Social network analysis of *Staphylococcus aureus* carriage in a general youth population

Supplementary Material

Dina B. Stensen, MD *[1,2], Rafael A. Nozal Cañadas, MSc *[3], Lars Småbrekke, PhD [4], Karina Olsen, PhD [5], Christopher Sivert Nielsen, PhD [6,7], Kristian Svendsen, PhD [4], Anne Merethe Hanssen, PhD [8], Johanna UE Sollid, PhD [8], Gunnar Skov Simonsen, PhD [5,8], Lars Ailo Bongo, PhD [3], Anne-Sofie Furberg, PhD [5,9]

[1]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway

[2]Division of Internal Medicine, University Hospital of North Norway, Sykehusvegen 38, 9019 Tromsø, Norway

[3]Department of Computer Science, UiT The Arctic University of Norway, Hansine Hansens veg 54, 9019 Tromsø, Norway.

[4]Department of Pharmacy, Faculty of Health Sciences, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway

[5]Department of Microbiology and Infection Control, Division of Internal Medicine, University Hospital of North Norway, Sykehusvegen 38, 9019 Tromsø, Norway

[6]Department of Chronic Diseases and Ageing, Norwegian Institute of Public Health, Marcus Thranes gate 6, 0473 Oslo, Norway

[7]Department of Pain Management and Research, Division of Emergencies and Critical Care, Oslo University Hospital, Postboks 4956 Nydalen, 0424 Oslo, Norway

[8]Department of Medical Biology, Faculty of Health Sciences, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway

[9]Faculty of Health and Social Sciences, Molde University College, Britvegen 2, 6410 Molde, Norway.

*These authors have contributed equally to the development of the manuscript

Corresponding author: Dina B. Stensen.

Institution: Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway.

Postal address: Hansine Hansens veg 18, 9019 Tromsø, Norway. E-mail: dina.b.stensen@uit.no

Table of contents

Statistical background

The use of Network analysis has increased exponentially over the last few years. In this brief introduction we include the statistical background on random graph analysis and autocorrelation networks models, and some included references to provide a deeper understanding of the topic.

In statistics, a general rule to solve problems is to find all possibilities and compare all the scenarios in which something happens against all scenarios in which something does not happen. Such a ratio will give you the probability of something occurring. In our case, it is impossible to compare random graphs with all possible random graphs to get the real probability. Such calculations are unattainable, and it is necessary to constrain the amount of possible random graphs based on some assumptions (1). The constraints added to the random graph will give a model which is similar enough to reality. In our case, we use the same frequency tables with a network with the same topology as constriction. We also assume that high contagiousness would cluster positives together with positives, and negatives together with negatives.

In this context, identical topology means having the same nodes (participants) and same edges (relationships) as the original network; but each node has randomly assigned attributes based on the probability distribution of each category (i.e., *S. aureus* persistent carrier status is assigned randomly to each node, following an arbitrary 30% prevalence probability, instead of using the original value).

Our bootstrapping (2) consists of counting how many relationships connect two nodes with the same attributes in our network (i.e., persistent carrier with persistent carrier or same *spa*-

types) in 1000 simulations. This gives us a distribution of 1000 values (with a mean and standard deviation) which we can compare to the real number of homophilic relationships in our network. We then perform simple hypothesis testing like t-test, where we consider a p-value of 0.05 or less to be statistically significant. As the numbers of these tests are low, there is no need for p-value correction for false positives.

Further, we can use the random network average number of relationships that we just created to compare with our network. We can create, again, a random network with the same topology, but using the conditional probability for each host factor independently. In this way we can check how much each of the categories deviates with respect to each other, and we can identify which category has higher or lower risk for the outcome variable.

Network autocorrelation models (3) are a special case of autoregression analysis in time series (4) where we want to find how much influence your neighbors (typically your social network of friends) have over you. We aim to find the ρ coefficient in the formula:

$$Y^{(t+1)} = \rho W_n Y^{(t)} + X\beta + \varepsilon$$

W is a weighted matrix (typically normalized to 1) indicating which neighbors have influence over you, X is your explanatory variable (in our case, sex, BMI, smoke, and so on), β is a vector of coefficients (similar to linear regression) and ε is a random noise vector. Y is your dependent variable vector (in our case persistent carrier status), which over time (t), will converge to a common value. The ρ coefficient represents how much you are following the pressure of your neighbor influence and ranges typically from 0 to infinity, although negative values are also valid depending on your context. A value close to 0 would mean that you are completely ignoring your neighbors and the explanatory variables really do not have any influence on you. Positive values indicate that people can exert influence over you, and in our case, your friends will increase your risk of being a carrier.

Significant negative values would indicate that you dislike your neighbors so much, that you will do the complete opposite of what he tells you to do. In our case, this is not a valid case for the ρ coefficient as you cannot protect someone from carriage, you simply will not transmit the bacteria. However, for the explanatory variables, negative values of ρ coefficient are valid in our case because we encode categorical variables with dummy variables.

