1	≤ 1 year	0.69	0.68-0.69
≤ 1	≤ 1.5 year	0.69	0.68-0.7
≤ 3	≤ 0.5 year	1.98	1.96-1.99
≤ 3	≤ 1 year	2.06	2.05-2.07
≤ 3	≤ 1.5 year	2.11	2.10-2.12
≤ 5	≤ 0.5 year	2.89	2.88-2.91
≤ 5	≤ 1 year	3.01	3.00-3.02
≤ 5	≤ 1.5 year	3.11	3.10-3.12

417

418 SUPPLEMENTARY FIGURES

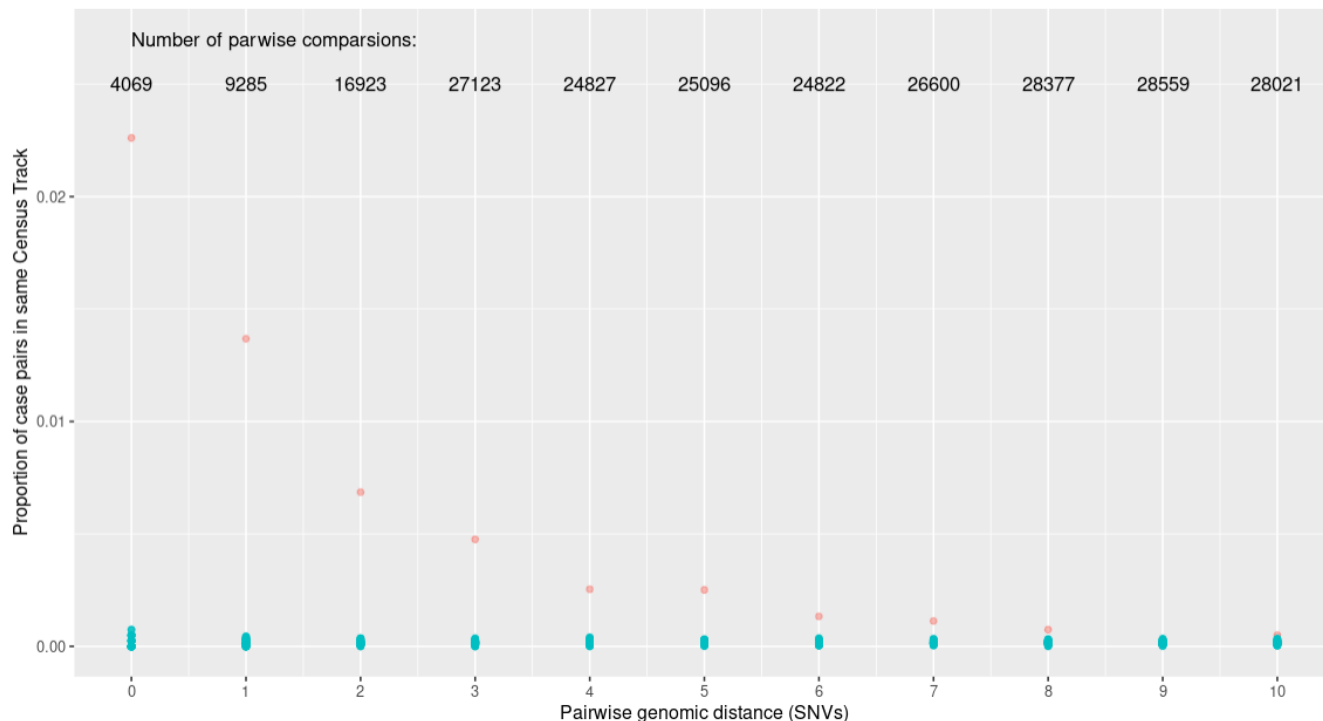
419 Figure S1



420
421 Figure S1. Pairwise genomic distances between all iGAS isolates of the same *emm* type were calculated
422 and combined. For a range of genomic distance thresholds, we calculated the proportion of pairwise
423 comparisons in which both iGAS cases occurred within a 90-day period among all pairwise comparisons
424 with the indicated genomic distance. The resulting proportions are plotted as the red points, which
425 appear to follow approximately exponential decay as genomic distance thresholds increase. The green
426 points represent the outcome of 100 permutations in which the same statistic was calculated when the
427 isolates detection dates were randomized.

