



Published in final edited form as:

Lancet Microbe. 2024 December ; 5(12): 100927. doi:10.1016/S2666-5247(24)00169-1.

Genomic cluster formation among invasive group A streptococcal infections in the USA: a whole-genome sequencing and population-based surveillance study

Yuan Li,
Joy Rivers,
Saundra Mathis,
Zhongya Li,
Sopio Chochua,
Benjamin J Metcalf,
Bernard Beall,
Lesley McGee

Respiratory Diseases Branch, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA (Y Li PhD, J Rivers BS, S Chochua MD PhD, B J Metcalf PhD, L McGee PhD); ASRT, Atlanta, GA, USA (S Mathis BS, Z Li BS); Eagle Global Scientific, Atlanta, GA, USA (B Beall PhD)

Summary

Background—Clusters of invasive group A streptococcal (iGAS) infection, linked to genomically closely related group A streptococcal (GAS) isolates (referred to as genomic clusters), pose public health threats, and are increasingly identified through whole-genome sequencing (WGS) analysis. In this study, we aimed to assess the risk of genomic cluster formation among iGAS cases not already part of existing genomic clusters.

Methods—In this WGS and population-based surveillance study, we analysed iGAS case isolates from the Active Bacterial Core surveillance (ABCs), which is part of the US Centers for Disease Control and Prevention's Emerging Infections Program, in ten US states from Jan 1, 2015, to Dec 31, 2019. We included all residents in ABCs sites with iGAS infections meeting the case definition and excluded non-conforming GAS infections and cases with whole-genome assemblies of the isolate containing fewer than 1·5 million total bases or more than 150 contigs. For iGAS cases we collected basic demographics, underlying conditions, and risk factors for infection from

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Correspondence to: Dr Yuan Li, Respiratory Diseases Branch, Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30033, USA, yqh8@cdc.gov.

Contributors

YL, LM, and BB contributed to study design. YL, JR, SM, ZL, SC, and BJM collected data. Data were analysed by YL, SC, and BJM, and interpreted by YL, SC, BJM, JR, SM, ZL, BB, and LM. YL designed the figures. The manuscript was written by YL and BB. BB and LM also supervised the study. All authors had full access to all the data in the study and accept responsibility for the decision to submit for publication. YL and LM verified underlying data of the study.

Declaration of interests

We declare no competing interests.

medical records, and for isolates we included *emm* types, antimicrobial resistance, and presence of virulence-related genes. Two iGAS cases were defined as genomically clustered if their isolates differed by three or less single-nucleotide variants. An iGAS case not clustered with any previous cases at the time of detection, with a minimum trace-back time of 1 year, was defined as being at risk of cluster formation. We monitored each iGAS case at risk for a minimum of 1 year to identify any cluster formation event, defined as the detection of a subsequent iGAS case clustered with the case at risk. We used the Kaplan–Meier method to estimate the cumulative incidence of cluster formation events over time. We used Cox regression to assess associations between features of cases at risk upon detection and subsequent cluster formation. We developed a random survival forest machine-learning model based on a derivation cohort (random selection of 50% of cases at risk) to predict cluster formation risk. This model was validated using a validation cohort consisting of the remaining 50% of cases at risk.

Findings—We identified 2764 iGAS cases at risk from 2016 to 2018, of which 656 (24%) formed genomic clusters by the end of 2019. Overall, the cumulative incidence of cluster formation was 0·057 (95% CI 0·048–0·066) at 30 days after detection, 0·12 (0·11–0·13) at 90 days after detection, and 0·16 (0·15–0·18) at 180 days after detection. A higher risk of cluster formation was associated with *emm* type (adjusted hazard ratio as compared with *emm*89 was 2·37 [95% CI 1·71–3·30] for *emm*1, 2·72 [1·82–4·06] for *emm*3, 2·28 [1·49–3·51] for *emm*6, 1·47 [1·05–2·06] for *emm*12, and 2·21 [1·38–3·56] for *emm*92), homelessness (1·42 [1·01–1·99]), injection drug use (2·08 [1·59–2·72]), residence in a long-term care facility (1·78 [1·29–2·45]), and the autumn–winter season (1·34 [1·14–1·57]) in multivariable Cox regression analysis. The machine-learning model stratified the validation cohort (n=1382) into groups at low (n=370), moderate (n=738), and high (n=274) risk. The 90-day risk of cluster formation was 0·03 (95% CI 0·01–0·05) for the group at low risk, 0·10 (0·08–0·13) for the group at moderate risk, and 0·21 (0·17–0·25) for the group at high risk. These results were consistent with the cross-validation outcomes in the derivation cohort.

Interpretation—Using population-based surveillance data, we found that pathogen, host, and environment factors of iGAS cases were associated with increased likelihood of subsequent genomic cluster formation. Groups at high risk were consistently identified by a predictive model which could inform prevention strategies, although future work to refine the model, incorporating other potential risk factors such as host contact patterns and immunity to GAS, is needed to improve its predictive performance.

Funding—Centers for Disease Control and Prevention.

Introduction

Streptococcus pyogenes, also referred to as Group A *Streptococcus* (GAS), is a Gram-positive human pathogen that can cause severe invasive diseases such as bacteraemia, necrotising fasciitis, and streptococcal toxic shock syndrome, as well as milder non-invasive diseases such as pharyngitis and impetigo. The case fatality rate of invasive GAS (iGAS) infection is high, around 10% overall and more than 30% among patients with necrotising fasciitis and streptococcal toxic shock syndrome in the USA.^{1,2} In recent years, incidence of iGAS disease in the USA increased substantially, from 3·2–4·3 cases per 100 000 people during 2002–12 to 7·6 cases per 100 000 people in 2019.² The incidence of iGAS

decreased to 6·1–6·3 per 100 000 people in 2020–21 but rapidly rebounded to 8·1 cases per 100 000 people in 2022.² Previous studies have documented frequent occurrence of iGAS outbreaks among long-term care facility residents, people experiencing homelessness, and people who inject drugs, potentially due to increased disease susceptibility, crowded living, and suboptimal hygiene conditions among these disadvantaged populations.^{3,4} During an outbreak, iGAS isolates from patients within the closely connected transmission network show highly similar genomic sequences (genomic clustering) because of their close relatedness to their common ancestor. Two iGAS cases can be defined as genomically clustered if genomic distance between case isolates is lower than a small single-nucleotide variant (SNV) threshold.^{5,6}

Genomic clusters of iGAS disproportionately affect disadvantaged communities⁷ and contribute heavily to the expansion of antimicrobial resistance in this pathogen.⁸ Conventional iGAS cluster detection mainly uses epidemiological links including timing, location, and contact of suspected cases.³ In the past 10 years, microbial whole-genome sequencing (WGS) data have been increasingly used to support cluster identification in both health-care and community settings.^{6,9,10} An iGAS case that is not yet part of existing clusters could be later identified as part of a new genomic cluster (cluster formation) due to the transmission of related isolates from a persistent bacterial lineage reaching a detectable level in surveillance of invasive infections. However, the risk of cluster formation and factors influencing such risk are largely unknown.