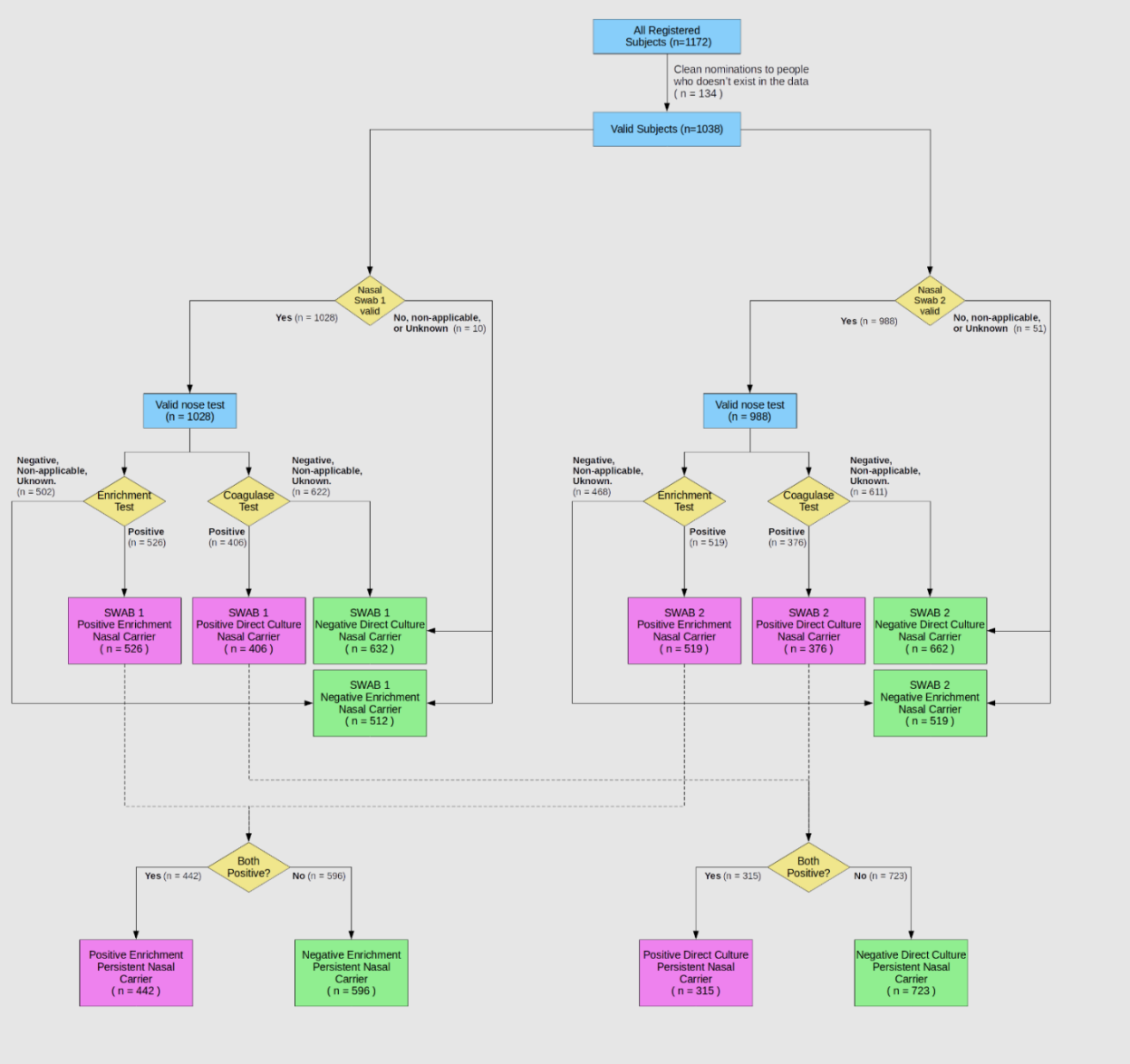
References:

1. Bevan, David. Introduction to random graphs by Alan Frieze and Michał Karoński, Theorem 11.6. The Mathematical Gazette, Volume 101, Issue 551. Cambridge University Press; 2016. 464 p. DOI: https://doi.org/10.1017/mag.2017.110

