# Supplemental Material

## Study 1

In addition to the analysis in the main text we perform a comparison of the summary statistics of the real and synthetic data. Table 1 shows the statistics based on the real data that we were able to use in the OLDW. Table 2 shows the statisitcs based on the synthetic data generated using HealthGAN.

**Table 1:** Reproduction of Table 2 in Vargason et. al. 2018 with Real Data. Region data obfuscated to preserve privacy of small cells.

| Characteristic | Total ASD cohort | ASD + GI subcohort | ASD–no GI subcohort | Total POP cohort | POP + GI subcohort | POP–no GI subcohort |
|---|---|---|---|---|---|---|
| Overall | 1631 | 1091 (66.89%) | 540 (33.11%) | 16,425 | 8301 (50.54%) | 8124 (49.46%) |
| Gender | | | | | | |
| Male | 1323 (81.12%) | 889 (81.48%) | 434 (80.37%) | 13,435 (81.80%) | 6799 (81.91%) | 6636 (81.68%) |
| Female | 308 (18.88%) | 202 (18.52%) | 106 (19.63%) | 2990 (18.20%) | 1502 (18.09%) | 1488 (18.32%) |
| Race/ethnicity | | | | | | |
| White | 1227 (75.23%) | 817 (74.89%) | 410 (75.93%) | 12,518 (76.21%) | 6171 (74.34%) | 6347 (78.13%) |
| Asian | 117 (7.17%) | 84 (7.70%) | 33 (6.11%) | 1176 (7.16%) | 610 (7.35%) | 566 (6.97%) |
| Black | 125 (7.66%) | 79 (7.24%) | 46 (8.52%) | 1118 (6.81%) | 575 (6.93%) | 543 (6.68%) |
| Hispanic | 162 (9.93%) | 111 (10.17%) | 51 (9.44%) | 1613 (9.82%) | 945 (11.38%) | 668 (8.22%) |
| Census division | | | | | | |
| Northeast | 296 (18.15%) | 190 (17.42%) | 106 (19.63%) | 1913 (11.65%) | 962 (11.59%) | 951 (11.71%) |
| West | 231 (14.16%) | 152 (13.93%) | 79 (14.63%) | 2541 (15.47%) | 1244 (14.99%) | 1297 (15.97%) |
| South | 686 (42.06%) | 480 (44.00%) | 206 (38.15%) | 7055 (42.95%) | 3747 (45.14%) | 3308 (40.72%) |
| Midwest | 418 (25.63%) | 269 (24.66%) | 149 (27.59%) | >4894 (>29.80%) | >2337 (>28.15%) | >2557 (>31.46%) |
| Other | 0 (0.00%) | 0 (0.00%) | 0 (0.00%) | <22 (<0.13%) | <11 (<0.13%) | <11 (<0.14%) |

**Table 2:** Reproduction of Table 2 in Vargason et. al. 2018 with Synthetic Data

| Characteristic | Total ASD cohort | ASD + GI subcohort | ASD–no GI subcohort | Total POP cohort | POP + GI subcohort | POP–no GI subcohort |
|---|---|---|---|---|---|---|
| Overall | 1341 | 886 (66.07%) | 455 (33.93%) | 16,715 | 8458 (50.60%) | 8257 (49.40%) |
| Gender | | | | | | |
| Male | 1111 (82.85%) | 729 (82.28%) | 382 (83.96%) | 14,312 (85.62%) | 7238 (85.58%) | 7074 (85.67%) |
| Female | 230 (17.15%) | 157 (17.72%) | 73 (16.04%) | 2403 (14.38%) | 1220 (14.42%) | 1183 (14.33%) |
| Race/ethnicity | | | | | | |
| White | 1067 (79.57%) | 694 (78.33%) | 373 (81.98%) | 13342 (79.82%) | 6595 (77.97%) | 6747 (81.71%) |
| Asian | 65 (4.85%) | 51 (5.76%) | 14 (3.08%) | 936 (5.60%) | 519 (6.14%) | 417 (5.05%) |
| Black | 137 (10.22%) | 91 (10.27%) | 46 (10.11%) | 1437 (8.60%) | 819 (9.68%) | 618 (7.48%) |
| Hispanic | 72 (5.37%) | 50 (5.64%) | 22 (4.84%) | 1000 (5.98%) | 525 (6.21%) | 475 (5.75%) |
| Census division | | | | | | |
| Northeast | 214 (15.96%) | 138 (15.58%) | 76 (16.70%) | 1998 (11.95%) | 1017 (12.02%) | 981 (11.88%) |
| West | 210 (15.66%) | 129 (14.56%) | 81 (17.80%) | 2336 (13.98%) | 1126 (13.31%) | 1210 (14.65%) |
| South | 559 (41.69%) | 381 (43.00%) | 178 (39.12%) | 7304 (43.70%) | 3873 (45.79%) | 3431 (41.55%) |
| Midwest | 358 (26.70%) | 238 (26.86%) | 120 (26.37%) | 5077 (30.37%) | 2442 (28.87%) | 2635 (31.91%) |

### Study 2

In addition to the analysis in the main text we perform a comparison of the summary statistics of the real and synthetic data. Table 3 shows the statistics based on the real data that we were able to use in the OLDW. Table 4 shows the statisitcs based on the synthetic data generated using HealthGAN.