Knowledge of iGAS cluster formation risk could be valuable in understanding disease transmission mechanisms, informing chemoprophylaxis and other infection control strategies, and prioritising prevention resources. Here, we aim to assess the risk of an iGAS infection becoming a new transmission cluster (cluster formation) and identify pathogen and host features associated with increased risk.

Methods

Study design

In this WGS and population-based surveillance study, we identified iGAS cases and isolates through the Active Bacterial Core surveillance (ABCs).² ABCs is an active, laboratory-based surveillance initiative for invasive pathogens of public health importance conducted in select counties in California, Colorado, Georgia, Maryland, New York, Oregon, and Tennessee and from the entire states of Connecticut, Minnesota, and New Mexico as part of the US Centers for Disease Control and Prevention (CDC)'s Emerging Infections Program.¹¹ An iGAS case was defined as an illness with GAS isolation from a normally sterile site or from a wound culture with accompanying necrotising fasciitis or streptococcal toxic shock syndrome. Isolates were collected and medical records reviewed for the extraction of demographics (including age, sex [male and female], ethnic origin, and race), clinical syndrome and outcome, and underlying medical conditions for each case for residents in ABCs sites from Jan 1, 2015, to Dec 31, 2019. Case information also included case detection time (positive culture date), residence location (state, county, and census tract), and patient population (eg, long-term care facility residents). Potential risk factors for iGAS infection included specific medical conditions (eg, cancer; chronic heart,

kidney, or lung disease; diabetes; immunocompromising condition; and skin disease) and populations at high risk (eg, people aged 65 years or older, people who inject drugs, people experiencing homelessness, and long-term care facility residents).² Inclusion criteria were all residents in ABCs sites with iGAS infections meeting the case definition. Exclusions were non-conforming GAS infections and cases with whole-genome assemblies of the isolate containing fewer than 1.5 million total bases or more than 150 contigs, to ensure sequencing quality (appendix pp 1–2).

ABCs case reporting and isolate collection were considered to be public health surveillance activities that were exempt from institutional ethics review by the CDC. Informed consent was not required by the CDC.

Procedures

WGS-based strain characterisation for all case isolates was conducted at the CDC's *Streptococcus* Laboratory (Atlanta, GA), using a MiSeq platform (Illumina, San Diego, CA, USA) with 250 bp paired end reads for sequencing and a validated bioinformatics pipeline for strain feature analysis (including *emm* type, antimicrobial resistance, exotoxins, and other virulence-related genes) and whole-genome assembly, as previously described.¹² Pairwise SNV distances within the same *emm* type were calculated with the MUMmer package (version 3.9.0)¹³ after accounting for recombination (appendix pp 2–3, 40–41).^{14,15}

Two iGAS cases were defined as genomically clustered if their isolates differed by three or less SNVs per 1.5 Mb of aligned genomes (2×10^{-6} SNV per site). The cutoff was based on analysis of the relationship between iGAS isolate genomic relatedness and case epidemiological relatedness (time, location, and patient population) in ABCs (appendix pp 3–4, 36–39).

An iGAS case was identified as at risk of forming a cluster upon detection if it did not cluster with any previous cases (figure 1). The primary outcome was cluster formation event, defined as the detection of a subsequent iGAS case in any surveillance site that was clustered with the case at risk. Using ABCs WGS data from 2015 to 2019, cases at risk were determined through a method that required at least 1 year of trace back for adequate comparison and 1 year of trace forward for event observation. Consequently, cases from 2015 were excluded from the at-risk status analysis, and cases from 2019 were excluded from the event observation. All cases at risk had detection dates from 2016 to 2018 (appendix pp 4–5).

Statistical analysis

We determined the proportion of cases at risk with observed cluster formation events, using the Wilson score interval method to estimate 95% CIs, as a descriptive statistic. To assess the cumulative incidence of cluster formation over time, we used the Kaplan–Meier estimator from the R survival package,¹⁶ defining cumulative incidence as one minus the survival probability (ie, remaining free from cluster formation). Additionally, the hazard rate for cluster formation events was calculated using a non-parametric approach with the R bshazard package¹⁷ (appendix pp 5–7).

In the univariable analysis (appendix p 7), we used the log-rank test to compare survival distributions between two groups based on the Kaplan–Meier estimator. We analysed 72 binary case features (appendix p 16), comparing cases positive for a feature against cases negative for that feature using the log-rank test. To adjust for multiple testing and control the false discovery rate at 0.05, we applied the Benjamini-Hochberg procedure, reporting only the corrected p values (appendix p 16).

In the multivariable analysis (appendix pp 7–9), we used Cox regression to evaluate the effect of multiple factors on cluster formation risk, adjusting for potential confounders among the features. Features with a p value of less than 0.2 from univariable analysis were initially considered as predictors in the multivariable Cox model. Variables highly correlated with *emm* type (Cramer's $V > .08$)¹⁸ were excluded. The selected predictors, showing no strong intercorrelation, were included in the initial full model (appendix p 29). A stepwise selection process,¹⁹ guided by the Akaike Information Criteria, identified the most relevant variables for the final Cox model. The proportional hazards assumption was verified using the Schoenfeld residuals test. Confounder-adjusted survival curves and 95% CIs were then calculated with the Corrected Group Prognosis method via the R *adjustedCurves* package.²⁰ Furthermore, we used a multivariable mixed-effects Cox model additionally incorporating long-term care facility identification number and county as random effects to account for clustering of cases in the same long-term care facility and clustering of cases among people experiencing homelessness and people who inject drugs in the same county.

Next, we employed machine-learning models to predict cluster formation risk (appendix pp 9–12), leveraging their flexibility and that they do not rely on the proportional hazards assumption and only use information on detection date preceding the event. The cases at risk were randomly divided into two equal-sized cohorts: derivation and validation, using the *sample* function in the R program. Model comparison, tuning, and training of machine-learning models were based solely on the derivation cohort. 13 machine-learning models were evaluated in the model comparison analysis (appendix p 33). The random survival forest model²¹ exhibited highest mean concordance index (C-index) and was selected in further analysis. Hyper-parameter tuning of random survival forests was conducted via the *tune* function in the *randomForestSRC* R package, using a resampling of the derivation cohort data, before training the model on the full derivation cohort with optimised parameters (*nodesize*=20, where *nodesize* is the node size; *mtry*=20, where *mtry* is the number of randomly sampled predictors). The model's output included a vector of mortality values representing estimated risk for each individual calibrated to the scale of the number of events,²¹ which was used as the predicted risk scores. Based on these scores in the derivation cohort, we defined three risk groups (low, moderate, and high) using the 25th and 75th percentiles as thresholds.