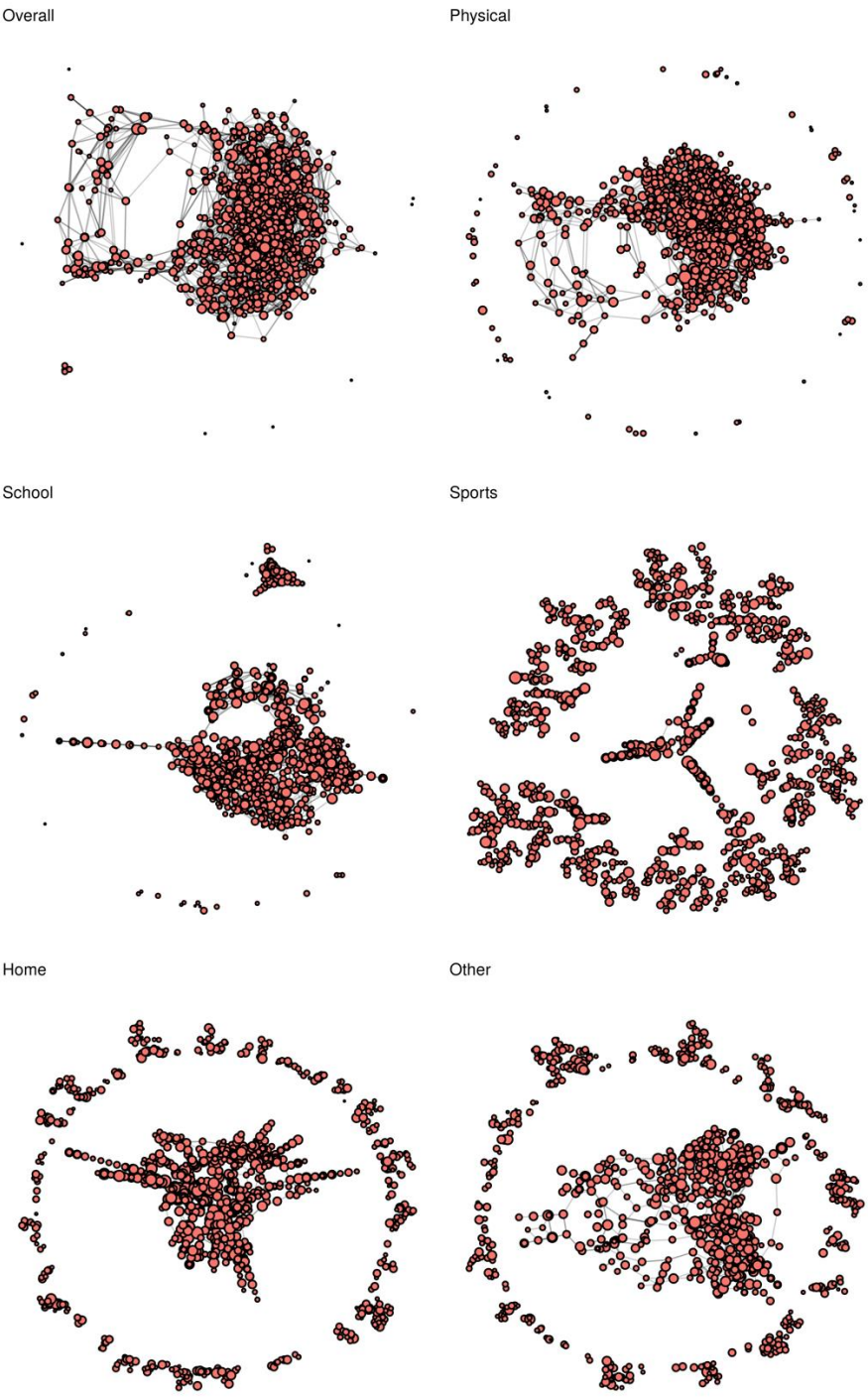
2. Snijders, Tom A B, Stephen P Borgatti. Non-Parametric Standard Errors and Tests for Network Statistics. Connections 22(2): 161-70. 1999. Available from: http://www.analytictech.com/borgatti/papers/snijders_borgatti-density_significance.pdf

3. O'Malley AJ, Marsden PV. The Analysis of Social Networks. Health Serv Outcomes Res Methodol. 2008 Dec 1;8(4):222-269. doi: 10.1007/s10742-008-0041-z. PMID: 20046802; PMCID: PMC2799303.

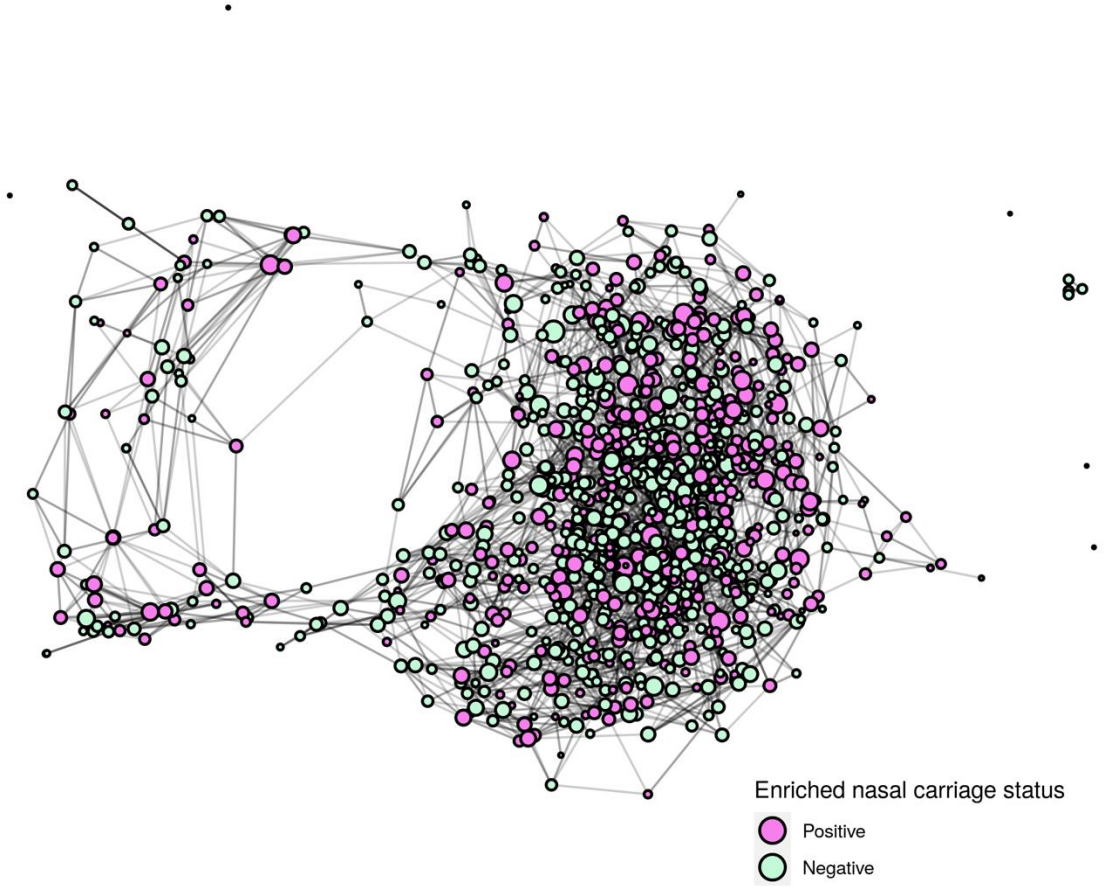
4. Anselin, L. Spatial Econometrics: Methods and Models. Netherland: Kluwer Academic; 1988.