428 Figure S2

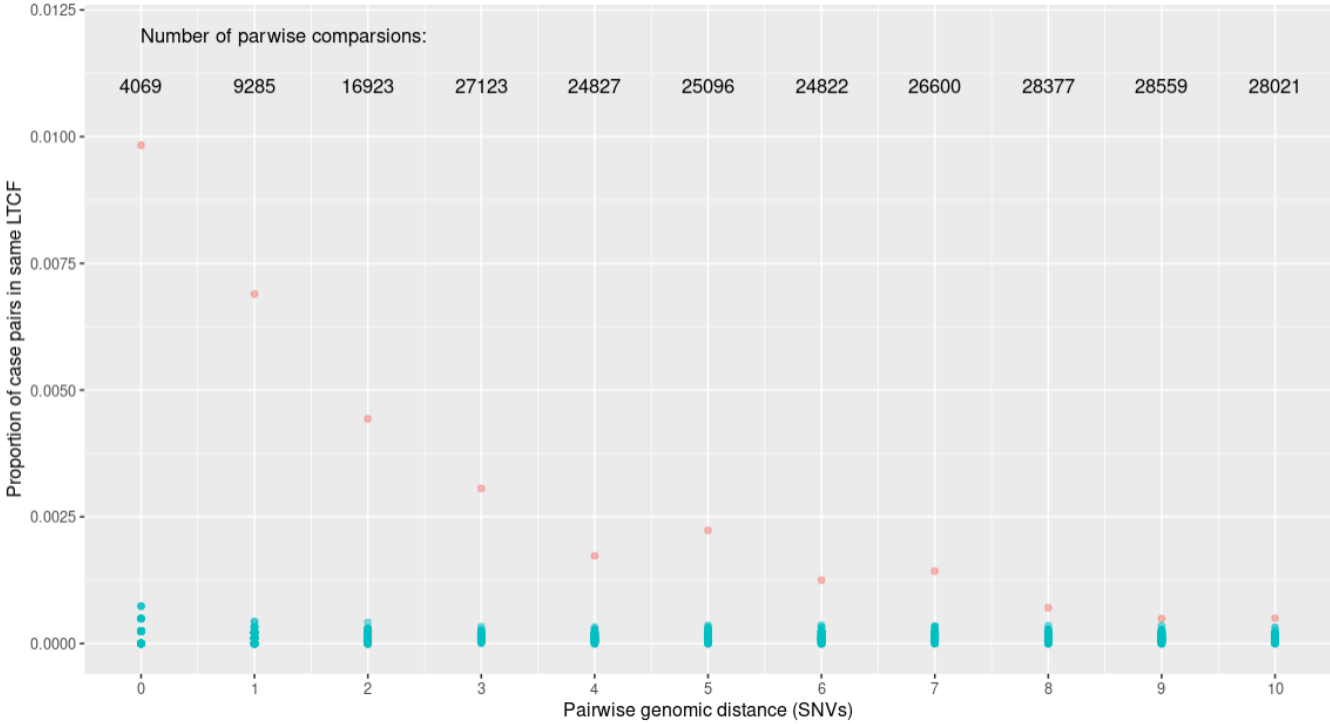
429



430

431 Figure S2. Pairwise genomic distances between all iGAS isolates of the same *emm* type were calculated
 432 and combined. For a range of genomic distance thresholds, we calculated the proportion of pairwise
 433 comparisons in which both iGAS cases occurred within same census tract (a geographic region defined
 434 for the purpose of taking a census in US with a population ranging from 8,000 to 1,200 people) among
 435 all pairwise comparisons with the indicated genomic distance. The resulting proportions are plotted as
 436 the red points, which appear to follow approximately exponential decay as genomic distance thresholds
 437 increases. The green points represent the outcome of 100 permutations in which the same statistic was
 438 calculated when the census tracts of the iGAS cases were randomized.

439 Figure S3

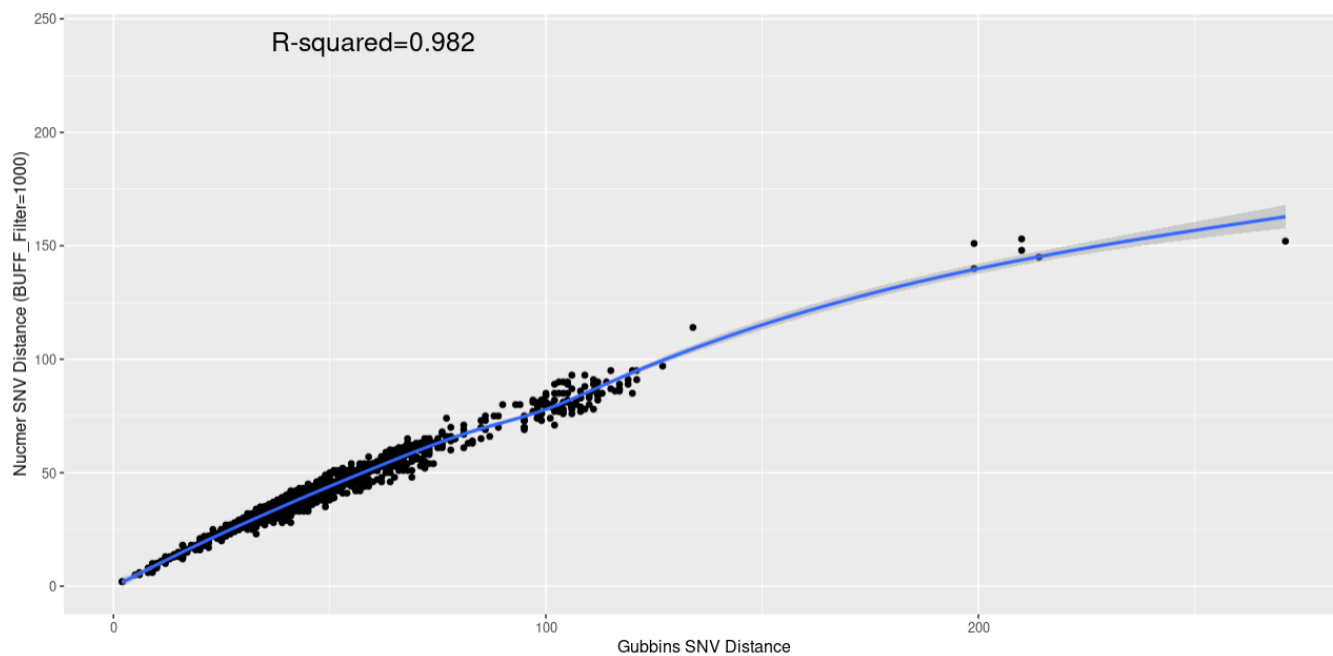


440

441 Figure S3. Pairwise genomic distances between all iGAS case isolates of the same *emm* type were
442 calculated and combined. For a range of genomic distance thresholds, we calculated the proportion of
443 pairwise comparisons in which both iGAS cases occurred within same long term care facility (LTCF)
444 among all pairwise comparisons with the indicated genomic distance. The resulting proportions are
445 plotted as the red points, which appear to follow approximately exponential decay as genomic distance
446 thresholds increases. The green points represent the outcome of 100 permutations in which the same
447 statistic was calculated when the LTCF IDs of the iGAS cases were randomized.