We used the validation cohort to validate the model, with the C-index measuring predictive performance and its 95% CI estimated via percentile bootstrap (1000 repetitions). The validation cohort was stratified into three risk groups based on the predefined risk score thresholds. Cluster formation risk for each group was calculated and compared. The validation entailed sensitivity to the cutoff (appendix p 34) and time cutoff (appendix p 35). Alternative predictive performance metrics for binary outcomes (cluster formation [yes

or no)) were also evaluated. Furthermore, we conducted leave-one-out cross-validation with the derivation cohort to thoroughly validate the random survival forest model, calculating the C-index and cluster formation risk for each group (appendix p 12).

All statistical analyses were conducted with R software version 3.4.3.²² All p values were two-sided, with those below 0.05 considered as statistically significant.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Between Jan 1, 2015, and Dec 31, 2019, 11 279 iGAS cases were identified through ABCs. 9699 (86%) of 11 279 iGAS isolates were characterised by WGS and passed the study's quality filter. 6143 iGAS cases from 2016 to 2018 were assessed for at-risk status of forming a new genomic cluster (figure 1), and 2764 cases were identified as at risk due to no evidence of clustering with any previous cases. The main analysis focused on these 2764 cases at risk (905 [33%] for 2016, 943 [34%] for 2017, and 916 [33%] for 2018) and their case features on detection date (appendix p 16).

The median age of cases at risk was 56 years (IQR 37–70); and 1266 (46%) of the 2764 cases were female and 1498 (54%) were male. Many cases had comorbidities or established risk factors for iGAS infection: 802 (29%) had diabetes; 537 (19%) had a BMI of 30 kg/m² or higher; 318 (12%) had chronic skin disease; 269 (10%) were either people experiencing homelessness (141 [5%]) or people who inject drugs (204 [7%]), 76 (3%) of whom were both; and 229 (8%) had alcohol abuse. The three most common syndromes among cases at risk were cellulitis (975 [35%]), bacteraemia without focus (594 [21%]), and pneumonia (378 [14%]). 1978 (72%) of the 2764 cases at risk were infected by one of the ten most common *emm* types: type 1 (398 [14%]), type 89 (358 [13%]), type 12 (274 [10%]), type 28 (242 [9%]), type 4 (195 [7%]), type 77 (125 [5%]), type 11 (125 [5%]), type 3 (116 [4%]), type 6 (85 [3%]), or type 92 (60 [2%]). WGS-based assessment of antimicrobial resistance found that levofloxacin non-susceptible case isolates were uncommon (56 [2%]), but there were notable rates of erythromycin non-susceptibility (543 [20%]), clindamycin non-susceptibility (502 [18%]), and tetracycline non-susceptibility (585 [21%]).

By the end of 2019, cluster formation events were observed among 656 of 2764 cases at risk (23.7% [95% CI 22.2–25.4]) while the other 2108 cases at risk remained unclustered (figure 2A). The median trace-forward time for all cases at risk was 732 days (IQR 410–1087). Among the 656 cases at risk that had observed cluster formation events, the median time to event was 94 days (32–235). 587 of 656 (89.5% [95% CI 86.9–91.6]) cluster formation events occurred between a case at risk and a subsequent iGAS case in the same surveillance site. The Kaplan–Meier survival curve in figure 2B shows the cumulative incidence of cluster formation over time. The estimated cumulative incidences at 30 days were 0.057 (95% CI 0.048–0.066), at 90 days were 0.12 (0.11–0.13), and at 180 days were 0.16 (0.15–0.18). The estimated instantaneous event rate (hazard) appeared to be high immediately

after detection of cases at risk and decreased consistently afterwards (figure 2C), which was consistent with the diminishingly ascending cumulative incidence (figure 2B).

The cluster formation events across 72 subgroups, each defined by a specific case feature (eg, infection by the *emm*1 strain or homelessness), are shown in the appendix (p 16). Subgroups with the highest cluster formation proportions were in individuals infected by *emm*92 (26 of 60 [43.3%; 95% CI 31.6–55.9]), people who inject drugs (86 of 204 [42.4%; 35.6–49.0]), and people experiencing homelessness (54 of 141 [38.3%; 30.7–46.5]). Conversely, the lowest proportions were observed in individuals infected by *emm*28 (32 of 242 [13.2%; 9.5–18.1]), *emm*4 (27 of 195 [13.8%; 9.7–19.4]), and a strain positive for *Alp* family adhesion (R28; 65 of 427 [15.2%; 12.1%–18.9%]).

For 29 of the 72 subgroups, the survival (cluster formation event free) function within the subgroup (feature-positive cases) differed significantly from that outside the subgroup (feature-negative cases) based on the log-rank test (corrected $p < 0.05$; appendix p 16), indicating the 29 binary case features were significantly associated with cluster formation risk in the univariable analysis. We did not observe that cluster formation risk varied by age as a continuous variable (appendix p 43).

Multivariable Cox regression confirmed that *emm* type, patient residence, injection drug use, and season of detection were significant risk factors after adjusting for all covariates in the Cox model (table). No evidence of violation of the proportional-hazards assumption was observed (appendix p 42; $p > 0.05$ for all covariates). Compared with *emm*89, higher risk of cluster formation was associated with *emm*3 (adjusted hazard ratio 2.72 [95% CI 1.82–4.06]), *emm*1 (2.37 [1.71–3.30]), *emm*6 (2.28 [1.49–3.51]), *emm*92 (2.21 [1.38–3.56]), and *emm*12 (1.47 [1.05–2.06]). Increased risk was observed among long-term care facility residents (1.78 [1.29–2.45]) and people experiencing homelessness (1.42 [1.01–1.99]), compared with cases at risk who lived in private residence before the onset of iGAS infection, as well as in people who inject drugs 2.08 (1.59–2.72) compared with people who do not inject drugs. Cases at risk detected in the autumn–winter seasons (October to March) were significantly more likely to form clusters (1.34 [1.14–1.57]), compared with cases at risk detected in the spring–summer seasons (April to September). We show the adjusted probability of cluster formation based on the fitted Cox model for the four independent risk factors: *emm* type (figure 3A), residence (figure 3B), injection drug use (figure 3C), and autumn–winter season (figure 3D).

Additional analysis based on one SNV and five SNV genomic cluster thresholds indicated similar risk factors (appendix pp 19–27), including infection by *emm* types 1 and 12, residence in a long-term care facility, injection drug use, and the autumn–winter season; homelessness was not a significant risk factor only when the less stringent SNV threshold (five SNV) was evaluated (appendix p 27). These factors, including homelessness, were also observed in the mixed Cox model (appendix p 31).

