Synthetic Data Generation

Wentao Li

May 7, 2021

Data overview

Settings

There are 8 Settings of data in total as followings:

- Setting 1: 2 sites, 500 patients each site, small variance
- Setting 2: 2 sites, 500 patients each site, large variance
- Setting 3: 10 sites, 500 patients each site, small variance
- Setting 4: 10 sites, 500 patients each site, large variance
- Setting 5: 2 sites, 30 patients each site, small variance
- Setting 6: 2 sites, 30 patients each site, large variance
- Setting 7: 10 sites, 30 patients each site, small variance
- Setting 8: 10 sites, 30 patients each site, large variance

Data information

And in each data setting, data are consisted of

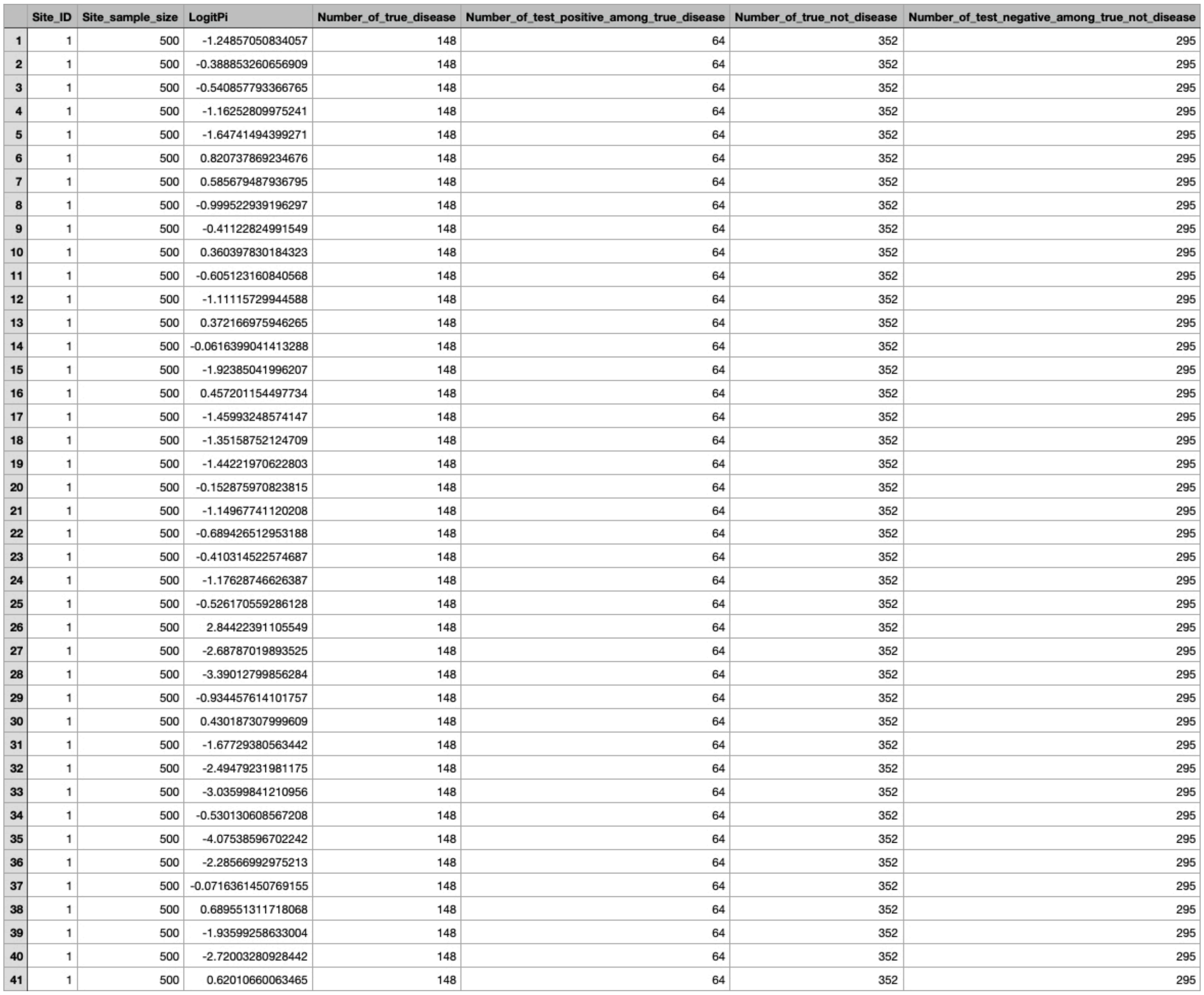
1. 4 categorical data, value in $\{0, 1\}$
2. 6 categorical data, value in range $[-1, 1.5] \in \mathbb{R}$
3. 1 outcome, value in $\{0, 1\}$
4. Site ID, represents the id of which site the entry belongs to