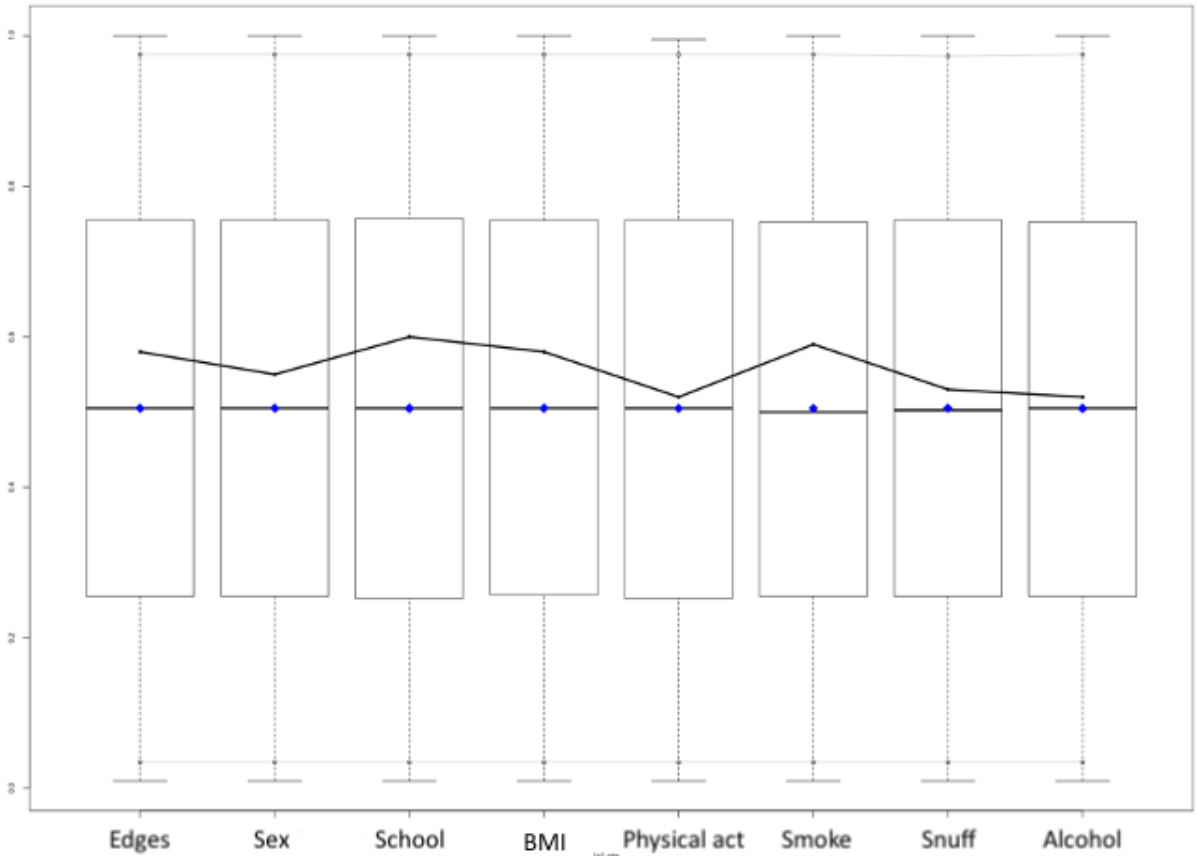
Supplementary Figure 1 Flowchart with inclusion criteria for definitions of *Staphylococcus aureus* persistent nasal carriage. The Fit Futures 1 study (N = 1038). Direct culture (bottom right) and enrichment culture (bottom left).