**Table 3:** Reproduction of Table 2 in Vargason et. al. 2019 with Real Data

| Characteristic | POP Cohort | ASD Cohort | High Cluster | Mid Cluster | Low Cluster |
|---|---|---|---|---|---|
| Overall | 140,114 | 1571 | 457 (29.09%) | 384 (24.44%) | 730 (46.47%) |
| Gender | | | | | |
| Male | 70,977 (50.66%) | 1293 (82.30%) | 394 (86.21%) | 310 (80.73%) | 589 (80.68%) |
| Female | 69,137 (49.34%) | 278 (17.70%) | 63 (13.79%) | 74 (19.27%) | 141 (19.32%) |
| Race/ethnicity | | | | | |
| White | 105,928 (75.60%) | 1184 (75.37%) | 340 (74.40%) | 288 (75.00%) | 556 (76.16%) |
| Asian | 10,904 (7.78%) | 106 (6.75%) | 34 (7.44%) | 21 (5.47%) | 51 (6.99%) |
| Black | 8780 (6.27%) | 119 (7.57%) | 31 (6.78%) | 41 (10.68%) | 47 (6.44%) |
| Hispanic | 14,502 (10.35%) | 162 (10.31%) | 52 (11.38%) | 34 (8.85%) | 76 (10.41%) |
| Census division | | | | | |
| Northeast | 16,614 (11.86%) | 262 (16.68%) | 82 (17.94%) | 61 (15.89%) | 119 (16.30%) |
| West | 21,320 (15.22%) | 220 (14.00%) | 48 (10.50%) | 64 (16.67%) | 108 (14.79%) |
| South | 60,468 (43.16%) | 647 (41.18%) | 220 (48.14%) | 139 (36.20%) | 288 (39.45%) |
| Midwest | 41,712 (29.77%) | 442 (28.13%) | 107 (23.41%) | 120 (31.25%) | 215 (29.45%) |
| Mean number of CMC categories diagnosed | 2.29 (1.15) | 3.76 (1.38) | 4.49 (1.10) | 4.25 (1.21) | 3.04 (1.27) |

**Table 4:** Reproduction of Table 2 in Vargason et. al. 2019 with Synthetic Data

| Characteristic | POP Cohort | ASD Cohort | High Cluster | Mid Cluster | Low Cluster |
|---|---|---|---|---|---|
| Overall | 138,543 | 1571 | 407 (25.91%) | 492 (31.32%) | 672 (42.78%) |
| Gender | | | | | |
| Male | 70,389 (50.81%) | 1266 (80.59%) | 313 (76.90%) | 443 (90.04%) | 510 (75.89%) |
| Female | 68,154 (49.19%) | 305 (19.41%) | 94 (23.10%) | 49 (9.96%) | 162 (24.11%) |
| Race/ethnicity | | | | | |
| White | 107,616 (77.68%) | 1242 (79.06%) | 310 (76.17%) | 414 (84.15%) | 518 (77.08%) |
| Asian | 10,973 (7.92%) | 120 (7.64%) | 30 (7.37%) | 26 (5.28%) | 64 (9.52%) |
| Black | 7825 (5.65%) | 88 (5.60%) | 27 (6.63%) | 20 (4.07%) | 41 (6.10%) |
| Hispanic | 12,129 (8.75%) | 121 (7.70%) | 40 (9.83%) | 32 (6.50%) | 49 (7.29%) |
| Census division | | | | | |
| Northeast | 17,012 (12.28%) | 240 (15.28%) | 65 (15.97%) | 68 (13.82%) | 107 (15.92%) |
| West | 19,902 (14.37%) | 244 (15.53%) | 50 (12.29%) | 78 (15.85%) | 116 (17.26%) |
| South | 60,135 (43.41%) | 649 (41.31%) | 183 (44.96%) | 206 (41.87%) | 260 (38.69%) |
| Midwest | 41,494 (29.95%) | 438 (27.88%) | 109 (26.78%) | 140 (28.46%) | 189 (28.12%) |
| Mean number of CMC categories diagnosed | 2.08 (1.22) | 3.34 (1.38) | 4.07 (1.21) | 3.80 (1.20) | 2.57 (1.19) |

In the original paper they go one step further than just the examining the CMCs overtime, they look at the CMC categories over time. In Figure 6 we can see this plot for the real data that we pulled. In Figure 7 we can see the synthetic version of this same plot. The relationships again look very similar across the different categories. The two exceptions to this is in the Seizure and Sleep disorder categories which have much lower prevalence in the real data as well as in the synthetic, therefore we get some abnormal results.

