

Table 1. HCCN 1 facility-level patient characteristics, N (%)

Characteristics	Health centers 1 – 132, range
Total	1 – 2,621
Age	*
15-20 years	0 – 200 (0% - 100%)
21-44 years	0 – 2,421 (0% - 100%)
Race	*
Asian	0 – 42 (0% - 28.2%)
Native Hawaiian or other Pacific Islander	0 – 12 (0% - 0.1%)
Black/African American	0 – 233 (0% - 100%)
American Indian/Alaska Native	0 – 74 (0% - 90.6%)
White	0 – 1,199 (0% - 100%)
More than one race	0 – 126 (0% - 100%)
Unreported/Refused to report	0 – 988 (0% - 100%)
Ethnicity	*
Hispanic/Latino	0 – 432 (0% - 75.5%)
Non-Hispanic/Latino	0 – 436 (0% - 100%)
Unreported/Refused to Report	0 – 1,753 (0% - 100%)

*Cells intentionally left empty

Table 2. HealthEfficient facility-level patient characteristics, N (%)

Characteristics	Health center 1	Health center 2	Health center 3
Total	144	153	172
Age	*	*	*
15-20 years	20 (13.9%)	30 (19.6%)	38 (22.1%)
21-44 years	124 (86.1%)	123 (80.4%)	134 (77.9%)
Race	*	*	*
Asian	5 (3.5%)	2 (1.3%)	26 (15.1%)
Native Hawaiian or other Pacific Islander	1 (0.7%)	5 (3.3%)	2 (1.2%)
Black/African American	60 (41.7%)	43 (28.1%)	56 (32.6%)
American Indian/Alaska Native	1 (0.7%)	3 (1.8%)	4 (2.3%)
White	65 (45.1%)	50 (32.7%)	33 (19.2%)
More than one race	0 (0%)	1 (0.7%)	11 (6.4%)
Unreported/Refused to report	12 (8.0%)	49 (32.0%)	40 (23.3%)
Ethnicity	*	*	*
Hispanic/Latino	60 (41.7%)	0 (0%)	0 (0%)
Non-Hispanic/Latino	80 (55.6%)	57 (37.3%)	127 (73.8%)
Unreported/Refused to Report	4 (2.8%)	96 (62.7%)	45 (26.2%)

Table 3. HCCN 1 clinician group/practice-level patient characteristics, N (%)

Characteristics	Sites 1 – 411 , range
Total	1 – 758
Age	*
15-20 years	1 – 91 (0% - 100%)
21-44 years	1 – 667 (0% - 100%)
Race	*
Asian	0 – 25 (0% - 100%)
Native Hawaiian or other Pacific Islander	0 – 12 (0% - 0.1%)
Black/African American	0 – 172 (0% - 100%)
American Indian/Alaska Native	0 – 54 (0% - 100%)
White	0 – 713 (0% - 100%)
More than one race	0 – 50 (0% - 100%)
Unreported/Refused to report	0 – 266 (0% - 100%)
Ethnicity	*
Hispanic/Latino	1 – 348 (0% - 100%)
Non-Hispanic/Latino	1 – 197 (0% - 100%)
Unreported/Refused to Report	1 – 403 (0% - 100%)

Table 4. HealthEfficient clinician group/practice-level patient characteristics, N (%)

Characteristics	Sites 1 – 14, range
Total	1 - 168
Age	*
15-20 years	0 – 37 (0% - 100%)
21-44 years	0 – 131 (0% - 100%)
Race	*
Asian	0 – 25 (0% - 14.9%)
Native Hawaiian or other Pacific Islander	0 – 5 (0% - 3.6%)
Black/African American	0 – 60 (0% - 100%)
American Indian/Alaska Native	0 – 4 (0% - 2.4%)
White	0 – 65 (0% - 100%)
More than one race	0 – 11 (0% - 100%)
Unreported/Refused to report	0 – 47 (0% - 66.7%)
Ethnicity	*
Hispanic/Latino	0 – 60 (0% - 41.7%)
Non-Hispanic/Latino	0 – 125 (0% - 100%)
Unreported/Refused to Report	0 – 91 (0% - 100%)

Table 5. Beta-binomial reliability estimates by age group, most or moderately effective methods

Level	Age group	Median N (all units)	Reliability (all units)	Median N (unit size ≥ 50)	Reliability (unit size ≥ 50)
Facility (HCCN 1)	15-44	51	0.590	157	0.862
Facility (HCCN 1)	21-44	47	0.573	139	0.848
Facility (HCCN 1)	15-20	3	0.308	84	0.834
Facility (HealthEfficient)	15-44	153	0.799	153	0.799
Facility (HealthEfficient)	21-44	124	0.833	124	0.833
Facility (HealthEfficient)	15-20	30	0.809	NA	NA
Clinician group/practice (HCCN 1)	15-44	9	0.400	102	0.813
Clinician group/practice (HCCN 1)	21-44	8	0.391	97	0.801
Clinician group/practice (HCCN 1)	15-20	1	0.201	59	0.805
Clinician group/practice (HealthEfficient)	15-44	1.5	0.231	144	0.709
Clinician group/practice (HealthEfficient)	21-44	1	0.263	124	0.771
Clinician group/practice (HealthEfficient)	15-20	0	0.524	NA	NA

Table 6. Beta-binomial reliability estimates by age group, LARC

Level	Age group	Median N (all units)	Reliability (all units)	Median N (unit size ≥ 50)	Reliability (unit size ≥ 50)
Facility (HCCN 1)	15-44	51	0.600	157	0.892
Facility (HCCN 1)	21-44	47	0.586	139	0.887
Facility (HCCN 1)	15-20	3	0.371	84	0.859
Facility (HealthEfficient)	15-44	153	0.722	153	0.722
Facility (HealthEfficient)	21-44	124	0.826	124	0.826
Facility (HealthEfficient)	15-20	30	0.501	NA	NA
Clinician group/practice (HCCN 1)	15-44	9	0.421	102	0.871
Clinician group/practice (HCCN 1)	21-44	8	0.407	97	0.856
Clinician group/practice (HCCN 1)	15-20	1	0.271	59	0.819
Clinician group/practice (HealthEfficient)	15-44	1.5	0.029	144	0.741
Clinician group/practice (HealthEfficient)	21-44	1	0.072	124	0.624
Clinician group/practice (HealthEfficient)	15-20	0	0.355	NA	NA

Table 7. Rates and reliabilities for postpartum most or moderately effective contraceptive method use and provision by facility, HealthEfficient, 2023.