The machine-learning approach for predicting cluster formation risk is shown in figure 4A. In model comparison analysis, the random survival forests model exhibited the highest performance (mean C-index 0.6071 [95% CI 0.5811–0.6330]), with four other models

also surpassing the Cox proportional hazards model, albeit modestly (appendix p 33). Consequently, the random survival forests model was chosen for this study and trained using data from the derivation cohort (n=1382). Notably, variables of highest importance to the random survival forests model (appendix p 44) largely overlapped with significant variables in the univariable analysis (appendix p 16).

The model's predictive performance was assessed in the validation cohort (n=1382), achieving a C-index of 0.63 (95% CI 0.60–0.66; figure 4B). The model output categorised the cohort into three risk groups (figure 4C). The three risk groups (low, moderate, and high) were defined using the 25th (12.8) and 75th (34.5) percentiles of derivation cohort risk scores as thresholds (appendix p 45). The proportion of cases with observed cluster formation was 14% (95% CI 10–18; 37 of 274) in the group at low risk, 22% (19–25; 161 of 738) in the group at moderate risk, and 35% (30–40; 130 of 370) in the group at high risk. The cumulative incidence curves representing moderate and high risk diverged sharply from the curve representing low risk in the first month (figure 4D). Continued separation of the high-risk curves from moderate-risk curves appeared to continue for another 3 months. The 90-day risk of cluster formation was 0.03 (95% CI 0.01–0.05) for the group at low risk, 0.10 (0.08–0.13) for the group at moderate risk, and 0.21 (0.17–0.25) for the group at high risk.

Leave-one-out cross-validation on the derivation cohort (appendix pp 46–49) and additional analyses based on one SNV and five SNV cluster thresholds (appendix p 34) confirmed the random survival forests model's consistent predictive performance.

Discussion

This study, using population-based and laboratory-based surveillance data in ABCs, showed that 24% of iGAS cases who were not part of an existing genomic cluster on detection date subsequently became part of a new genomic cluster, typically within a year. This proportion was much higher in certain subgroups (eg, people experiencing homelessness, people who inject drugs, and people infected by some *emm* types) and in the autumn–winter season. Using case information available on the detection date (before any cluster formation) the predictive model consistently identified iGAS cases with the highest future cluster formation risk. This insight could inform potential targeted intervention strategies. This study revealed previously unknown predictive values of known risk factors and, to our knowledge, estimated cluster formation probabilities for the first time. This study also identified specific strain features associated with transmission risk, providing a basis for generating hypotheses for future research on risk prediction and understanding mechanisms.

There are three potential scenarios for the formation of an iGAS cluster: (1) direct transmission of GAS from one invasive case to another; (2) a local spread of non-invasive GAS following the detection of the initial iGAS case; and (3) occasional invasive disease resulting from local GAS carriage. The formation of an iGAS cluster is more probable under scenarios 1 and 2, whereby the iGAS cases are linked by few transmission events. Scenario 3 is less probable because invasive disease is a rare outcome of GAS carriage. In cases of sporadic iGAS, the isolates would be separated by numerous carriage instances before reconnecting to their most recent common ancestor, resulting in a substantial SNV

distance between them. However, a higher than usual level of GAS transmission due to high prevalence of GAS carriage in a population could result in the appearance of iGAS genomic clusters by chance. Our findings and the predictive model can be interpreted as assessing the combined effects from these scenarios, influenced by their respective likelihoods.

GAS spreads mainly through individuals with milder GAS diseases (such as pharyngitis and impetigo) or asymptomatic carriers. GAS displays greater *emm* type diversity and a higher occurrence of non-throat-tropic strains in developing countries or disadvantaged populations in developed nations.^{4,23,24} Our study found a link between higher iGAS cluster formation risk and disadvantaged groups in the USA. High-risk *emm* types, such as *emm1* and *emm92*, might be more effective at causing iGAS after initial infections or more transmissible in disadvantaged groups. Regional variations in cluster formation risk (appendix p 16) could be due to complex population immunity and economic conditions such as poverty rates, housing density, and health-care access.

Our study has several limitations. First, GAS strain distribution and disease epidemiology vary globally,^{25,26} making the generalisability of identified risk factors and groups uncertain. Furthermore, clusters might form outside the surveillance area, potentially underestimating events. Second, the machine-learning model, derived from half the data and validated with the rest, represents an ideal scenario for prediction accuracy, but results might vary with datasets of different characteristics and structures, possibly affecting accuracy. Additionally, key predictive features such as direct contact with infected individuals and host immune responses to previous GAS exposures are missing from the data currently available, contributing to modest predictive accuracy (C-index around 0.63) in machine-learning models. Incorporating more relevant demographic and immunological information could enhance performance. Despite its modest accuracy, the model offers some use in risk stratification and serves as a baseline for future improvements. Third, genomic clusters defined by different SNV threshold values still share varying degrees of ancestry with each other. In the univariable and multivariable analyses, the shared ancestry between genomic clusters were not adjusted for. In this study, iGAS genomic clustering was defined by bacterial genomic distance, yet the clustering was influenced by more than just bacterial factors. New cluster predictors—such as residence in a long-term care facility, homelessness, and injection drug use—are inherently epidemiological. The iGAS genomic clustering often reflected the closer proximities of people that varied seasonally and was linked to underlying epidemiological transmission dynamics.

In conclusion, we used specific definitions of laboratory confirmed iGAS disease genomic clusters to quantify the risk of cluster formation and identify subgroups of patients with significantly higher risks. Predicting iGAS cluster formation risk might enhance prevention strategies by identifying cases at high risk before clusters occur, allowing for more effective allocation of public health resources. This approach supports targeted surveillance, improved infection control in health care, and focused educational campaigns in high-risk areas.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by the CDC as part of normal responsibilities. Major funding for this work was provided through support from CDC's Emerging Infection Program and the CDC Advanced Molecular Detection initiative. The use of product names in this manuscript does not imply their endorsement by the US Department of Health and Human Services. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

Data sharing

Whole-genome sequencing data for the study isolates have been deposited in the National Center for Biotechnology Information Sequence Read Archive under the BioProject number PRJNA395240 and are publicly accessible. Patient data in aggregate format are available at the ABCs Bact Facts Interactive Data Dashboard (<https://www.cdc.gov/abcs/bact-facts-interactive-dashboard.html>) and the data are available to the public without restrictions. Individual-level patient information is shared in accordance with the CDC–Agency for Toxic Substances and Disease Registry Policy on Releasing and Sharing Data (<https://stacks.cdc.gov/view/cdc/7563>). Requests for individual-level patient information should be addressed to Active Bacterial Core Surveillance at abcs@cdc.gov. Codes for data analysis in this study are deposited in GitHub (<https://github.com/livyuan/THELANCETMICROBE-D-23-00215R1>) and are publicly accessible.