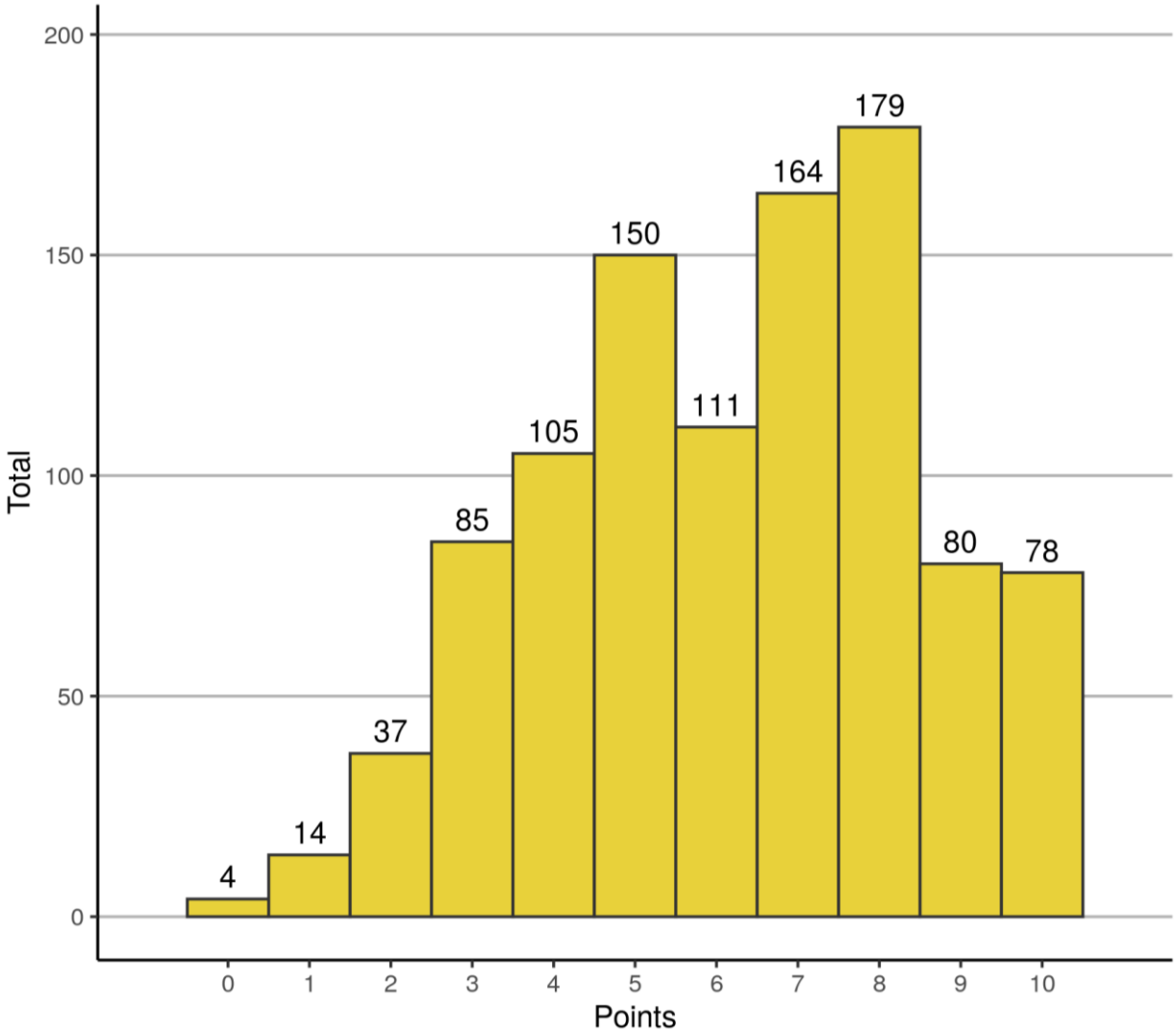
Supplementary Figure 2 Overview of the different social networks. The Fit Futures 1 study (N = 1038). From top to bottom and left to right: overall social network, physical contact, together at school, together in sports, together at home, and together in other settings. Each node represents a student. Each edge represents an undirected nomination (connection). The size of the node is proportional to the number of connections.



Supplementary Figure 3 Overall network. The Fit Futures 1 study (N = 1038). *Staphylococcus aureus* persistent nasal carriage status determined by enrichment culture is highlighted for each student (Positive = *S. aureus* detected in two nasal swab samples; Negative = *S. aureus* detected in one or none of two nasal swab samples). Node size is proportional to the number of connections (undirected friendship).



Supplementary Figure 4 Goodness of fit for the ERGM (Exponential Random Graph Model) analysis. The Fit Futures 1 study (N = 1038). Y-axis = proportion of statistics, X-axis = model statistics.



Supplementary Figure 5 Histogram of representativeness of the social network. The Fit Futures 1 study (N = 1038). 0 being not representative at all, and 10 being a perfect overview. For specific networks, having a mean score from top to bottom of school (6.61), other (6.52), physical (6.42) overall (6.29) sports (6.24) and home (6.13) (information not included in figure).

Supplementary Table 1 Characteristics of the study population by *Staphylococcus aureus* persistent nasal carriage determined by direct and enrichment culture. The Fit Futures 1 study (N = 1038).

	Direct culture			Enrichment culture		
	Positive [d]	Negative [d]	Prevalence	Positive [d]	Negative [d]	Prevalence
Sex			**< 0.001**			**< 0.001**
Male	193	337	36.4 %	255	275	48.1 %
Female	122	386	24.0 %	187	321	36.8 %
Study program			0.99			0.08
General	118	272	30.3 %	163	227	41.8 %
Sports	31	73	29.8 %	55	49	52.9 %
Vocational	166	378	30.5 %	224	320	41.2 %
Smoking			0.93			0.48
Daily	14	34	29.2 %	24	24	50.0 %
Sometimes	59	129	31.4 %	76	112	40.4 %
Never	236	546	30.2 %	333	449	42.6 %
Snuff use			0.79			0.30
Daily	73	172	29.8 %	107	138	43.7 %
Sometimes	43	88	32.8 %	63	68	48.1 %
Never	192	450	29.9 %	263	379	41.0 %
BMI category[a]			0.21			0.22
< 18.5 kg/m2	35	75	31.8 %	55	55	50.0 %
18.5-<25 kg/m2	201	509	28.3 %	289	421	40.7 %
25-<30 kg/m2	54	93	36.7 %	68	79	46.3 %
≥30 kg/m2	22	45	32.8 %	27	40	40.3 %
Physical activity[b]			0.15			0.07
None	80	149	34.9 %	107	122	46.7 %
Light	99	239	29.3 %	129	209	38.2 %