Unit ID	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years
*	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)
1	11	20	0.55	0.753	NA	73	124	0.589	0.830	0.830	84	144	0.583	0.786	0.786
2	16	30	0.533	0.82	NA	39	123	0.317	0.829	0.829	55	153	0.359	0.796	0.796
3	23	38	0.605	0.853	NA	63	134	0.47	0.840	0.840	86	172	0.5	0.815	0.815
Total or Mean	50	88	0.568	*	*	175	381	0.459	*	*	225	469	0.482	*	*
*	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability
*	Median n	30	*	0.809	NA	Median n	124	*	0.833	0.833	Median n	153	*	0.799	0.799
*	Min n	20	*			Min n	123	*			Min n	144	*		

Table 8. Rates and reliabilities for postpartum LARC provision by facility, HealthEfficient, 2023.

Unit ID	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years
*	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)
1	0	20	0	0.414	NA	14	124	0.113	0.823	0.823	14	144	0.097	0.706	0.706
2	6	30	0.2	0.515	NA	17	123	0.138	0.822	0.822	23	153	0.15	0.718	0.718
3	6	38	0.158	0.573	NA	19	134	0.142	0.834	0.834	25	172	0.145	0.741	0.741
Total or Mean	12	88	0.136	*	*	50	381	0.131	*	*	62	469	0.132	*	*
*	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability
*	Median n	30	*	0.501	NA	Median n	124	*	0.826	0.826	Median n	153	*	0.722	0.722
*	Min n	20	*			Min n	123	*			Min n	144	*		

Table 9. Rates and reliabilities for postpartum most or moderately effective contraceptive method use and provision by clinician group/practice, HealthEfficient, 2023.

Unit ID	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years
*	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)
1	0	0	NA	NA	NA	0	1	0	0.045	NA	0	1	0	0.038	NA
2	1	1	1	0.167	NA	0	0	NA	NA	NA	1	1	1	0.038	NA
3	0	0	NA	NA	NA	0	1	0	0.045	NA	0	1	0	0.038	NA
4	0	0	NA	NA	NA	0	1	0	0.045	NA	0	1	0	0.038	NA
5	0	0	NA	NA	NA	0	2	0	0.086	NA	0	2	0	0.074	NA
6	0	2	0	0.286	NA	0	1	0	0.045	NA	0	3	0	0.107	NA
7	0	0	NA	NA	NA	0	1	0	0.045	NA	0	1	0	0.038	NA

Unit ID	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years
*	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	Most/Mod	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)
8	11	20	0.55	0.800	NA	73	124	0.589	0.853	0.774	84	144	0.583	0.852	0.701
9	0	0	NA	NA	NA	2	3	0.667	0.123	NA	2	3	0.667	0.107	NA
10	0	1	0	0.167	NA	0	0	NA	NA	NA	0	1	0	0.038	NA
11	23	37	0.622	0.881	NA	61	131	0.466	0.860	0.783	84	168	0.5	0.870	0.733
12	0	0	NA	NA	NA	0	3	0	0.123	NA	0	3	0	0.107	NA
13	0	0	NA	NA	NA	0	1	0	0.045	NA	0	1	0	0.038	NA
14	15	27	0.556	0.844	NA	39	112	0.348	0.840	0.755	54	139	0.388	0.847	0.694
Total or Mean	50	88	0.568	*	*	175	381	0.459	*	*	225	469	0.482	*	*
*	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability
*	Median n	0	*	0.524	NA	Median n	1	*	0.263	0.771	Median n	1.5	*	0.231	0.709
*	Min n	0	*			Min n	0	*			Min n	1	*		

Table 10. Rates and reliabilities for postpartum LARC provision by clinician group/practice, HealthEfficient, 2023.

Unit ID	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	15 – 20 Years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	21 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years	15 - 44 years
*	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)	LARC	Total N	Rate	Reliability (all units)	Reliability (unit size≥50)
1	0	0	NA	NA	NA	0	1	0	0.003	NA	0	1	0	0.001	NA
2	0	1	0	0.060	NA	0	0	NA	NA	NA	0	1	0	0.001	NA
3	0	0	NA	NA	NA	0	1	0	0.003	NA	0	1	0	0.001	NA
4	0	0	NA	NA	NA	0	1	0	0.003	NA	0	1	0	0.001	NA
5	0	0	NA	NA	NA	0	2	0	0.006	NA	0	2	0	0.002	NA
6	0	2	0	0.113	NA	0	1	0	0.003	NA	0	3	0	0.003	NA
7	0	0	NA	NA	NA	0	1	0	0.003	NA	0	1	0	0.001	NA
8	0	20	0	0.561	NA	14	124	0.113	0.276	0.628	14	144	0.097	0.123	0.734
9	0	0	NA	NA	NA	1	3	0.333	0.009	NA	1	3	0.333	0.003	NA
10	0	1	0	0.060	NA	0	0	NA	NA	NA	0	1	0	0.001	NA
11	6	37	0.162	0.703	NA	18	131	0.137	0.287	0.641	24	168	0.143	0.141	0.763
12	0	0	NA	NA	NA	0	3	0	0.009	NA	0	3	0	0.003	NA
13	0	0	NA	NA	NA	0	1	0	0.003	NA	0	1	0	0.001	NA
14	6	27	0.222	0.633	NA	17	112	0.152	0.256	0.604	23	139	0.165	0.119	0.727
Total or Mean	12	88	0.136	*	*	50	381	0.131	*	*	62	469	0.132	*	*
*	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability	*	*	*	Overall Reliability	Overall Reliability
*	Median n	0	*	0.355	NA	Median n	1	*	0.072	0.624	Median n	1.5	*	0.029	0.741
*	Min n	0	*			Min n	0	*			Min n	1	*		

Table 11. HCCN 1 facility-level reliability, postpartum most or moderately effective method use and provision.