References

1. Nelson GE, Pondo T, Toews K-A, et al. Epidemiology of invasive group A streptococcal infections in the United States, 2005–2012. *Clin Infect Dis* 2016; 63: 478–86. [PubMed: 27105747]
2. Centers for Disease Control and Prevention. Active bacterial core surveillance. 2020. <https://www.cdc.gov/abcs/index.html> (accessed March 22, 2024).
3. Jordan HT, Richards CL Jr, Burton DC, Thigpen MC, Van Beneden CA. Group A streptococcal disease in long-term care facilities: descriptive epidemiology and potential control measures. *Clin Infect Dis* 2007; 45: 742–52. [PubMed: 17712760]
4. Valenciano SJ, Onukwube J, Spiller MW, et al. Invasive group A streptococcal infections among people who inject drugs and people experiencing homelessness in the United States, 2010–2017. *Clin Infect Dis* 2021; 73: e3718–26. [PubMed: 32803254]
5. Coelho JM, Kapatai G, Jironkin A, et al. Genomic sequence investigation *Streptococcus pyogenes* clusters in England (2010–2015). *Clin Microbiol Infect* 2019; 25: 96–101. [PubMed: 29698817]
6. 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: acmi000351.
7. 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: 546–53. [PubMed: 35511035]
8. Li Y, Rivers J, Mathis S, et al. Continued increase of erythromycin- and clindamycin-nonsusceptibility among invasive group A streptococci driven by genomic clusters, USA, 2018–2019. *Clin Infect Dis* 2023; 76: e1266–69. [PubMed: 35684991]
9. Gilchrist CA, Turner SD, Riley MF, Petri WA Jr, Hewlett EL. Whole-genome sequencing in outbreak analysis. *Clin Microbiol Rev* 2015; 28: 541–63. [PubMed: 25876885]
10. Cordery R, Purba AK, Begum L, et al. Frequency of transmission, asymptomatic shedding, and airborne spread of *Streptococcus pyogenes* in schoolchildren exposed to scarlet fever: a prospective, longitudinal, multicohort, molecular epidemiological, contact-tracing study in England, UK. *Lancet Microbe* 2022; 3: e366–75. [PubMed: 35544097]

11. Langley G, Schaffner W, Farley MM, et al. Twenty years of active bacterial core surveillance. *Emerg Infect Dis* 2015; 21: 1520–28. [PubMed: 26292067]
12. 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: e01422–17. [PubMed: 28928212]
13. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biol* 2004; 5: R12. [PubMed: 14759262]
14. 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* 2022; 225: 1841–51. [PubMed: 34788828]
15. Lee IPA, Andam CP. Frequencies and characteristics of genome-wide recombination in *Streptococcus agalactiae*, *Streptococcus pyogenes*, and *Streptococcus suis*. *Sci Rep* 2022; 12: 1515. [PubMed: 35087075]
16. Therneau T A package for survival analysis in R. 2014. <https://cran.r-project.org/web/packages/survival/index.html> (accessed Oct 14, 2024).
17. Rebora P, Salim A, Reilly M. bshazard: a flexible tool for nonparametric smoothing of the hazard function. *R J* 2014; 6: 114.
18. Knofczynski GT, Mundfrom D. Sample sizes when using multiple linear regression for prediction. *Educ Psychol Meas* 2008; 68: 431–42.
19. 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: 1735–49. [PubMed: 27022822]
20. Denz R, Klaaßen-Mielke R, Timmesfeld N. A comparison of different methods to adjust survival curves for confounders. *Stat Med* 2023; 42: 1461–79. [PubMed: 36748630]
21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; 2: 841–60.
22. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020.
23. Steer AC, Law I, Matatolu L, Beall BW, Carapetis JR. Global *emm* type distribution of group A streptococci: systematic review and implications for vaccine development. *Lancet Infect Dis* 2009; 9: 611–16. [PubMed: 19778763]
24. Tyrrell GJ, Bell C, Bill L, Fathima S. Increasing incidence of invasive group A *Streptococcus* disease in First Nations population, Alberta, Canada, 2003–2017. *Emerg Infect Dis* 2021; 27: 443–51. [PubMed: 33496247]
25. Bessen DE, Smeesters PR, Beall BW. Molecular epidemiology, ecology, and evolution of group A streptococci. *Microbiol Spectr* 2018; published online Sept 6. 10.1128/microbiolspec.CPP3-0009-2018.
26. Jespersen MG, Lacey JA, Tong SYC, Davies MR. Global genomic epidemiology of *Streptococcus pyogenes*. *Infect Genet Evol* 2020; 86: 104609. [PubMed: 33147506]

Research in context

Evidence before this study

We searched PubMed for studies published on or before March 25, 2024, with no language restrictions, using the following search terms: (“group A Streptococcus” OR “Streptococcus pyogenes”) AND (“cluster” OR “outbreak” OR “transmission”) AND (“genome sequencing” OR “genomic epidemiology”). This search yielded 69 articles, of which 31 were primary research studies that used whole-genome sequencing (WGS) to investigate group A streptococcal (GAS) disease outbreaks and transmissions, primarily in health-care settings. Most studies applied WGS to define clusters of closely related isolates (genomic clusters), confirm clonality of outbreak isolates, exclude unrelated cases, and characterise outbreak strain features including antimicrobial resistance, virulence, and phylogenetic relationship. Several studies attempted to infer transmission routes, identify possible common sources of infection, and determine protracted outbreaks based on WGS data. Two studies analysed potential transmission events from index cases of GAS infection. Sriskandan and colleagues used WGS to show that recent transmission accounted for high GAS carriage among contacts of scarlet fever outbreak cases in five schools in Greater London, UK. Another study from Houston, TX, USA reported six potential transmission events, involving both invasive and non-invasive infections, that were supported by highly closely related genomes. However, none of these previous studies systematically and comprehensively assessed the risk of an invasive GAS (iGAS) case to subsequently become a genomic cluster of iGAS infections, particularly in the context of community transmissions where most iGAS cases occur. Important questions regarding clinical and microbiological risk factors for iGAS cluster formation remain largely unexplored.

Added value of this study

To our knowledge, this study is the first national-level genomic epidemiological analysis of laboratory-confirmed genomic cluster formation among iGAS infections in population-based surveillance. We used WGS on isolates collected from 9699 iGAS cases from ten US states to identify 2764 cases that were not part of a genomic cluster at the time of positive culture, thus representing isolates at risk of initiating genomic clusters. Next, we followed up each of the 2764 cases at risk in the surveillance population over time until we observed a cluster formation event, defined as a subsequent iGAS case isolate that differed by three or less single-nucleotide variants from the case isolate at risk. Our study leverages genomic surveillance to capture the dynamics of cluster formation events primarily happening in the community, expanding knowledge on iGAS transmission beyond the current literature, which largely focuses on outbreak investigations. The large study size allowed us to comprehensively analyse iGAS strain and patient features associated with higher risk of cluster formation. Further, we developed and validated a machine-learning model, which offers a novel way to predict the risk of iGAS cluster formation. This study focuses on genomic clusters, which are not necessarily the same as epidemiological clusters.