448 Figure S4

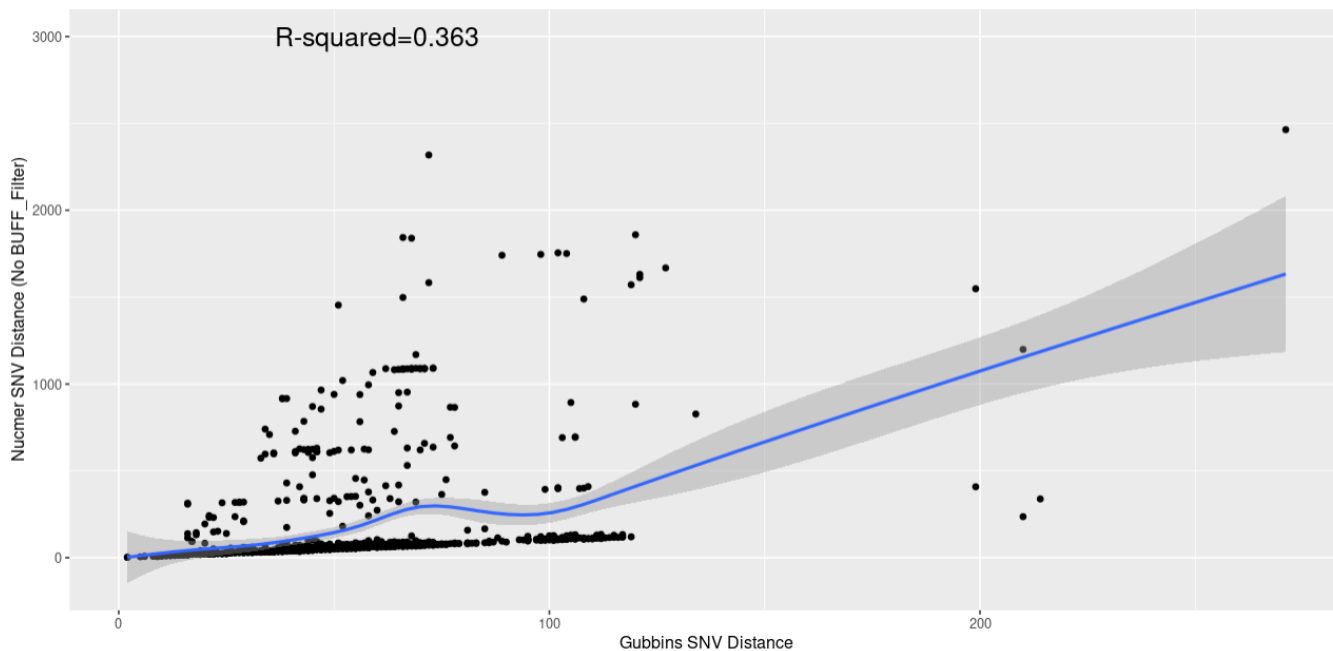
449



450

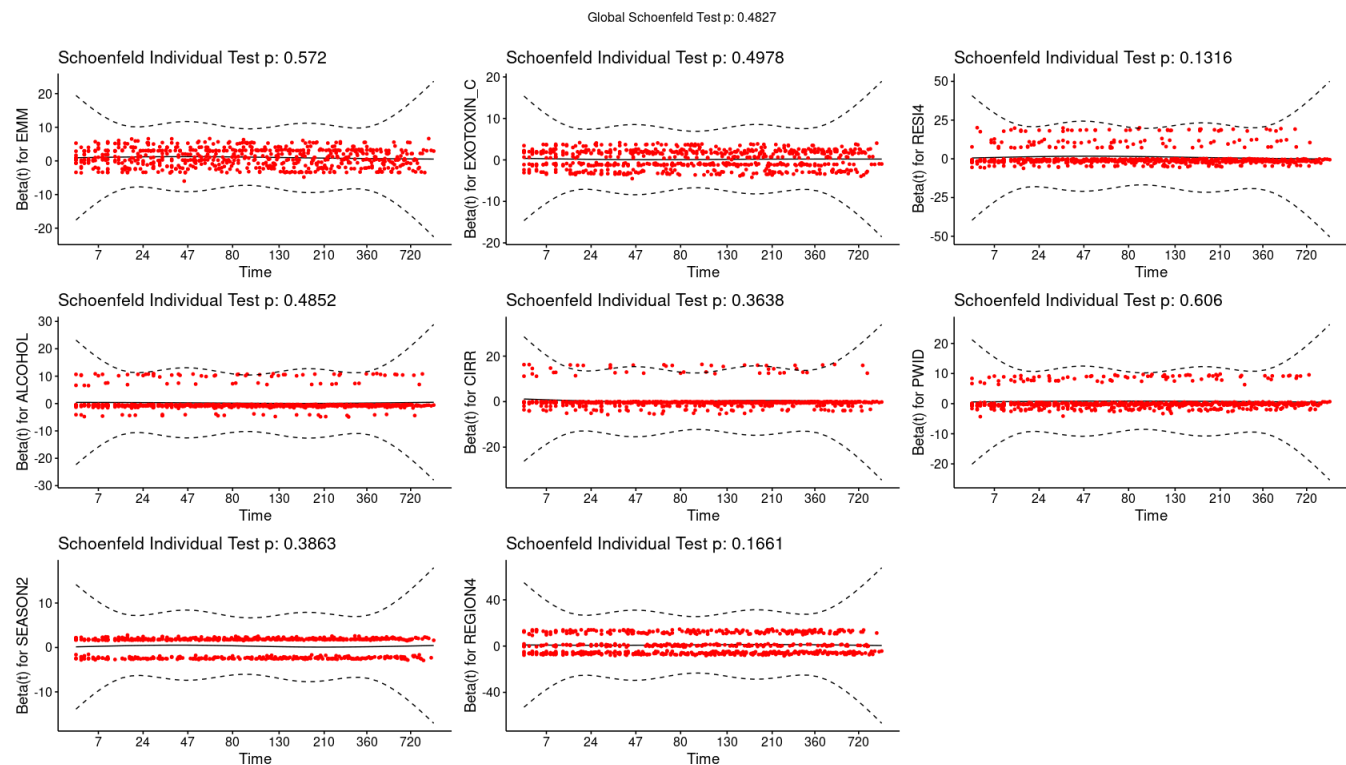
451 Figure S4. Pairwise genomic distances calculated using the Nucmer software after removing SNVs
452 within 1000bp were compared with those calculated using the Gubbins software for 1000 randomly
453 selected pairwise comparisons. An excellent correlation ($R^2=0.982$) was observed.

454 Figure S5



455

456 Figure S5. Pairwise genomic distances calculated using the Nucmer software without removing SNVs
457 within 1000bp were compared with those calculated using the Gubbins software for 1000 randomly
458 selected pairwise comparisons. A weak correlation ($R^2=0.363$) was observed.