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.590	0.047	0.047	0.120	0.313	0.467	0.653	0.775	0.832	0.887	0.945	0.972	0.992
Reliability (unit size≥50)	0.862	0.683	0.714	0.768	0.793	0.822	0.852	0.885	0.922	0.947	0.961	0.976	0.991
N of Entities (all units)	132	1	14	17	11	11	13	14	13	13	14	12	1
N of Entities (unit size≥50)	67	1	7	7	7	6	7	7	6	8	6	6	1
N of Persons (all units)	20648	1	14	47	103	197	505	984	1304	2085	5083	10326	2621
N of Persons (unit size≥50)	19832	50	409	539	623	646	945	1261	1701	3382	3414	6912	2621

Table 12. HCCN 1 facility-level reliability, postpartum LARC provision.

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.600	0.051	0.051	0.129	0.330	0.487	0.671	0.788	0.843	0.894	0.949	0.975	0.993
Reliability (unit size≥50)	0.892	0.744	0.771	0.817	0.838	0.862	0.886	0.912	0.942	0.960	0.971	0.982	0.993
N of Entities (all units)	132	1	14	17	11	11	13	14	13	13	14	12	1
N of Entities (unit size≥50)	67	1	7	7	7	6	7	7	6	8	6	6	1
N of Persons (all units)	20648	1	14	47	103	197	505	984	1303	2084	5083	10321	2621
N of Persons (unit size≥50)	19832	50	409	539	623	645	945	1261	1701	3382	3414	6912	2621

Table 13. HCCN 1 clinician group/practice-level reliability, postpartum most or moderately effective method use and provision.

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.400	0.048	0.048	NA*	0.091	0.151	0.250	0.414	0.581	0.748	0.835	0.923	0.974
Reliability (unit size≥50)	0.813	0.670	0.687	0.726	0.748	0.769	0.796	0.823	0.860	0.883	0.915	0.950	0.969
N of Entities (all units)	411	1	99	NA*	51	23	33	41	41	42	40	41	1
N of Entities (unit size≥50)	117	1	15	9	12	11	12	12	12	11	12	11	1
N of Persons (all units)	20648	1	99	NA*	102	82	222	592	1161	2542	4157	11691	758
N of Persons (unit size≥50)	18120	50	810	588	874	903	1156	1376	1809	2043	3226	5335	758

*There were only 9 deciles due to the distribution of reliability at this level.

Table 14. HCCN 1 clinician group/practice-level reliability, postpartum LARC provision.

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.421	0.055	0.055	NA*	0.105	0.173	0.281	0.452	0.618	0.777	0.855	0.933	0.978
Reliability (unit size≥50)	0.871	0.762	0.775	0.807	0.824	0.840	0.860	0.880	0.906	0.922	0.944	0.968	0.980
N of Entities (all units)	411	1	99	NA*	51	23	33	41	41	42	40	41	1
N of Entities (unit size≥50)	117	1	15	9	12	11	12	12	12	11	12	11	1
N of Persons (all units)	20641	1	99	NA*	102	82	222	592	1161	2542	4157	11691	758
N of Persons (unit size≥50)	18120	50	810	588	874	903	1156	1376	1809	2043	3226	5335	758

*There were only 9 deciles due to the distribution of reliability at this level.

Table 15. HealthEfficient facility-level reliability, postpartum most or moderately effective method use and provision. (Due to a small number of entities at this level, we are only presenting the minimum and maximum reliability)

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.799	0.786	*	*	*	*	*	*	*	*	*	*	0.815
N of Entities (all units)	3	1	*	*	*	*	*	*	*	*	*	*	1
N of Persons (all units)	469	144	*	*	*	*	*	*	*	*	*	*	172

* Cell left blank intentionally

Table 16. HealthEfficient facility-level reliability, postpartum LARC provision. (Due to a small number of entities at this level, we are only presenting the minimum and maximum reliability)

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.722	0.706	*	*	*	*	*	*	*	*	*	*	0.741
N of Entities (all units)	3	1	*	*	*	*	*	*	*	*	*	*	1
N of Persons (all units)	469	144	*	*	*	*	*	*	*	*	*	*	172

* Cell left blank intentionally

Table 17. HealthEfficient clinician group/practice-level reliability, postpartum most or moderately effective method use and provision. (Due to a small number of entities at this level after applying the unit size cutoff of 50, we are only presenting the minimum and maximum reliability for unit size≥50)

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.231	0.044	0.038	NA	NA	NA	NA	0.099	NA	0.847	0.852	0.870	0.870
Reliability (unit size≥50)	0.709	0.694	*	*	*	*	*	*	*	*	*	*	0.733
N of Entities (all units)	14	1	7	NA	NA	NA	NA	4	NA	1	1	1	1
N of Entities (unit size≥50)	3	1	*	*	*	*	*	*	*	*	*	*	1
N of Persons (all units)	469	1	7	NA	NA*	N	NA	11	NA	139	144	168	168
N of Persons (unit size≥50)	451	139	*	*	*	*	*	*	*	*	*	*	168

* Cell left blank intentionally
NA: There were only 5 deciles due to the distribution of reliability at this level.

Table 18. HealthEfficient clinician group/practice-level reliability, postpartum LARC provision. (Due to a small number of entities at this level after applying the unit size cutoff of 50, we are only presenting the minimum and maximum reliability for unit size≥50)

Item	Overall	Min	Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	Max
Reliability (all units)	0.029	0.001	0.001	NA	NA	NA	NA	0.020	NA	0.508	0.517	0.141	0.141
Reliability (unit size≥50)	0.741	0.727	*	*	*	*	*	*	*	*	*	*	0.763
N of Entities (all units)	14	1	7	NA	NA	NA	NA	4	NA	1	1	1	1
N of Entities (unit size≥50)	3	1	*	*	*	*	*	*	*	*	*	*	1
N of Persons (all units)	469	1	7	NA	NA	NA	NA	11	NA	139	144	168	168
N of Persons (unit size≥50)	451	139	*	*	*	*	*	*	*	*	*	*	168

* Cell left blank intentionally
NA: There were only 5 deciles due to the distribution of reliability at this level.