Medium	67	192	25.9 %	105	154	40.5 %
Hard	63	131	32.5 %	93	101	47.9 %
Alcohol intake			0.32			0.780
Never	88	192	31.4 %	115	165	41.1 %
<= 1 Month	134	286	31.9 %	183	237	43.6 %
≥2 Month	86	232	27.0 %	134	184	42.1 %
Hormonal contraceptives[c]			0.76			0.68
Non-user	78	249	23.9 %	121	206	37.2 %
Progestin only	3	17	15.0 %	5	15	25.0 %
Combination contraceptives, low estradiol	12	38	24.0%	19	31	38.0 %
Combination contraceptives, high estradiol	26	73	27.1 %	39	60	39.4 %

[a] BMI = body mass index

[b] Physical activity in leisure time: None = reading, watching TV, or other sedentary activity; Low level = walking, cycling, or other forms of exercise at least 4 hours a week; Medium level = participation in recreational sports, heavy outdoor activities with minimum duration of 4 hours a week; High level = Participation in heavy training or sports competitions regularly several times a week.

[c] Hormonal contraceptives: Non-user = No current use of hormonal contraceptives (women only); Progestinonly = Use of hormonal contraceptives with progestin (Cerazette, Nexplanon, Depo-provera, Implanon); Combination contraceptives, low estradiol = Use of hormonal contraceptives with progestin and ethinyl estradiol less than or equal to 20µg (Mercilon, Yasminelle, Loette 28, Nuvaring). Combination contraceptives, high estradiol = Use of hormonal contraceptives with progestin and ethinyl estradiol greater than or equal to 30µg (Marvelon, Yasmin, Microgynon, Oralcon, Diane, Synfase, Evra, Zyrona). Women taking contraceptives, but who were unable to recognize the brand were removed from the analysis.

[d] Positive = two consecutive nasal swab cultures positive for *Staphylococcus aureus* Negative = one or none of two consecutive nasal swab cultures positive for *Staphylococcus aureus*

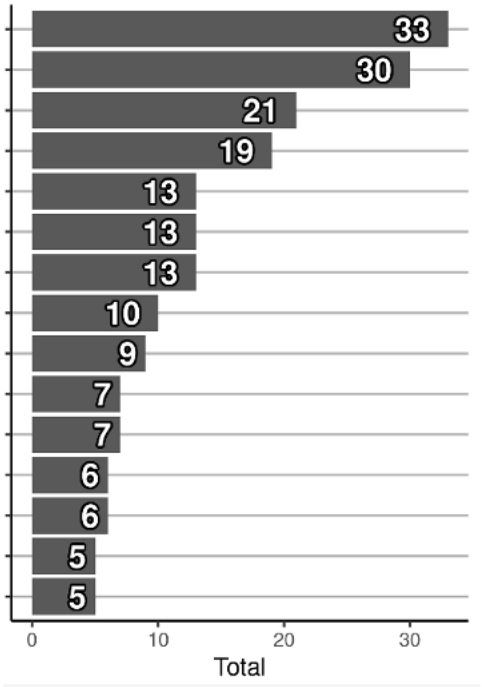
Supplementary Table 2 ERGM (Exponential Random Graph Model) analysis of relationships within groups of participants with the same characteristics. The Fit Futures 1 study (N = 1038).

	Homophily (%)	Estimate (logit)	Std Error	P-value
Edges	--	-8.41	0.08	**< 0.001**
Sex	84.05	1.47	0.05	**< 0.001**
School	87.85	2.16	0.06	**< 0.001**
BMI[a]	54.23	0.18	0.04	**< 0.001**
Smoke	68.06	0.22	0.04	**< 0.001**
Snuff	57.71	0.31	0.04	**< 0.001**
Alcohol	45.26	0.42	0.04	**< 0.001**
Physical activity	40.22	0.43	0.04	**< 0.001**

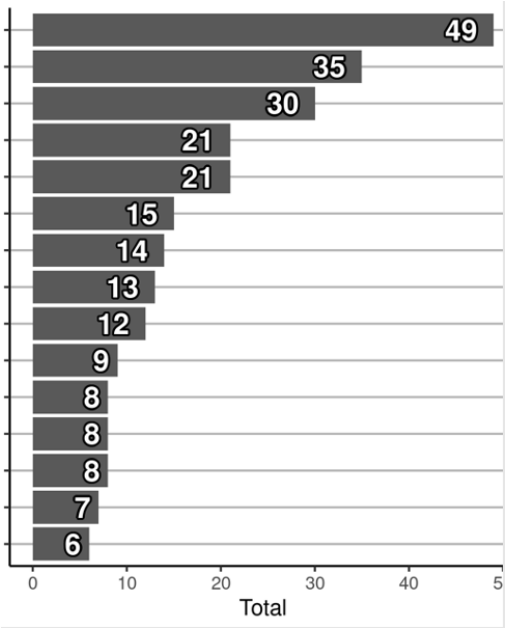
[a] BMI = body mass index

Supplementary Table 3 The most prevalent *spa*-types for *Staphylococcus aureus* throat carriage. The Fit Futures 1 study (N = 746). Only persistent carriers are shown. The plots are the results for the direct culture (above) and for enrichment culture (below).