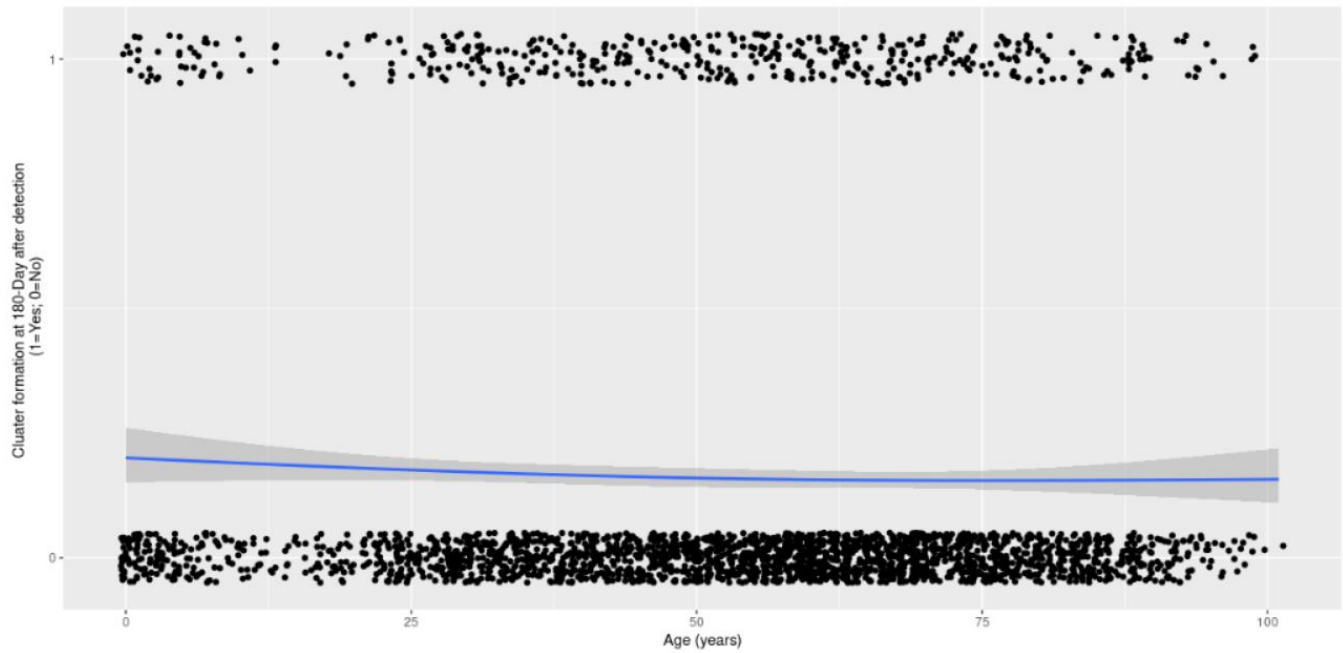


460

461

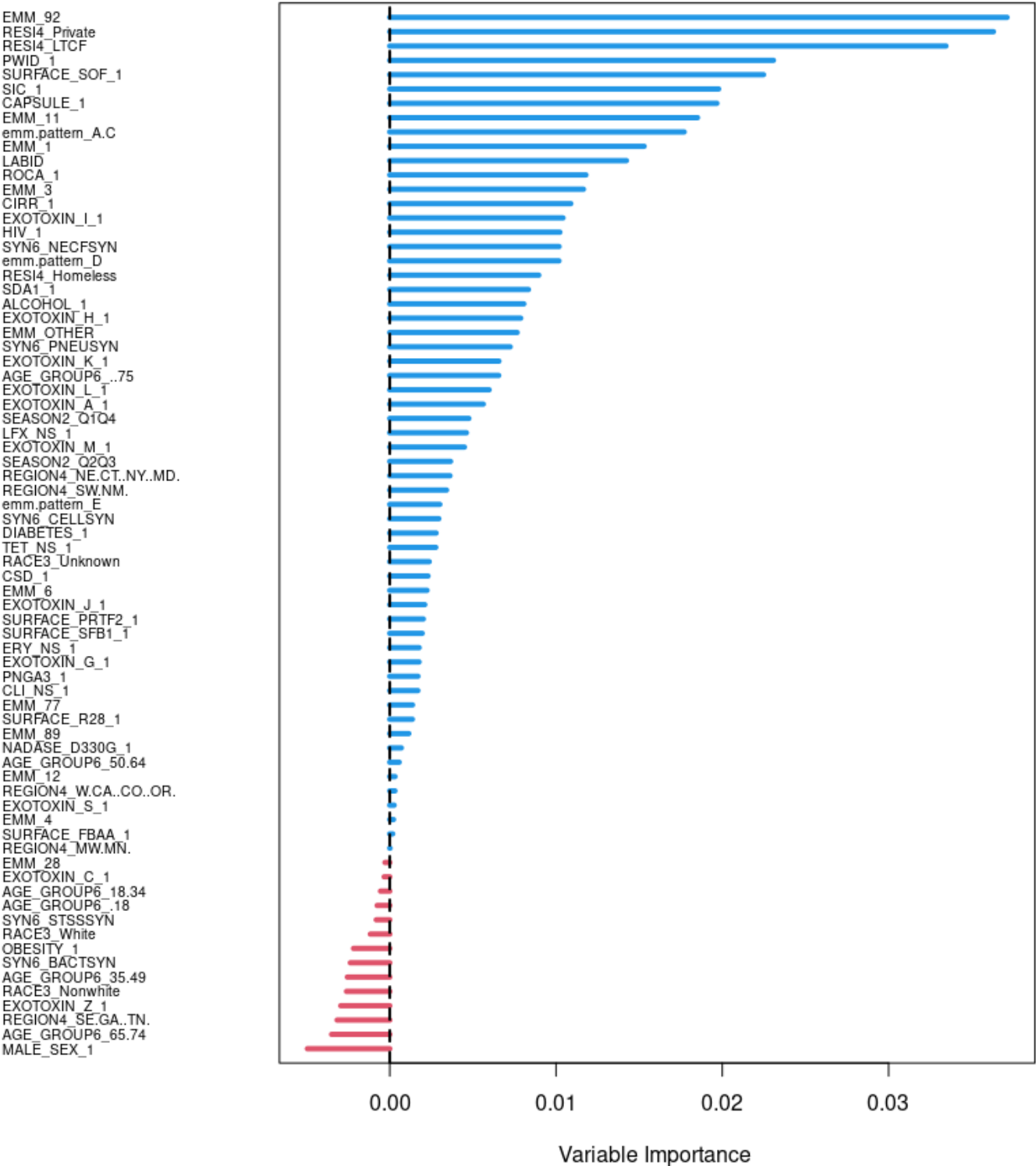
462 Figure S6. Assessment of the Schoenfeld residuals in the multivariable Cox regression model. No
463 evidence of violation of the proportional-hazards assumption was observed ($p > 0.05$ for all covariates).

464 Figure S7



465

466 Figure S7. The functional form of age as a continuous variable with binary cluster formation outcome at
467 180-day after detection.