5. Site sample size, represents the number of samples in this specific setting

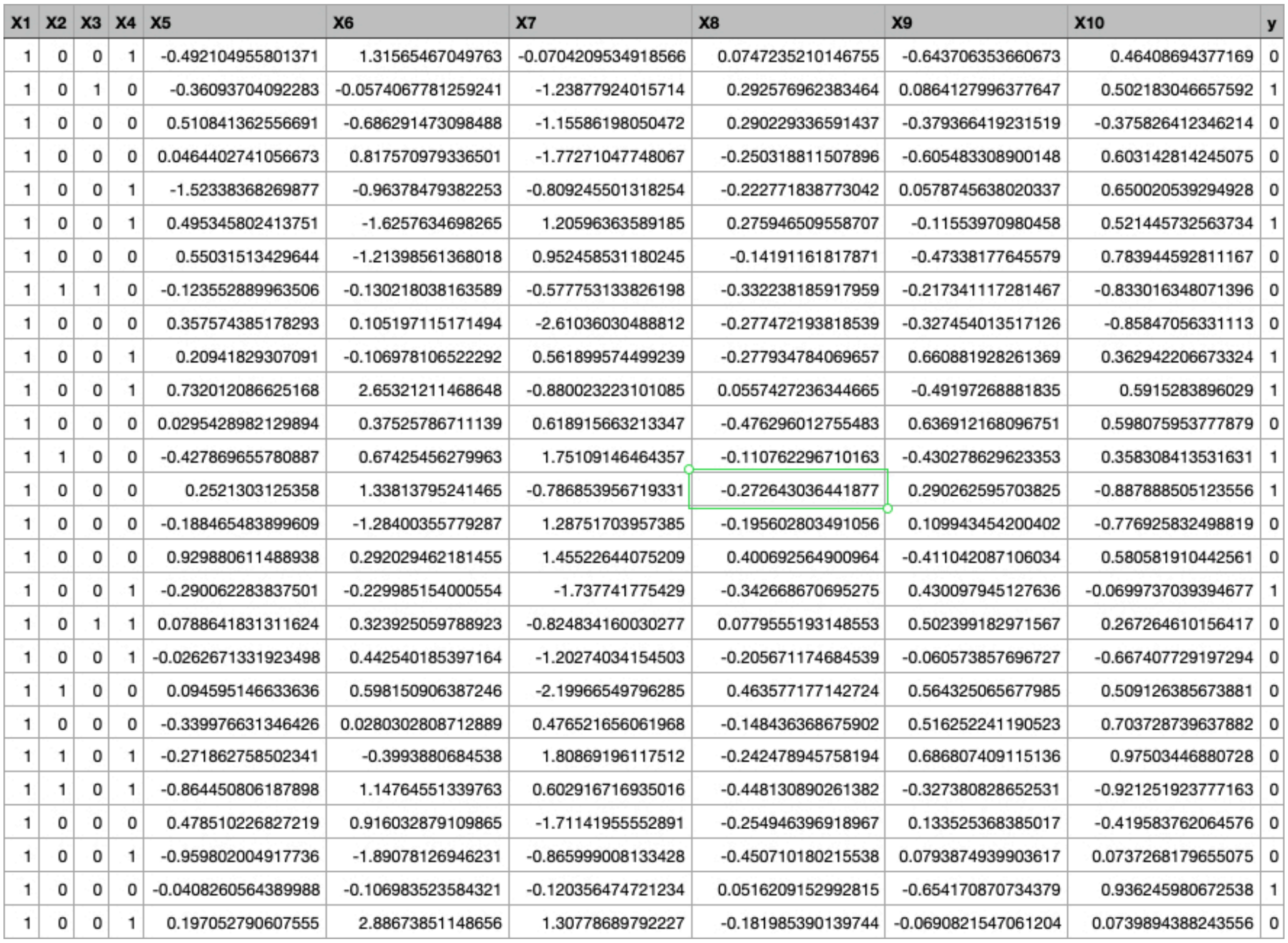
6. Log-odds ratio for each sample

7. Number of true positive, true negative, false positive, false positive, false negative

Screenshot



	Site_ID	Site_sample_size	LogitPi	Number_of_true_disease	Number_of_test_positive_among_true_disease	Number_of_true_not_disease	Number_of_test_negative_among_true_not_disease
1	1	500	-1.24857050834057	148	64	352	295
2	1	500	-0.388853260656909	148	64	352	295
3	1	500	-0.540857793366765	148	64	352	295
4	1	500	-1.16252809975241	148	64	352	295
5	1	500	-1.64741494399271	148	64	352	295
6	1	500	0.820737869234676	148	64	352	295
7	1	500	0.585679487936795	148	64	352	295
8	1	500	-0.999522939196297	148	64	352	295
9	1	500	-0.41122824991549	148	64	352	295
10	1	500	0.360397830184323	148	64	352	295
11	1	500	-0.605123160840568	148	64	352	295
12	1	500	-1.11115729944588	148	64	352	295
13	1	500	0.372166975946265	148	64	352	295
14	1	500	-0.0616399041413288	148	64	352	295
15	1	500	-1.92385041996207	148	64	352	295
16	1	500	0.457201154497734	148	64	352	295
17	1	500	-1.45993248574147	148	64	352	295
18	1	500	-1.35158752124709	148	64	352	295
19	1	500	-1.44221970622803	148	64	352	295
20	1	500	-0.152875970823815	148	64	352	295
21	1	500	-1.14967741120208	148	64	352	295
22	1	500	-0.689426512953188	148	64	352	295
23	1	500	-0.410314522574687	148	64	352	295
24	1	500	-1.17628746626387	148	64	352	295
25	1	500	-0.526170559286128	148	64	352	295
26	1	500	2.84422391105549	148	64	352	295