Implications of all the available evidence

The available evidence highlights that a combination of host, pathogen, and environment factors collectively influence risk for iGAS infections and cluster formation. This study shows the value of WGS in assessing the incidence and risk factors of iGAS initiating transmission clusters in a population. By integrating WGS and epidemiological data into a predictive model, we can potentially identify groups at high risk with greater precision, thereby informing more targeted public health interventions and optimising resource allocation. The framework employed in this study, which entailed measuring cluster formation risk based on surveillance data, also supports future evaluations of protection against disease transmission by vaccines, prophylactic antibiotics, and non-pharmaceutical interventions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

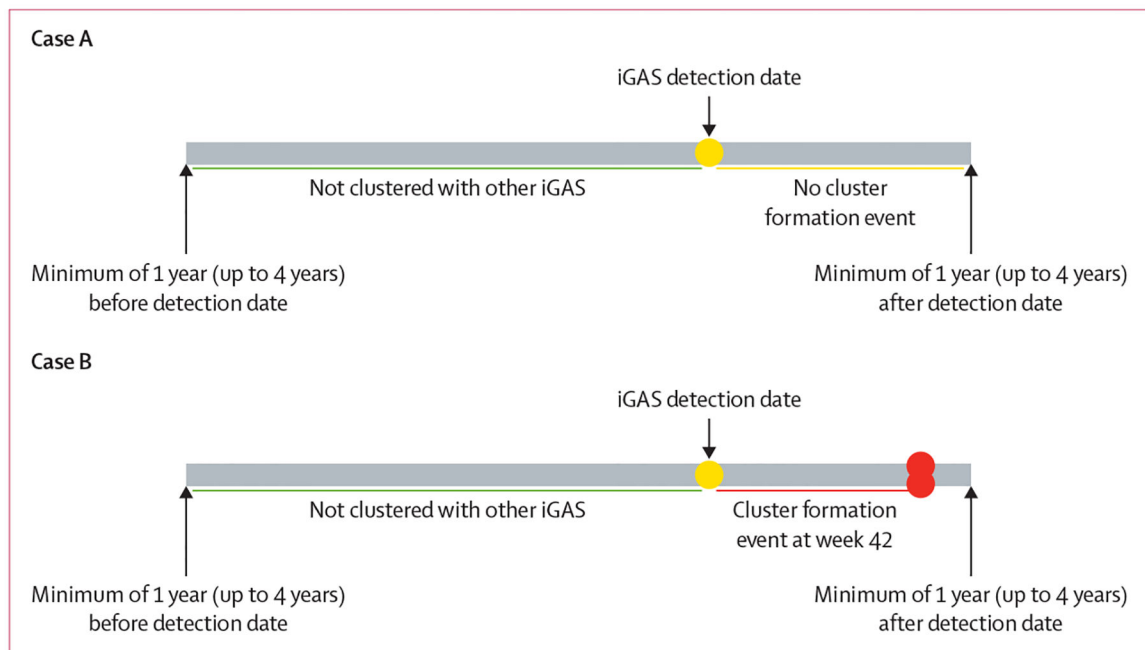


Figure 1: Schematic of the study design

Cases A and B are both at risk of genomic cluster formation because none of them are clustered with other iGAS before their detection dates. The detection date is the date of positive GAS culture. Case A remains not clustered with any other iGAS cases by the end of the study period and thus has no cluster formation event observed. Case B becomes clustered with another iGAS isolate 42 weeks after detection date (with cluster formation event). Note that this schematic view is not drawn to scale. GAS=group A *Streptococcus*. iGAS=invasive GAS.

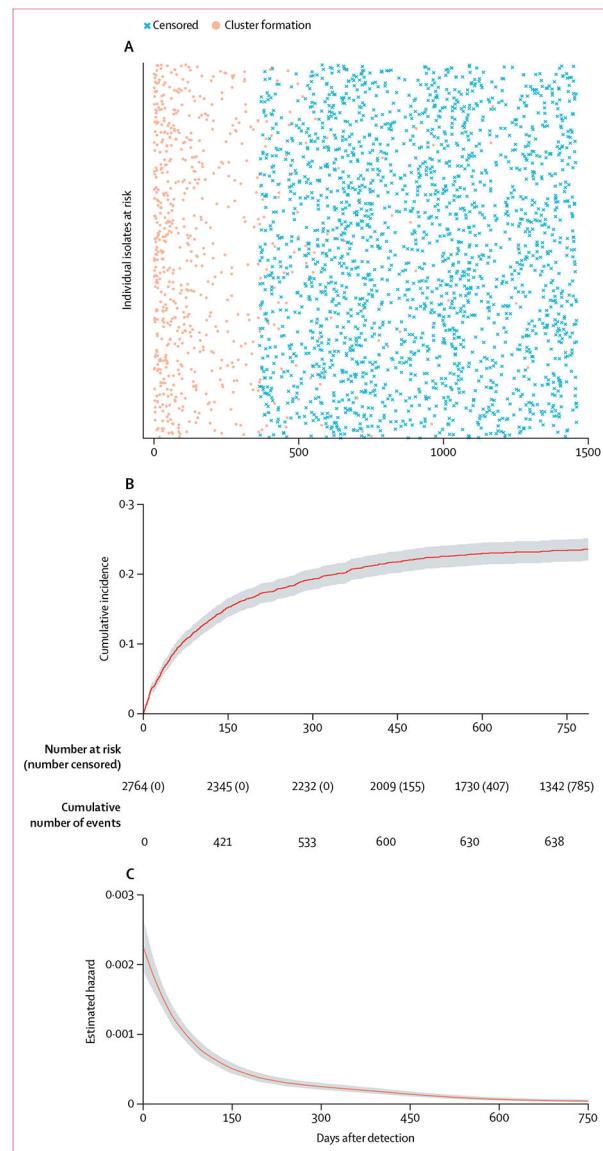
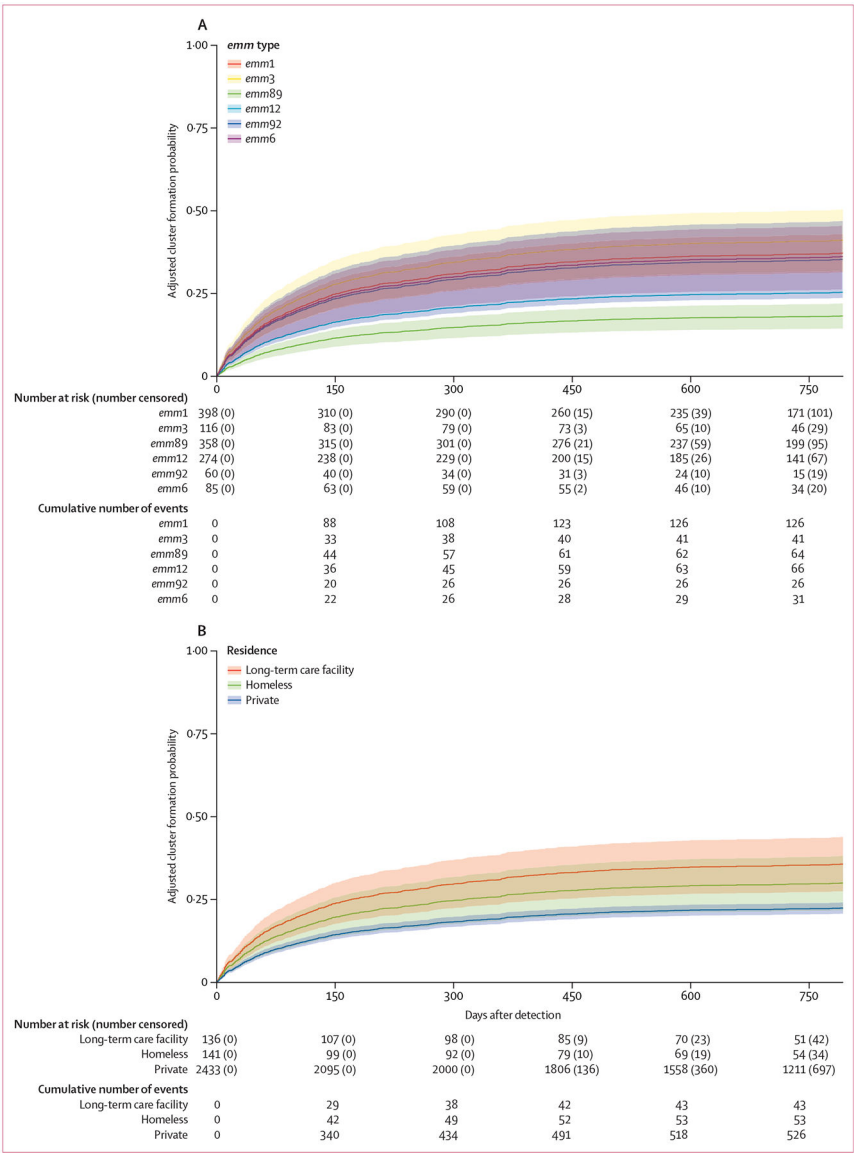