Appendix A:

An Alternative Reliability Analysis Method for Assessing Quality Measures

Sam Field, PhD; Pat Malone, PhD; Philip Hastings, PhD; Eric Booth, MA

Introduction

We derive an alternative reliability parameter using health service quality indicators as an example. We argue that this formulation is more consistent with the underlying data-generating process than a commonly utilized beta-binomial approach from the health service quality literature (Adams, 2009). Our alternative approach is more widely applicable and can include situations where the quality measures for a health care provider are averaged over a small number of observations/patients. We describe a straightforward implementation of our approach using the R statistical software package.

The Beta-Binomial model

The measures of service quality we employ all take the form of a binomial proportion:

$$\frac{y_i}{n_i},$$

where y_i is a count of patients who were provided a particular service in cluster i

$$i$$

, and

$$n_i$$

is the total number of patients in the cluster who were eligible to receive that service. The beta-binomial model begins with the assumption that the observed counts of services received by patients within cluster i

$$i$$

arise as a binomial random variable with parameters π_i and n_i

$$\pi_i$$

and

$$n_i$$

.

$$\begin{aligned}
 p(Y = y_i | n_i, \pi_i) \\
 = \textit{Binomial}(n_i, \pi_i)
 \end{aligned}$$

The approach further assumes that the cluster-level proportion parameters

$$\pi_i \text{ , or “true”}$$

quality scores for providers, are sampled from a population of quality scores across clusters that follow a beta distribution with parameters

$$\begin{aligned}
 &\alpha_0 \\
 &\beta_0
 \end{aligned}
 \text{ and }$$

$$\begin{aligned}
 p(\Pi = \pi_i | \alpha_0, \beta_0) \\
 = \textit{Beta}(\alpha_0, \beta_0)
 \end{aligned}$$

When a parameter of a random variable is itself a random variable, the statistical distribution of that parameter is known as a prior distribution. In this case, the beta distribution is a prior distribution for the

$$\pi_i \text{ in }$$

parameter the binomial distribution. Furthermore, the substitution of

$$\begin{aligned}
 &\alpha_0 \\
 &\beta_0 \\
 &\pi_i
 \end{aligned}$$

in the binomial

distribution leads to the beta-binomial distribution.

$$\begin{aligned}
 p(Y = y_i | n_i, \alpha_0, \beta_0) \\
 = \textit{BetaBinomial}(n_i, \alpha_0, \beta_0)
 \end{aligned}$$

Such a mixture of two statistical distribution does not always produce a third statistical distribution (i.e. beta-binomial) that is well-defined. When it does, the resulting distribution is called a "compound distribution" and

the prior distribution for the parameter is known as a "conjugate prior". Thus, the beta distribution is a conjugate prior of the binomial distribution, and the beta-binomial distribution is the resulting compound distribution.

In practice, the parameters

$$\alpha_0$$

and

$$\beta_0$$

are estimated

from the observed quality scores as

$$\alpha_0$$

and

$$\beta_0$$

. As an example,

we plot the density of the beta distribution with maximum likelihood estimates (MLE)

$$\alpha_0$$

and

$$\beta_0$$

of

$$\alpha_0$$

and

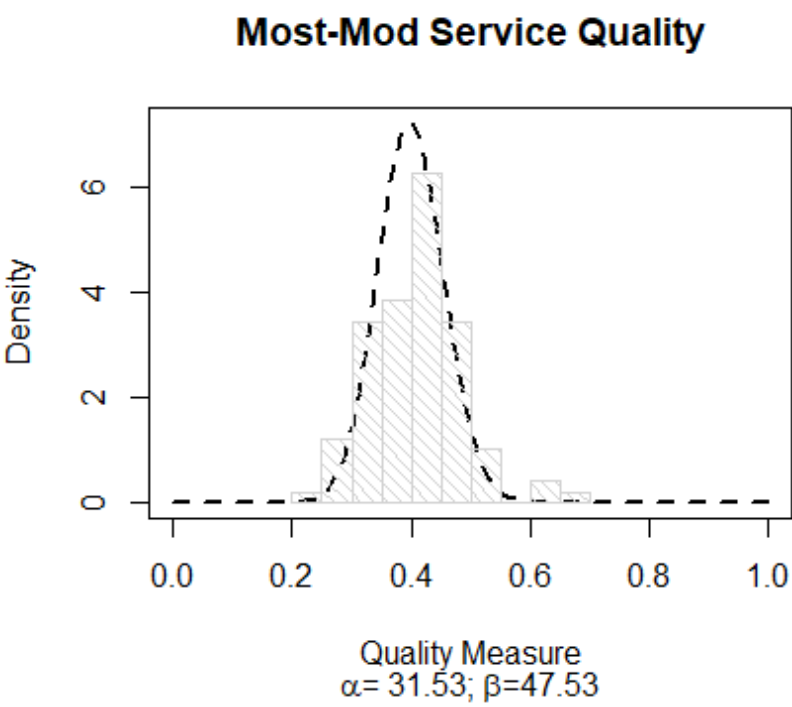
$$\beta_0$$

obtained from

a beta-binomial regression of a sample of service quality measures taken from 99 U.S. counties. The measures obtained indicate the proportion of eligible patients in each county that received at least one “Most or Moderately Effective Contraceptive Method” (Most-Mod) service over the course of a year.

Figure 1 depicts the distribution of observed quality measures as a histogram, while the fitted beta distribution is depicted as a continuous density plot. The mean, median, and mode of the distribution is close to .4 indicating that in the average county, approximately 40% of women receive at least one Most-Mod service. In addition, most of the county-level variation is restricted to an interval of .2 to .6. As can be seen from the plot, the fit to the beta distribution is approximate. Specifically, the fitted distribution does not capture a small cluster of counties with quality scores > .6. As an approximation, however, the beta distribution does appear to fit the observed data adequately.

Figure 1: Histogram with density plot overlay depicting the county-level distribution of Most-Mod service quality measures.