Modality	Count	Relative	Cumulative
	33	0.09	0.09
t024	30	0.08	0.17
t084	21	0.06	0.23
t012	19	0.05	0.29
t2524	13	0.04	0.32
t065	13	0.04	0.36
t015	13	0.04	0.39
t026	10	0.03	0.42
t002	9	0.02	0.45
t346	7	0.02	0.47
t021	7	0.02	0.48
t706	6	0.02	0.5
NT	6	0.02	0.52
t4938	5	0.01	0.53
t267	5	0.01	0.55



Modality	Count	Relative	Cumulative
	49	0.11	0.11
t024	35	0.08	0.19
t084	30	0.07	0.26
t065	21	0.05	0.31
t012	21	0.05	0.35
t2524	15	0.03	0.39
t015	14	0.03	0.42
t026	13	0.03	0.45
t002	12	0.03	0.48
t346	9	0.02	0.5
t267	8	0.02	0.51
t160	8	0.02	0.53
t021	8	0.02	0.55
t246	7	0.02	0.57
t706	6	0.01	0.58



Supplementary Table 2 Detailed summary of 1000 simulations for each social network. The Fit Futures 1 study (N = 1038).

Network	Total Relationships	Equal relationships	MIN	Q1	Median	Q3	MAX	SD	Direct culture P-value
Overall	3767	2260	2012	2136	2177	2214	2353	57	0.07
Physical	2823	1698	1492	1596	1628	1658	1756	45	0.06
School	2979	1814	1559	1687	1718	1747	1866	46	**0.02**
Sports	598	365	285	333	345	357	404	17	0.12
Home	1247	731	644	703	720	737	812	25	0.34
Others	1095	663	567	616	632	648	705	22	0.08
Network	Total relationships	Equal relationships	MIN	Q1	Median	Q3	MAX	SD	Enrichment culture P-value
Overall	3767	2013	1784	1899	1926	1953	2040	40	**0.02**
Physical	2823	1502	1339	1418	1442	1465	1576	34	**0.04**
School	2979	1610	1401	1499	1524	1548	1647	36	**0.01**
Sports	598	314	257	296	306	315	367	15	0.29
Home	1247	644	570	623	638	652	717	22	0.39
Others	1095	588	507	545	558	571	625	19	0.06
Network	Total relationships	Equal relationships	MIN	Q1	Median	Q3	MAX	SD	*Spa*-type P-value
Overall	1948	136	20	45	51	58	90	9.6	**< 0.001**
Physical	1459	111	16	33	38	44	84	8.0	**< 0.001**
School	1539	100	15	35	41	46	76	8.2	**< 0.001**
Sports	335	21	0	7	9	12	22	3.7	**< 0.001**
Home	664	63	4	14	17	21	38	5.1	**< 0.001**
Others	563	45	4	12	15	18	30	4.5	**< 0.001**

Columns 4-9[4] contain the simulation summary statistics of the 1000 simulation result, in order, the minimum value of same-to-same relationships, first quartile, median rounded to the nearest integer, third quartile, maximum value, and standard deviation rounded to the nearest integer. The last column is the result of applying a t-test with the equal relationship against a distribution formed with the average of the 1000 simulations, and the standard deviation of the 1000 simulations. Significant p-values are highlighted in bold.

Supplementary Table 3 Average popularity in the overall network for each host risk factor. The Fit Futures 1 study (N = 1038).

	Average Popularity [a] (3.62)	Relative physical isolation [b] (%)	Relative frequency all (%)
Sex	**0.008**		
Male	3.46	70	51.1
Female	3.81	30	48.9
BMI-category[c]	**0.001**		
< 18.5 kg/m2	3.61	9.23	10.60
18.5-<25 kg/m2	3.72	62.31	68.40
25-<30 kg/m2	3.63	14.62	14.16
> 30 kg/m2	2.64	13.08	6.45
Smoking	**0.003**		
Daily	2.75	10	4.62
Sometimes	3.90	13.85	18.11
Never	3.60	72.31	75.34
Snuff use	**0.003**		
Daily	3.82	19.23	23.60
Sometimes	4.05	7.69	12.62
Never	3.45	69.23	61.85
Study program	**0.004**		
General	3.83	33.08	37.57
Sports	3.95	5.38	10.02
Vocational	3.43	62.31	52.41
Physical activity [d]	0.254		
None	3.45	29.23	22.06
Light	3.56	30.77	32.56
Medium	3.73	21.54	24.95
Hard	3.80	14.62	18.69

Alcohol intake	**< 0.001**		
Never	3.05	41.54	26.97
<= 1 Month	3.76	30.77	40.46
> 2 Month	3.93	23.85	30.64
Direct culture persistent carriage	0.347		
Positive	3.72	30	30.35
Negative	3.59	70	69.65
Enrichment culture persistent carriage	**0.007**		
Positive	3.87	38.46	42.58
Negative	3.47	61.54	57.42
Hormonal Contraceptives (Women only, n = 505)[e]	Average Popularity (3.81)	Relative physical isolation (%)	Relative frequency all (%)
	0.006		
Non-user	4.00	58.97	64.88
Progestin only	2.65	13.16	3.97
Low Estrogen	3.44	13.16	9.92
High Estrogen	3.63	15.79	19.64

P-values are given next to variable names and represent a significant difference from the popularity average. P-values are calculated from t-test for two categories or ANOVA for more than two categories.