Figure 2: Cluster formation events among the 2764 iGAS isolates at risk
 (A) Red data points indicate isolates with cluster formation event (n=656) and blue data points indicate isolates censored at the end of 2019 (n=2108). (B) Cumulative incidence of iGAS cluster formation estimated by Kaplan–Meier survival analysis. (C) Estimated hazard function for iGAS cluster formation. Shaded areas in B and C indicate the 95% CIs.
 iGAS=invasive group A *Streptococcus*.



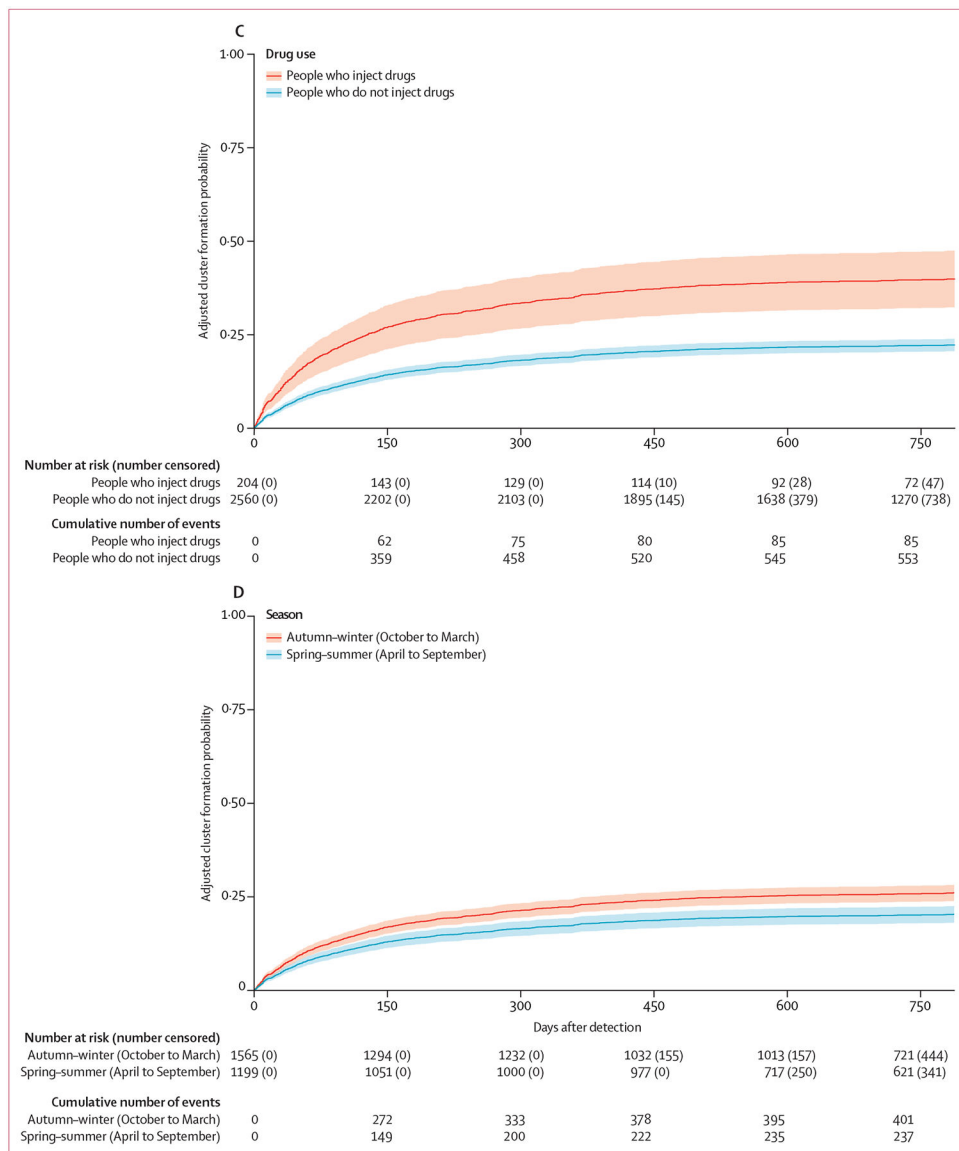


Figure 3: Adjusted probability of cluster formation based on the multivariable Cox regression
Subgroups that differed significantly from the reference group in four categorical variables are shown, including *emm* type (A), residence (B), injection drug use (C), and detection season (D). Shaded areas indicate the 95% CIs.