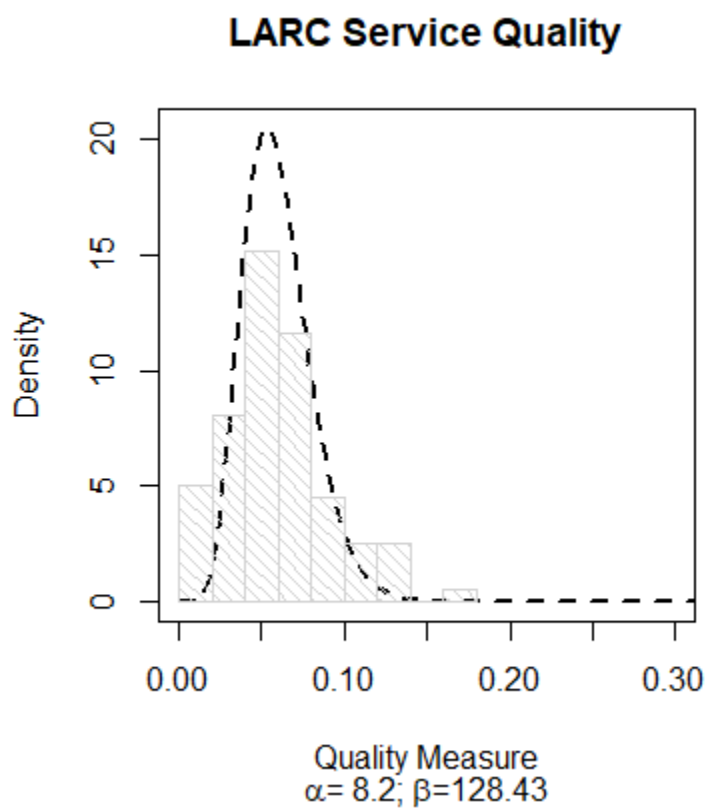


Although not evident in Figure 1, the range (i.e. x-axis) of the beta distribution is restricted to the 0,1 interval, which makes it a particularly suitable prior for a binomial parameter,

$$\pi_i$$

. This is easier to see if we plot a service quality indicator with an either very high or very low incidence. In Figure 2, we present the same plot for a service quality measure with a lower frequency, long-acting reversible contraception (LARC). In contrast to the estimated prior distribution for the Most-Mod service quality measure, the distribution of LARC quality measures is shifted considerably to the left with a noticeable right skew.

Figure 2: Histogram with density plot overlay depicting the county-level distribution of LARC service quality measures.



As Adams (2009) notes, the beta distribution is very flexible when it comes to fitting the observed distribution of service quality measures. Various combinations of the

α_0

and

$$\beta_0$$

parameters

generate a wide range of shapes - including U-shaped distributions where high and low-quality providers are widely separated from each other (Liu, et. al., 2013). Since this flexibility only requires the estimation of two parameters, the risk of over-fitting the observed data is minimal in cases where the number of clusters is large (e.g., > 30).

The predictive posterior distribution for health service quality

From a Bayesian perspective, predictions regarding the “true” cluster-level service quality score for any given cluster (e.g., provider or county) is based on the posterior predictive distribution (PPD) of

$$\pi_i$$

The PPD is the

distribution of possible values for

$$\pi_i$$

for each

cluster, conditional on the observed quality scores in that cluster,

$$\frac{y_i}{n_i}$$

as well as the

estimated parameters in the prior beta distribution (

$$\frac{\alpha}{0}$$

,

$$\frac{\beta}{0}$$

. Because of the conjugacy property discussed above, the PPD is analytically tractable. Specifically, the PPD for

$$\pi_i$$

is proportional to another beta distribution.

$$\begin{aligned} p(\Pi = \pi_i | n_i, y_i, \alpha_0, \beta_0) \\ \propto \textit{Beta}((\alpha_0 + y_i), (\beta_0 + (n_i - y_i))) \end{aligned}$$

In the current context, the PPD is the distribution of possible “true” cluster-level service quality scores based on sampling a single quality measure (i.e.,

$$\frac{y_i}{n_i}$$

) from a population of true quality scores that are assumed to follow a beta distribution with known or estimated

$$\alpha_0$$

parameters and

$$\beta_0$$

For the purpose of deriving reliability measures, we focus on the mean,

$$E(\pi_i)$$

, of this distribution:

$$E(\pi_i | n_i, y_i, \alpha_0, \beta_0) \\ = E(Beta((\alpha_0 + y_i), (\beta_0 + (n_i - y_i))))$$

We can substitute the analytically derived mean of the beta distribution,

$$\frac{\alpha_0}{(\alpha_0 + \beta_0)}$$

side of the equation and simplify the result. , in the right

$$E(\pi_i | n_i, y_i, \alpha_0, \beta_0) \\ = \frac{(\alpha_0 + y_i)}{(\alpha_0 + y_0) + (\beta_0 + (n_i - y_i))} \\ = \frac{(\alpha_0 + y_i)}{(\alpha_0 + \beta_0 + n_i)}$$

The empirical Bayes shrinkage estimator

Our derivation of the cluster-specific reliability estimate,

$$\lambda_i$$

commonly used identity for the empirical Bayes shrinkage estimator. Using this identity in the equation below (substituting empirically-estimated values , employs a

$$\frac{\alpha_0}{\beta_0}$$

and

population parameters

for unobserved

α_0 and β_0 , we define the mean of the posterior predictive distribution as equal to

a combination of the observed proportions, $\frac{y_i}{n_i}$, and the mean of the prior beta distribution for the "true" service quality scores,

π_i , weighted by the reliability λ_i ,

$$\frac{((\alpha_0 + y_i))}{(\alpha_0 + \beta_0 + n_i)} = \lambda_i \left(\frac{y_i}{n_i}\right) + \frac{(1 - \lambda_i)(\alpha_0}{(\alpha_0 + \beta_0}$$

As the reliability λ_i approaches 1 for a given cluster (nearing perfect reliability), the mean of the posterior predictive distribution for that cluster approaches the observed proportion. Conversely, for

λ_i values less than 1, the mean of the predictive posterior is "shrunk" towards the [estimated] mean of the prior distribution

$$\frac{\alpha_0}{(\alpha_0 + \beta_0)}.$$

For any given cluster, the more the shrinkage estimator is pulled towards the mean of the prior distribution, the less reliable the observed quality measures are for that cluster.