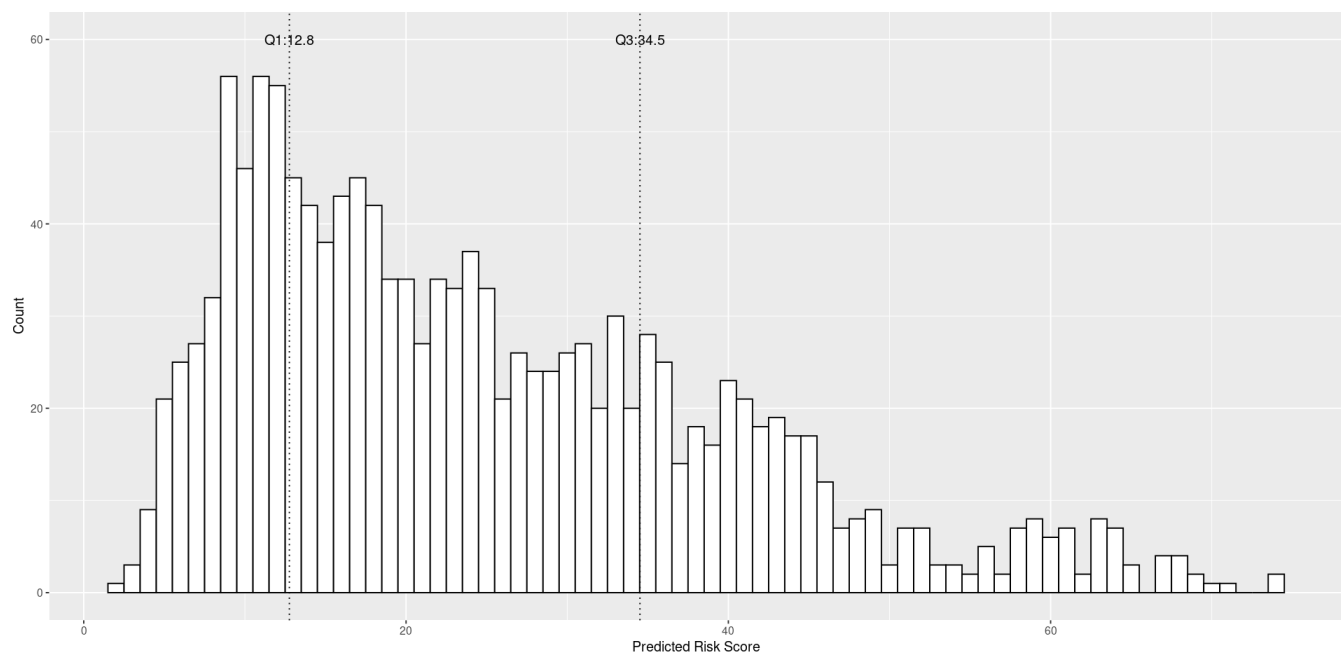
469

470 Figure S8. Variable importance plot of the random survival forest model trained by the derivation cohort.

471 The variables are ordered top-to-bottom as most-to-least important in predicting cluster formation risk.

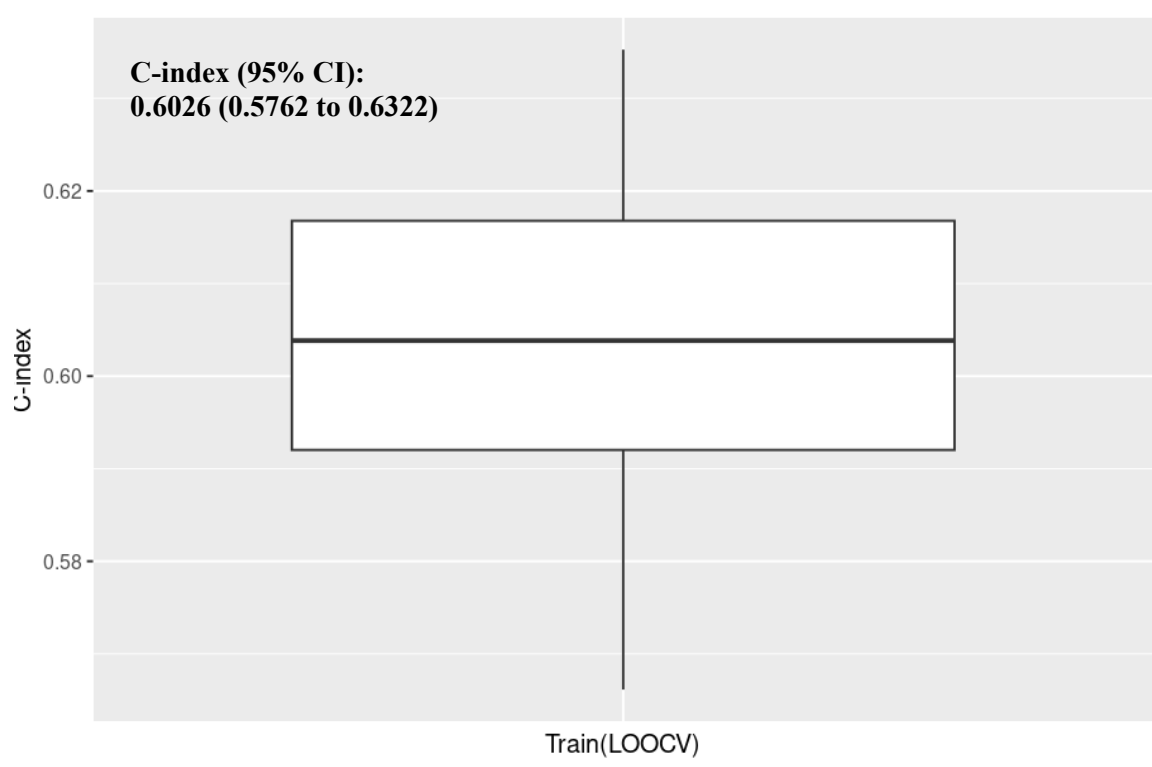
472

473 Figure S9



474

475 Figure S9. Distribution of the predicted risk score in the derivation cohort based on the trained random
476 survival forest model. Dotted lines indicate the 25th percentile (12.8) and the 75th percentile (34.5).