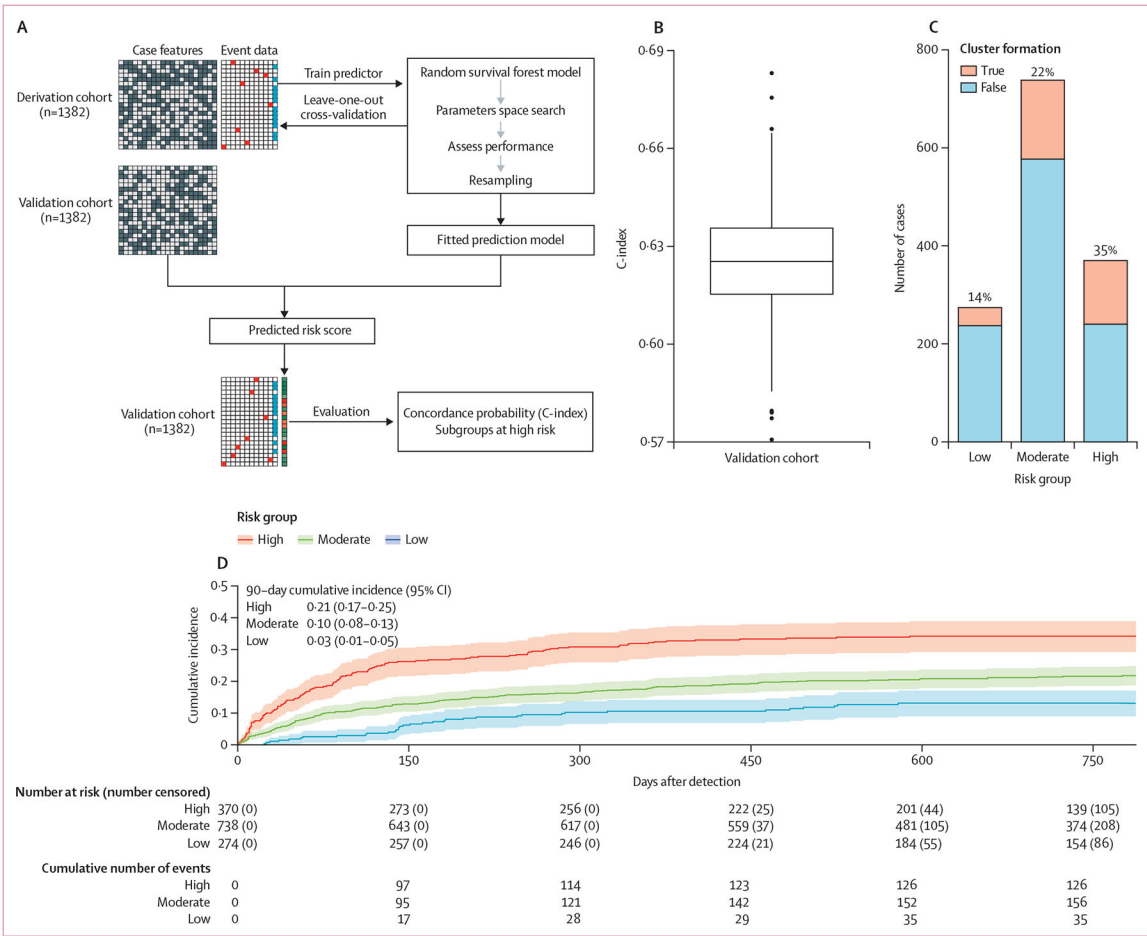


Figure 4: Predictive model for iGAS cluster formation

(A) Illustration of the machine-learning scheme derivation and validation. A random survival forest model was tuned and trained using the derivation cohort data. The predictive performance of the trained model was evaluated in a validation cohort. (B) Model predictive performance evaluated by the C-index. The middle line of the box represents the median C-index of 1000 bootstrapping; upper and lower lines of the box represent the 75th percentile (Q3) and 25th percentile (Q1), respectively; the upper and lower whiskers represent the maximum and minimum data points within 1.5 times the interquartile range from Q3 and Q1, respectively. (C) Risk groups predicted by the trained model in the validation cohort (low, moderate, and high) and the proportion of cases with observed cluster formation in each risk group. (D) Cumulative incidence curves for each predicted risk group in the validation cohort. Shaded areas indicate the 95% CIs. iGAS=invasive group A streptococcus. C-index=concordance index.

Table:

Multivariable Cox regression analysis of variables associated with cluster formation

	Adjusted hazard ratio (95% CI)	p value
<i>emm</i> type		
<i>emm1</i>	2.37 (1.71–3.30)	<0.0001
<i>emm89</i>	Ref	..
<i>emm12</i>	1.47 (1.05–2.06)	0.025
<i>emm28</i>	0.71 (0.46–1.08)	0.11
<i>emm4</i>	0.71 (0.45–1.12)	0.14
<i>emm11</i>	1.28 (0.83–1.98)	0.26
<i>emm77</i>	1.34 (0.86–2.10)	0.20
<i>emm3</i>	2.72 (1.82–4.06)	<0.0001
<i>emm6</i>	2.28 (1.49–3.51)	0.0002
<i>emm92</i>	2.21 (1.38–3.56)	0.0010
Other type than the ten most common types listed above	1.15 (0.87–1.54)	0.33
Exotoxin C		
Negative	Ref	..
Positive	1.18 (0.96–1.45)	0.11
Residence		
Private	Ref	..
Homeless	1.42 (1.01–1.99)	0.041
Long-term care facility	1.78 (1.29–2.45)	0.0004
Other than listed above	1.46 (0.88–2.41)	0.14
Alcohol abuse *	1.25 (0.96–1.62)	0.099
Cirrhosis *	1.35 (0.99–1.85)	0.062
Injection drug use *	2.08 (1.59–2.72)	<0.0001
Detection season		
Spring-summer (April to September)	Ref	..
Autumn-winter (October to March)	1.34 (1.14–1.57)	0.0003
Detection region (included states)		
Northeast (CT, NY, and MD)	Ref	..
Midwest (MN)	1.22 (0.96–1.56)	0.11
Southeast (GA and TN)	0.91 (0.72–1.15)	0.43
Southwest (NM)	1.18 (0.89–1.57)	0.24
West (CA, CO, and OR)	0.89 (0.71–1.12)	0.31

* Reference group consists of invasive group A streptococcal cases at risk that did not present with this comorbidity.