The classical test theory definition of reliability

In classical test score theory, reliability is defined as a ratio of true score variance to observed score variance (Novick, 1965):

$$\lambda_i = \frac{\sigma_{true}^2}{\sigma_{obs}^2}$$

In the beta-binomial model, the variance of the true score in the numerator is equal to the variance of the prior distribution for

$$\pi_i$$

. This is distributed as beta with an analytically derived variance:

$$\begin{aligned} \sigma_{true}^2 &= var(\pi_i) \\ &= \frac{(\alpha_0 \beta_0)}{(\alpha_0 + \beta_0)^2 * (\alpha_0 + \beta_0 + 1)} \end{aligned}$$

The variance in the observed service incidence,

$$y_i$$

, is the variance of the compound distribution - the beta-binomial. However, we need to derive the variance of the observed

$$\frac{y_i}{n_i}$$

proportions, . In the first step, we note that the variance of any random variable multiplied by a constant is the variance of the random variable times the square of the constant. Thus,

$$\begin{aligned}\sigma_{obs}^2 &= var\left(\frac{y_i}{n_i}\right) \\ &= var(y_i) \times \left(\frac{1}{n_i}\right)^2\end{aligned}$$

In the second step we replace

$$var(y_i)$$

with the

analytically derived variance of beta-binomial distribution. Thus,

$$\begin{aligned}\sigma_{obs}^2 &= var(y_i) \times \left(\frac{1}{n_i}\right)^2 \\ &= \left(\frac{n_i(\alpha_0\beta_0)(\alpha_0 + \beta_0 + n_i)}{(\alpha_0 + \beta_0)^2(\alpha_0 + \beta_0 + 1)}\right) \\ &\times \left(\frac{1}{n_i}\right)^2\end{aligned}$$

In the final step, the ratio of true score to observed score variance in the beta-binomial model,

$$\lambda_i$$

becomes:

$$\begin{aligned}\lambda_i &= \frac{\sigma_{true}^2}{\sigma_{obs}^2} \\ &= \frac{(\alpha_0\beta_0)}{(\alpha_0 + \beta_0)^2 * (\alpha_0 + \beta_0 + 1)} \bigg/ \left(\frac{n_i(\alpha_0\beta_0)(\alpha_0 + \beta_0 + n_i)}{(\alpha_0 + \beta_0)^2(\alpha_0 + \beta_0 + 1)}\right) \times \left(\frac{1}{n_i}\right)^2\end{aligned}$$

After simplifying we are left with a straight-forward expression for

$$\lambda_i$$

$$\lambda_i = \frac{n_i}{\alpha_0 + \beta_0 + n_i}$$

Algebraic proof of the identity above is available from the authors upon request. ¹

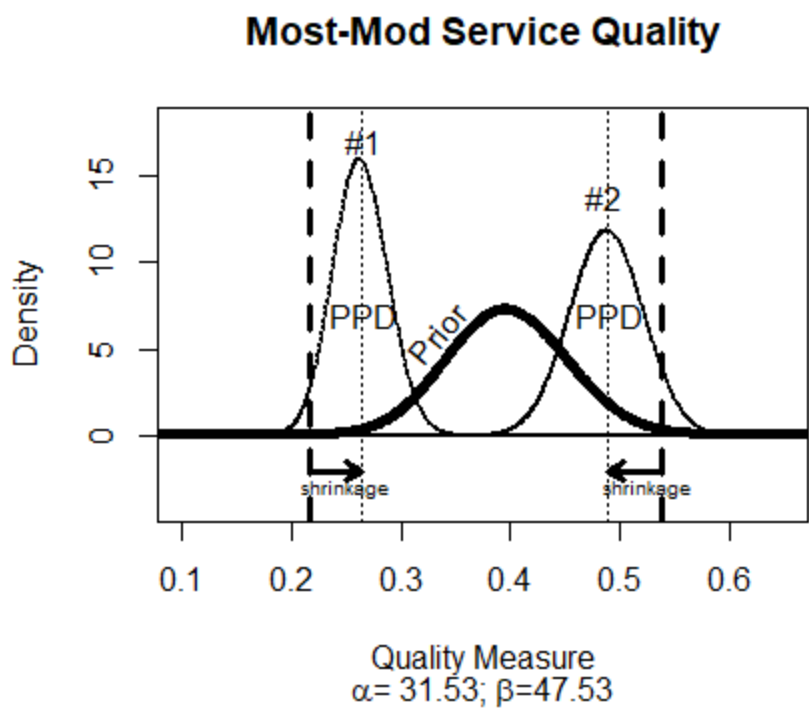
To illustrate the PPD and the empirical Bayes shrinkage estimator, we return to the 99-county example and the Most-Mod quality measures. In Figure 3, the PPDs for 2 different counties are plotted on top of the fitted prior distribution previously seen in Figure 1. We also indicate the means of each PPD as dotted vertical lines and of the observed quality measure (

$$\frac{y_i}{n_i}$$

) as thick, dashed vertical lines. The arrows indicate the direction and magnitude of the shrinkage of the EB estimate towards the mean of the prior distribution.

¹ Carlin and Louis (2000) also derive an expression for the beta-binomial reliability parameter. Although their derivation is based on a different parameterization of the beta prior distribution, when expressed as a function of α_0 and β_0 , their result is equivalent to ours (pp. 67-68).

Figure 3: PPD for two counties with a prior distribution overlay.



Although the general shapes of the plotted PPDs are similar, the more peaked distribution seen for the county on the left (county #1) reflects greater certainty about the location of that county’s “true” quality score. As our formulation of reliability indicates, this difference between the two counties is entirely a function of their difference in size. The numbers of patients in the left and right clusters are 230 and 141, respectively.

An alternative expression for Beta-Binomial reliability

In the sections above, we have derived a formulation for reliability that is consistent with the definition of the Bayesian shrinkage estimator as the mean of the posterior predictive distribution of the beta-binomial model. It is also consistent with the classical test theory, which views reliability as the proportion of total measurement variance that is attributable to true score variance. However, as we discuss below, it does not appear to be consistent with a formulation of beta-binomial reliability that often appears in the health care quality literature.