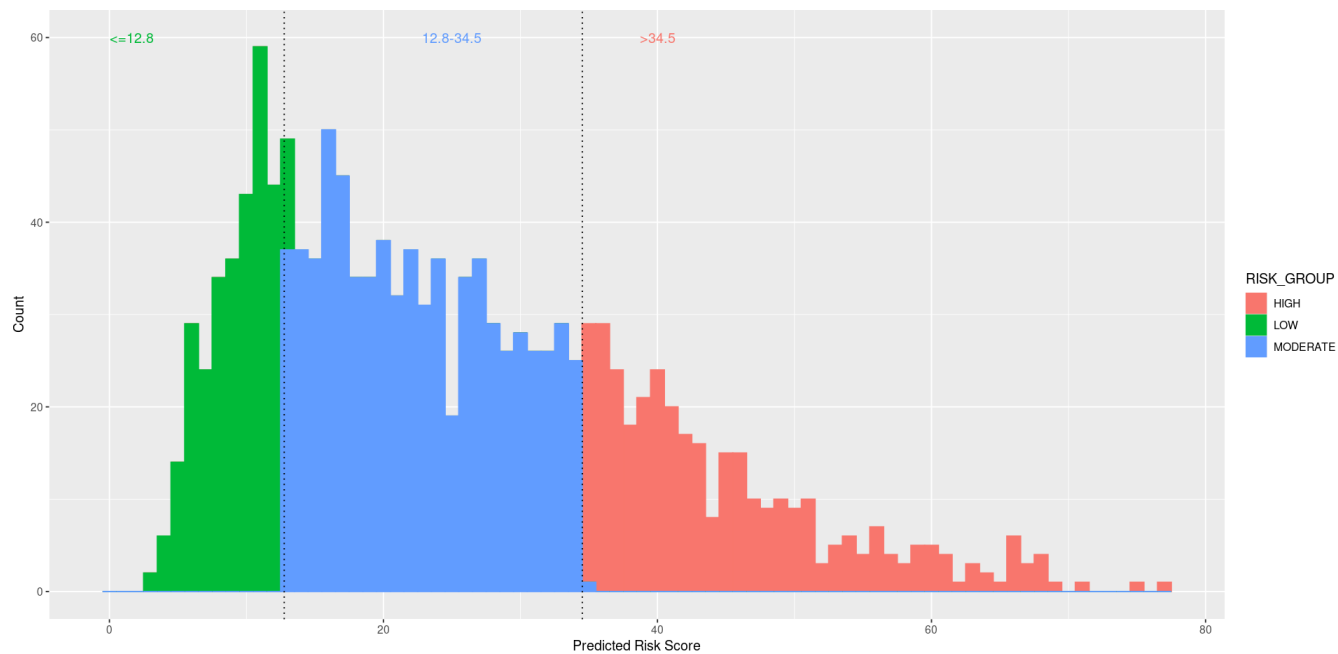


478

479 Figure S10. Predictive performance (C-index) of the random survival forest model evaluated by leave-
 480 one-out cross-validation in the derivation cohort. The variation of C-Index was estimated with the use of
 481 the nonparametric percentile bootstrap method with 1000 repetitions.

482 Figure S11

483



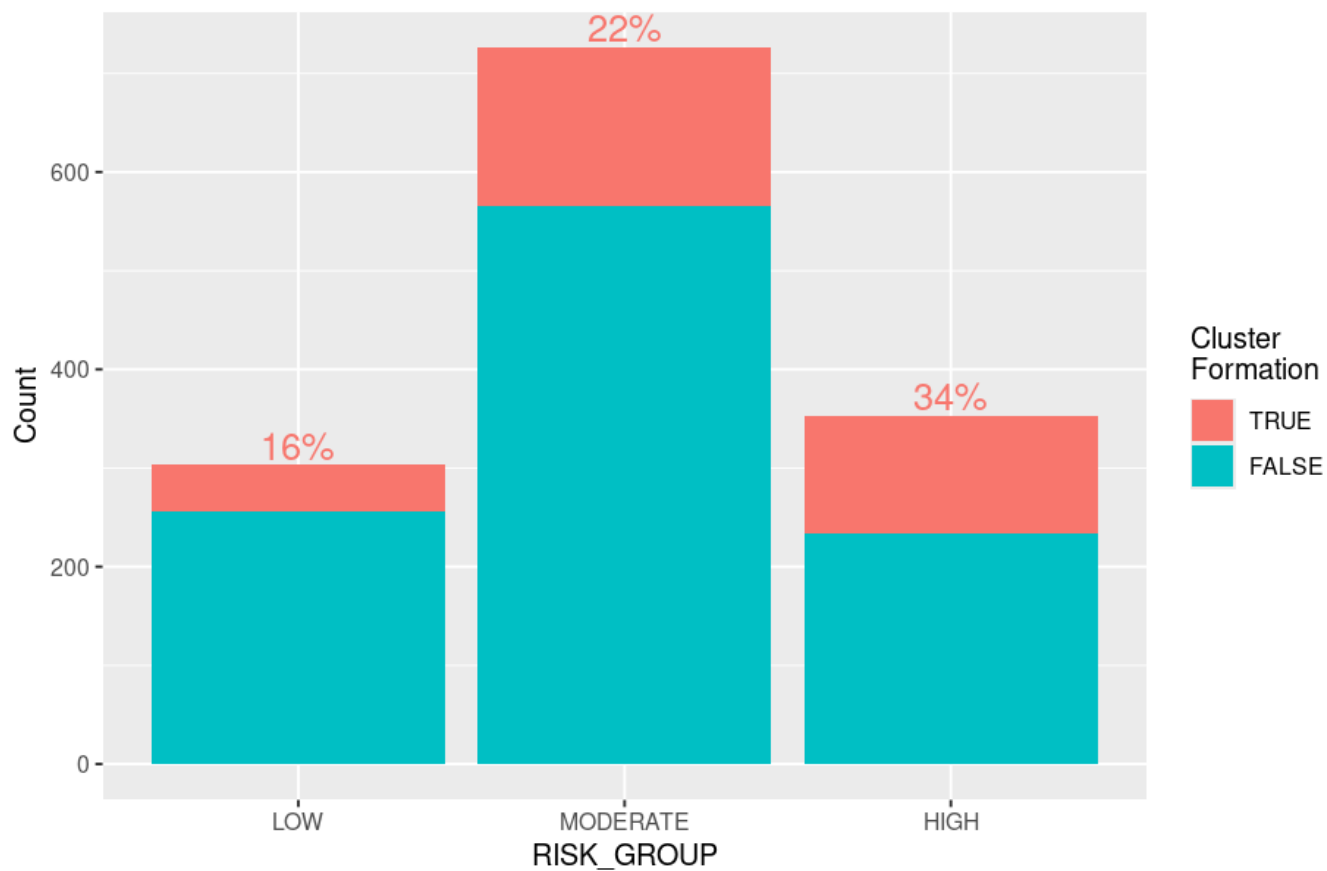
484

485

486 Figure S11. Distribution of risk scores generated by the random survival forest model evaluated by
487 leave-one-out cross-validation in the derivation cohort. Dotted lines indicate the predefined thresholds
488 (Figure S9) for clarifying at-risk cases in to 3 risk groups : Low risk (≤ 12.8), Moderate risk (12.8-34.5),
489 and High risk (> 34.5).