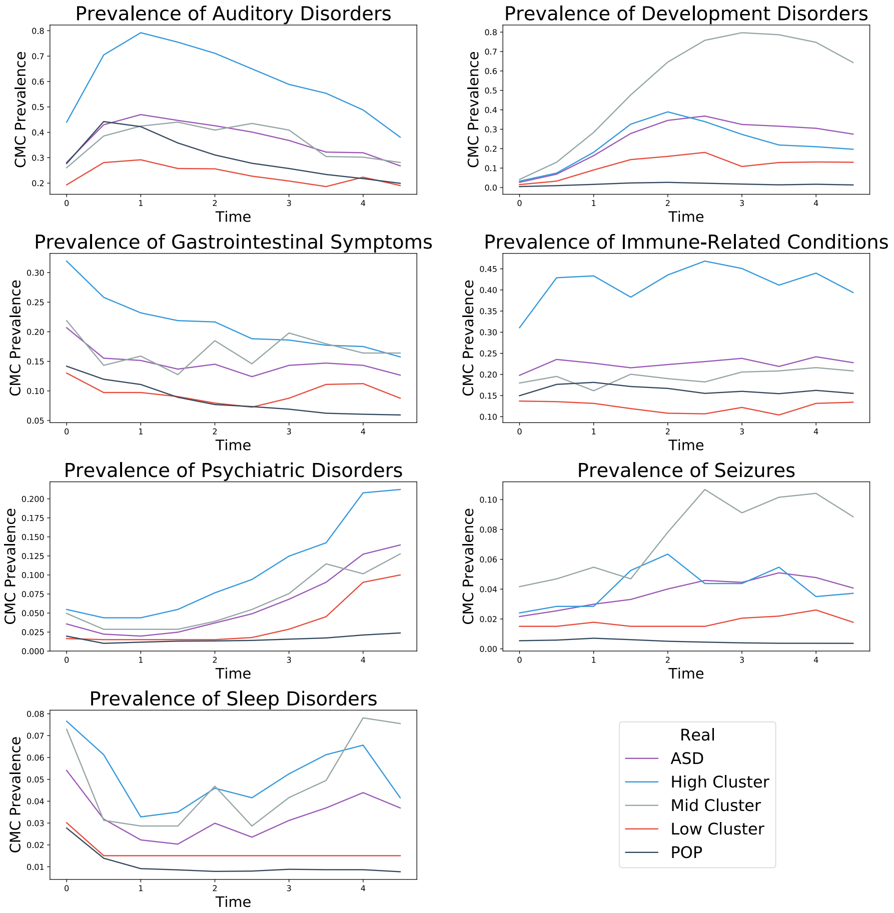

**Fig. 6:** CMC Over Time by Different Clusters by Category: Real. Some rare values are obfuscated to preserve privacy of small cells.

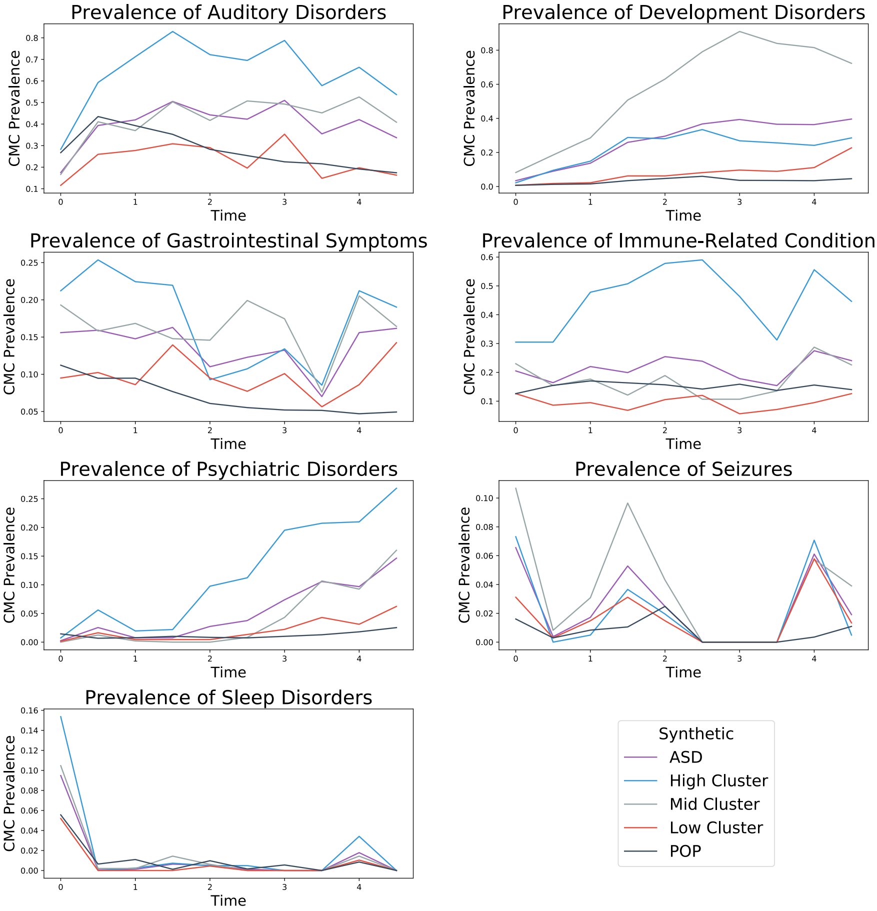

**Fig. 7:** CMC Over Time by Different Clusters by Category: Synthetic