In a widely cited technical report from the RAND Corporation that was written for health care quality researchers, Adams (2009) offered an alternative formulation for beta-binomial reliability. Their approach was based on a least-squares formulation for reliability (Raudenbush & Bryk, 2002). Specifically,

$$\lambda_i = \frac{var(\pi_i)}{\left(var(\pi_i) + \frac{var\left(\frac{y_i}{n_i}\right)}{n} \right)}$$
$$var(\pi_i)$$

For , Adams
used the variance of the prior beta distribution.

$$var(\pi_i) = \frac{(\alpha_0\beta_0)}{(\alpha_0 + \beta_0)^2 * (\alpha_0 + \beta_0 + 1)}$$

$$var\left(\frac{y_i}{n_i}\right)$$

while was set
equal to the variance of the binomial distribution,

$$var\left(\frac{y_i}{n_i}\right) = \frac{y_i}{n_i} \times \left(1 - \frac{y_i}{n_i}\right)$$

Thus, the Adams formulation for beta-binomial reliability is:

$$\lambda_i = \frac{\left(\frac{(\alpha_0\beta_0)}{(\alpha_0 + \beta_0)^2 * (\alpha_0 + \beta_0 + 1)}\right)}{\left(\frac{(\alpha_0\beta_0)}{(\alpha_0 + \beta_0)^2 * (\alpha_0 + \beta_0 + 1)} + \frac{\left(\frac{y_i}{n_i} \times \left(1 - \frac{y_i}{n_i}\right)\right)}{n_i}\right)}$$

To demonstrate the practical consequences of the two different approaches to calculating beta-binomial reliability, we return to the 99-county example and the Most-Mod quality measure. In Figure 4, we have plotted county-level reliability parameters calculated using the formulation that we described against the results obtained under Adams’ (2009) approach. The diagonal represents the line of equality. The appearance of the plotting characters varies along two dimensions. The size of the characters is proportional to the natural log of the cluster size, ln(

$$n_i)$$
), while

character type identifies counties with observed measures (

$$\frac{y_i}{n_i}$$

) that are either

within or outside the interquartile range of the sample.

Figure 4: Comparison of reliability parameters: Most-Mod quality measure.

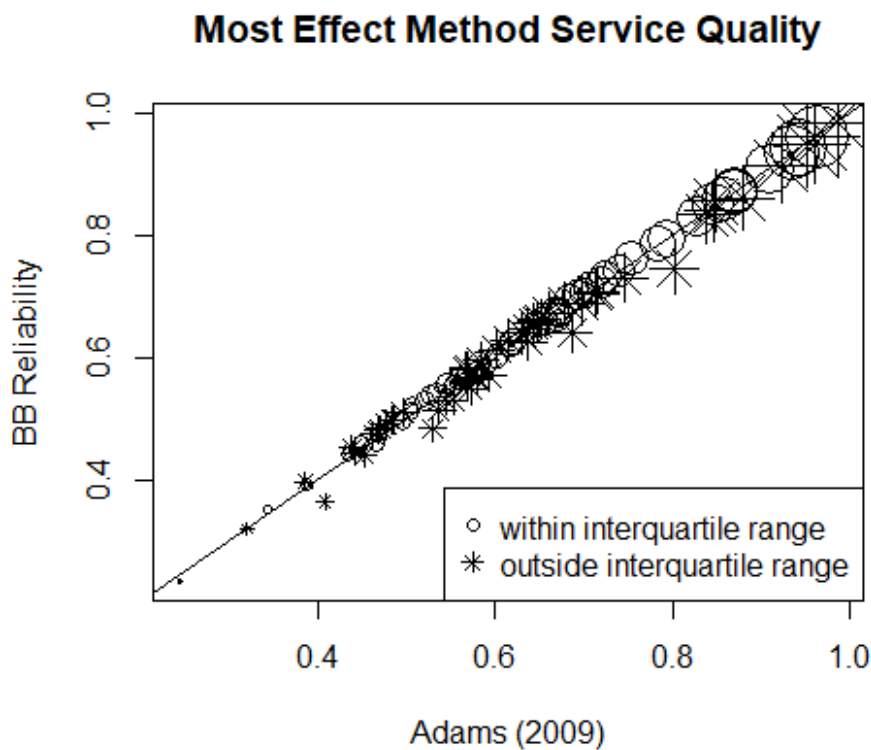
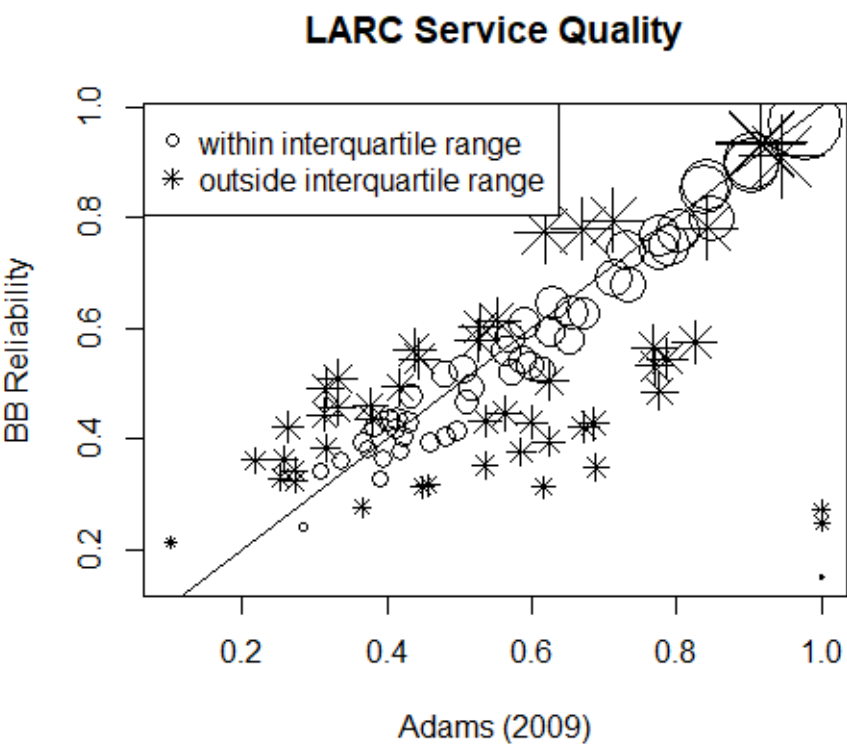


Figure 4 reveals very little difference between the two approaches. However, when we turn to the same plot for the LARC quality measure shown in Figure 5, substantial differences between the two methodologies emerge. Specifically, the disagreement between the methods appears greater in counties where the observed quality measure lies outside versus inside the inter-quartile range of the sample.