490 Figure S12

491

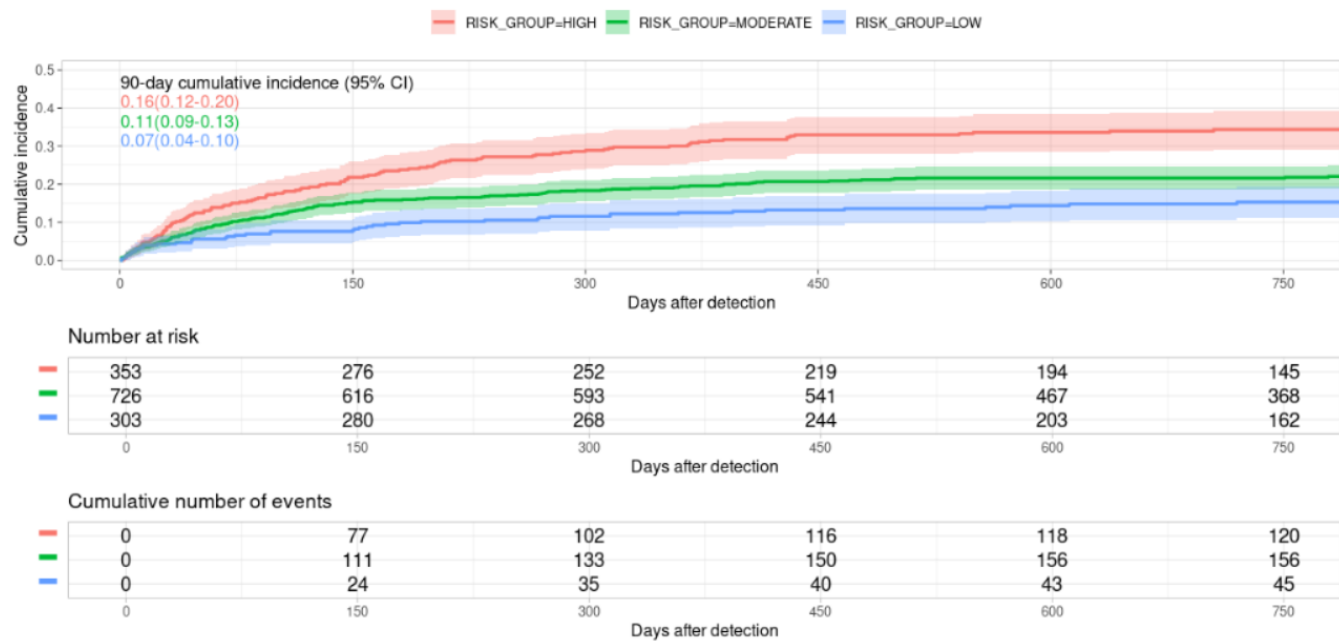


492

493 Figure S12. At-risk cases with observed cluster formation for each risk groups predicted by the random
494 survival forest model in leave-one-out cross-validation.

495 Figure S13

496

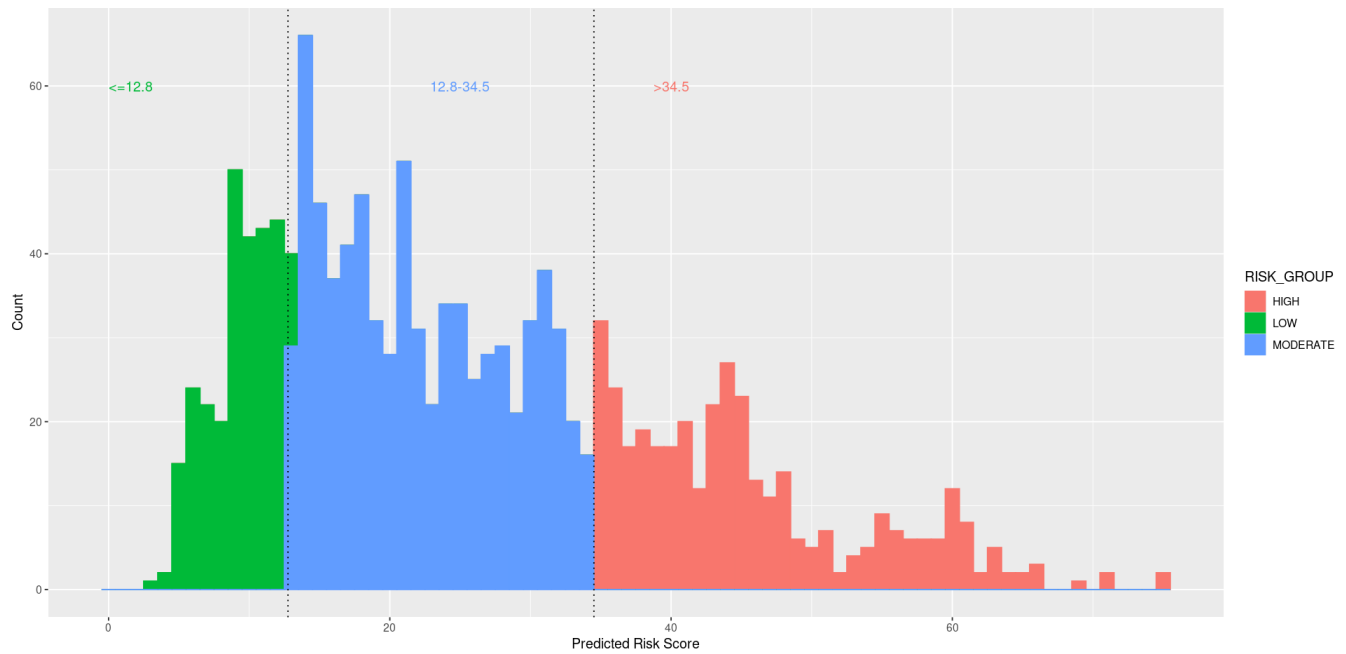


497

498 Figure S13. Cumulative incidence curves for each risk group predicted by the random survival forest
499 model in leave-one-out cross-validation.