[a] Average popularity = Average number of friends nominating a participant as their friend

[b] Relative physical isolation = Number of participants not being nominated at all

[c] BMI = body mass index

[d] Physical activity: None = reading, watching TV, or other sedentary activity; Low level = walking, cycling, or other forms of exercise at least 4 hours a week; Medium level = participation in recreational sports, heavy outdoor activities with minimum duration of 4 hours a week; High level = Participation in heavy training or sports competitions regularly several times a week.

[e] Hormonal contraceptives: Non-user = No current use of hormonal contraceptives (women only); Progestin-only = Use of hormonal contraceptives with progestin (Cerazette, Nexplanon, Depo-provera, Implanon); Combination contraceptives low estradiol = Use of hormonal contraceptives with progestin and ethinyl estradiol less than or equal to 20µg (Mercilon, Yasminelle, Loette 28, Nuvaring). Combination contraceptives high estradiol = Use of hormonal contraceptives with progestin and ethinyl estradiol greater than or equal to 30µg (Marvelon, Yasmin, Microgynon, Oralcon, Diane, Synfase, Evra, Zyrona). Women taking contraceptives, but who were unable to recognize the brand were removed from the analysis

Supplementary Table 4 Average number of positive friends with respect to *Staphylococcus aureus* persistent nasal carrier status. The Fit Futures 1 study (N = 1038).

	Average number of friends	P-value[a]
Direct culture		**0.002**
Persistent carrier	1.85	
Non-carrier	1.56	
Enrichment culture		**< 0.001**
Persistent carrier	2.54	
Non-carrier	2.21	
[a]Student's t-test		

Supplementary Table 5 Logistic regression model of *Staphylococcus aureus* persistent nasal carrier status with respect to positive friends. The Fit Futures 1 study (N = 1038).

	Estimate	Std Error	P-value
Direct culture			
Intercept	- 1.09	0.11	**<0.001**
Number of friends that are persistent carriers	0.16	0.05	**0.0016**
Enrichment culture			
Intercept	- 0.63	0.12	**<0.001**
Number of friends that are persistent carriers	0.14	0.04	**<0.001**

Supplementary Table 6 Attendance dates for each high school. The Fit Futures 1 study (N = 1038).

Week	Year	H1	H2	H3	H4	H5	H6	H7	H8	Friends
38	2010	32	0	0	0	0	0	0	0	64.79 %
39	2010	24	0	0	0	0	0	0	0	56.39 %
40	2010	36	0	0	0	0	0	0	0	47.82 %
41	2010	36	0	0	0	0	0	0	0	57.22 %
42	2010	35	0	0	0	0	0	0	0	49.29 %
43	2010	30	0	0	0	0	0	0	0	54.56 %
44	2010	6	16	0	0	0	0	0	0	42.80 %
45	2010	0	40	0	0	0	0	0	0	57.79 %
46	2010	0	42	0	0	0	0	0	0	62.78 %
47	2010	0	32	0	0	0	0	0	0	60.57 %
48	2010	0	6	0	0	0	0	0	28	60.69 %
49	2010	4	0	0	0	0	0	0	34	48.51 %
50	2010	4	4	0	0	0	0	0	31	39.06 %
51	2010	0	0	0	0	0	0	0	0	100.00 %
52	2010	0	0	0	0	0	0	0	0	100.00 %
1	2011	0	2	0	0	0	0	6	27	32.14 %
2	2011	0	0	0	0	0	0	43	0	61.51 %
3	2011	0	0	0	0	0	0	45	0	47.00 %
4	2011	0	0	0	0	0	0	40	0	47.87 %
5	2011	0	0	0	0	0	0	46	0	54.24 %
6	2011	0	0	30	0	0	0	10	0	45.17 %

7	2011	0	0	41	0	0	0	0	0	50.57 %
8	2011	0	0	44	0	0	0	2	0	56.52 %
9	2011	0	0	43	0	0	0	0	0	53.10 %
10	2011	0	0	0	0	0	0	0	0	100 %
11	2011	0	0	8	12	17	0	0	0	69.10 %
12	2011	0	0	0	4	18	19	0	0	54.51 %
13	2011	0	0	0	15	24	5	0	0	45.04 %
14	2011	0	0	0	22	26	0	0	0	43.44 %
15	2011	0	0	2	31	0	2	0	0	43.76 %
16	2011	0	0	0	0	0	0	0	0	100.00 %
17	2011	0	0	0	14	0	0	0	0	33.10 %

The first attendance date was 2010-September-20th, which corresponds to Week 38 of 2010. The last attendance date was 2011-April-27th, which corresponds to week 17 of 2011. Notice the public holidays in Norway, during weeks 51 and 52 of 2010 is Christmas holidays, and week 16 of 2011 is Easter holiday. H1 to H8 correspond to each of the high school identifiers. The "Friends" column shows the average proportion of friends nominated by each participant who attended the Fit Futures 1 study in the same week as the subject himself/ herself. The weighted average for all weeks is 52.07%.