This pattern of results is expected because the reliability calculations that we describe depend *exclusively* on the size of the cluster while the Adams (2009) approach depends on both the size and the observed incident count. The results were not as evident with the Most-Mod quality measure because the binomial variance is relatively insensitive to variation in the quality measures within the central range of the 0,1 interval.

Figure 5: Comparison of reliability parameters: LARC quality measure.



When the variation in the cluster-level quality measures moves toward the boundaries, the between-cluster differences in the binomial variance component of the reliability parameter get much larger, and identically sized clusters can have dramatically different reliabilities. This is not possible with our calculations as identically sized clusters sampled from the same prior distribution will have equal reliability. It is also worth noting that the three small counties that appear in the bottom-right corner of the plot had perfect reliability under the Adams (2009) approach. This is because the variance of the binomial distribution equals zero when the observed measure is either 0 or 1.

Although the Adams (2009) formulation has been used to calculate beta-binomial reliability in previous empirical studies of health care quality (e.g., Adams and Paddock, 2017; Blair, et. al., 2015; Kazis, et. al., 2017; Staggs and Cramer, 2016), it has two distinct disadvantages when compared to the approach described here. First, we argue from statistical principle that the variance of the observed quality measures should be based on the beta-binomial compound distribution, and not the sum of the prior distribution (beta) and likelihood distribution (binomial) variances. Thus, the reliability formulation offered by Adams does not appear to be consistent with the underlying statistical model.

Second, under the Adams approach a component of measurement variance is determined by the observed data itself - the score,

$$\frac{y_i}{n_i}$$

. For clusters in which the observed measures equal 0 or 1, the binomial variance used in the calculations will equal zero, and the reliability measure will, consequently, be unstable in small clusters. The reliability calculations are also unstable when the bulk of the cluster-level variation lies close to the 0,1 boundaries - a situation that can be present even when the clusters are large. Indeed, Figure 5 demonstrates that the Adams (2009) approach can produce both substantially inflated and deflated estimates of reliability under such conditions. In contrast, the formulation that we offer does not depend on

$$y_i$$

and is equally valid across all cluster sizes.

Implementation

The approach to reliability calculation that we advocate is mathematically straightforward and can be implemented in many statistical software packages. In our implementation of the method, we use the statistical software R; specifically, we use the *vglm()* function in the “VGAM” package (Yee and Mohler, 2020). The procedure reparametrizes the beta-binomial distribution from

betabinomial($n_i \alpha_0, \beta_0$)

betabinomial(n_i, μ, γ)

to

, where

$$\mu = \frac{\alpha_0}{(\alpha_0 + \beta_0)}$$

, or the mean of

the prior distribution, and

$$\gamma = \frac{1}{(\alpha_0 + \beta_0 + 1)}$$

. The second

$$\gamma$$

parameter, γ , is interpreted as the overdispersion parameter or intra-cluster correlation coefficient (ICC), and it is possible to write reliability as a function of

$$\gamma$$

$$\lambda_i = \frac{n_i}{\left(\frac{1}{\hat{\gamma}} + (n_i - 1)\right)}$$

:

.

As the ICC coefficient,

$$\gamma$$

, approaches 1,

the reliability parameter also approaches 1, regardless of the within-cluster sample size.

The mean of the beta distribution,

$$\mu$$

, is linked to a linear combination of covariates via the inverse logit link.

$$\mu = e^{Xb} / (1 + e^{Xb})$$

Where X is a matrix of covariates (including a constant) and b is a vector of regression parameters. In our implementation, we exclude covariates and estimate an intercept-only model. The ICC is also estimated on the logit scale.

References

Adams, John L. (2009) *The Reliability of Provider Profiling: A Tutorial*. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/technical_reports/TR653.html.

Adams, J. L., & Paddock, S. M. (2017). Misclassification Risk of Tier-Based Physician Quality Performance Systems. *Health services research*, 52(4), 1277-1296.

Blair, R., Liu, J., Rosenau, M., Brannan, M., Hazelwood, N., Gray, K. F., ... & Schmitt, A. (2015). Development of Quality Measures for Inpatient Psychiatric Facilities. *Development*, 2, 04.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis (Vol. 88)*. Boca Raton: Chapman & Hall/CRC.

Kazis, L. E., Rogers, W. H., Rothendler, J., Qian, S., Selim, A., Edelen, M. O., ... & Butcher, E. (2017). *Outcome performance measure development for persons with multiple chronic conditions*. RAND.

Liu, C. F., Burgess Jr, J. F., Manning, W. G., & Maciejewski, M. L. (2013). Beta-Binomial Regression and Bimodal Utilization. *Health services research*, 48(5), 1769-1778.

Staggs, V. S., & Cramer, E. (2016). Reliability of pressure ulcer rates: how precisely can we differentiate among hospital units, and does the standard signal-noise reliability measure reflect this precision? *Research in nursing & health*, 39(4), 298-305.

Yee, T. & Moler, C. (2020_ VGAM: Vector Generalized Linear and Additive Models. <https://CRAN.R-project.org/package=VGA>

R syntax to calculate beta-binomial reliability

This is a very sparse R function that calculates reliability measures for binomial count
data assuming a beta-binomial distribution. The arguments in the function "y" and "n"
correspond to the total number of "successes" and "trials" respectively.

```
beta_rel <- function(y,n){  
  
  fit <- vglm(cbind(y,n-y) ~ 1, betabinomial) #estimate beta-binomial model  
  parms <- coef(fit, matrix = TRUE) #extract logit(mu) and logit(gamma)  
  
  #Inverse logit link  
  mu <- exp(coef(fit, matrix = TRUE)[1])/(1+exp(coef(fit, matrix = TRUE)[1]))  
  gamma <- exp(coef(fit, matrix = TRUE)[2])/(1+exp(coef(fit, matrix = TRUE)[2]))  
  theta <- gamma/(1-gamma)  
  
  #Derive Beta parameters  
  alpha=mu/theta;  
  beta=(1-mu)/theta;  
  
  #Calculate reliability  
  rel <- n/(alpha+beta+n)  
  
  cbind(y,n,id,rel)  
}
```