500 Figure S14

501



502

503

504 Figure S14. Distribution of risk scores generated by the random survival forest model in the validation
505 cohort. Dotted lines indicate the predefined thresholds (Figure S9) for clarifying at-risk cases in to 3 risk
506 groups: Low risk (≤ 12.8), Moderate risk (12.8-34.5), and High risk (> 34.5).

References

1. CDC. Active Bacterial Core surveillance. 2020. <https://www.cdc.gov/abcs/index.html>; Last accessed March 22, 2024.
2. Coelho J, Kapatai G, Jironkin A, et al. Genomic sequence investigation Streptococcus pyogenes clusters in England (2010–2015). *Clinical Microbiology and Infection* 2019; **25**(1): 96-101.
3. Turner CE, Bedford L, Brown NM, et al. Community outbreaks of group A Streptococcus revealed by genome sequencing. *Sci Rep* 2017; **7**(1): 8554.
4. Alamarat ZI, Sanson MA, McNeil JC, Baker CJ, Flores AR. Use of whole-genome sequencing to detect transmission of group A Streptococcus in Houston, TX. *Access Microbiol* 2022; **4**(5): acmi000351.
5. Metcalf B, Nanduri S, Chochua S, et al. Cluster Transmission Drives Invasive Group A Streptococcus Disease Within the United States and Is Focused on Communities Experiencing Disadvantage. *J Infect Dis* 2022; **226**(3): 546-53.
6. Metcalf BJ, Chochua S, Walker H, et al. Invasive Pneumococcal Strain Distributions and Isolate Clusters Associated With Persons Experiencing Homelessness During 2018. *Clin Infect Dis* 2021; **72**(12): e948-e56.
7. Chochua S, Metcalf BJ, Li Z, et al. Population and Whole Genome Sequence Based Characterization of Invasive Group A Streptococci Recovered in the United States during 2015. *mBio* 2017; **8**(5).
8. Inouye M, Dashnow H, Raven L-A, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine* 2014; **6**: 1-16.
9. Li Y, Dominguez S, Nanduri SA, et al. Genomic Characterization of Group A Streptococci Causing Pharyngitis and Invasive Disease in Colorado, USA, June 2016 - April 2017. *J Infect Dis* 2021.
10. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic acids research* 2015; **43**(3): e15-e.
11. Greenwood M. The natural duration of cancer. *Reports on public health and medical subjects* 1926; **33**: 1-26.
12. Rebora P, Salim A, Reilly M. bshazard: A flexible tool for nonparametric smoothing of the hazard function. *R Journal* 2014; **6**(2).
13. Knofczynski GT, Mundfrom D. Sample sizes when using multiple linear regression for prediction. *Educational and psychological measurement* 2008; **68**(3): 431-42.
14. Yeh RW, Secemsky EA, Kereiakes DJ, et al. Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 year after percutaneous coronary intervention. *Jama* 2016; **315**(16): 1735-49.
15. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**(3): 515-26.
16. Chang I-M, Gelman R, Pagano M. Corrected group prognostic curves and summary statistics. *Journal of chronic diseases* 1982; **35**(8): 669-74.
17. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer* 2003; **89**(5): 781-6.
18. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; **2**(3): 841-60.
19. Team RC. R: A Language and Environment for Statistical Computing. 2020.
20. Stevens DL, Bryant AE. Severe Streptococcus pyogenes Infections. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet] 2nd edition* 2022.
21. Valenzuela TD, Hooton TM, Kaplan EL, Schlievert P. Transmission of 'toxic strep'syndrome from an infected child to a firefighter during CPR. *Annals of emergency medicine* 1991; **20**(1): 90-2.
22. Lacy MD, Horn K. Nosocomial transmission of invasive group a streptococcus from patient to health care worker. *Clinical infectious diseases* 2009; **49**(3): 354-7.
23. Gamba M-A, Martinelli M, Schaad H, et al. Familial transmission of a serious disease—producing group A streptococcus clone: case reports and review. *Clinical infectious diseases* 1997; **24**(6): 1118-21.

- 556 24. DiPersio JR, File Jr TM, Stevens DL, Gardner WG, Petropoulos G, Dinsa K. Spread of serious disease—
557 producing M3 clones of group A streptococcus among family members and health care workers. *Clinical*
558 *infectious diseases* 1996; **22**(3): 490-5.
- 559 25. Alcolea-Medina A, Snell LB, Alder C, et al. The ongoing Streptococcus pyogenes (Group A Streptococcus)
560 outbreak in London, United Kingdom, in December 2022: a molecular epidemiology study. *Clinical Microbiology*
561 *and Infection* 2023; **29**(7): 887-90.
- 562 26. Nanduri S, Metcalf B, Arwady M, et al. Prolonged and large outbreak of invasive group A Streptococcus
563 disease within a nursing home: repeated intrafacility transmission of a single strain. *Clinical Microbiology and*
564 *Infection* 2019; **25**(2): 248. e1-. e7.
- 565 27. Bessen DE, Beall BW, Davies MR. Molecular basis of serotyping and the underlying genetic organization
566 of Streptococcus pyogenes. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet] 2nd*
567 *edition* 2022.
- 568 28. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. 2008.
- 569 29. Hothorn T, Zeileis A. partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine*
570 *Learning Research* 2015; **16**(1): 3905-9.
- 571 30. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate
572 descent. *Journal of statistical software* 2010; **33**(1): 1.
- 573 31. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data
574 with competing risks. *Bioinformatics* 2009; **25**(7): 890-6.
- 575 32. Jaeger BC, Long DL, Long DM, et al. Oblique random survival forests. *The annals of applied statistics*
576 2019; **13**(3): 1847.
- 577 33. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*
578 *(Methodological)* 1972; **34**(2): 187-202.
- 579 34. Friedman JH. Stochastic gradient boosting. *Computational statistics & data analysis* 2002; **38**(4): 367-78.
- 580 35. Bühlmann P, Yu B. Boosting with the L 2 loss: regression and classification. *Journal of the American*
581 *Statistical Association* 2003; **98**(462): 324-39.
- 582 36. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in C++
583 and R. *arXiv preprint arXiv:150804409* 2015.
- 584 37. Loh WY. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge*
585 *discovery* 2011; **1**(1): 14-23.
- 586 38. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd
587 international conference on knowledge discovery and data mining; 2016; 2016. p. 785-94.

588