27	1	500	-2.68787019893525	148	64	352	295
28	1	500	-3.39012799856284	148	64	352	295
29	1	500	-0.934457614101757	148	64	352	295
30	1	500	0.430187307999609	148	64	352	295
31	1	500	-1.67729380563442	148	64	352	295
32	1	500	-2.49479231981175	148	64	352	295
33	1	500	-3.03599841210956	148	64	352	295
34	1	500	-0.530130608567208	148	64	352	295
35	1	500	-4.07538596702242	148	64	352	295
36	1	500	-2.28566992975213	148	64	352	295
37	1	500	-0.0716361450769155	148	64	352	295
38	1	500	0.689551311718068	148	64	352	295
39	1	500	-1.93599258633004	148	64	352	295
40	1	500	-2.72003280928442	148	64	352	295
41	1	500	0.62010660063465	148	64	352	295



X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	y
1	0	0	1	-0.492104955801371	1.31565467049763	-0.0704209534918566	0.0747235210146755	-0.643706353660673	0.46408694377169	0
1	0	1	0	-0.36093704092283	-0.0574067781259241	-1.23877924015714	0.292576962383464	0.0864127996377647	0.502183046657592	1
1	0	0	0	0.510841362556691	-0.686291473098488	-1.15586198050472	0.290229336591437	-0.379366419231519	-0.375826412346214	0
1	0	0	0	0.0464402741056673	0.817570979336501	-1.77271047748067	-0.250318811507896	-0.605483308900148	0.603142814245075	0
1	0	0	1	-1.52338368269877	-0.96378479382253	-0.809245501318254	-0.222771838773042	0.0578745638020337	0.650020539294928	0
1	0	0	1	0.495345802413751	-1.6257634698265	1.20596363589185	0.275946509558707	-0.11553970980458	0.521445732563734	1
1	0	0	0	0.55031513429644	-1.21398561368018	0.952458531180245	-0.14191161817871	-0.47338177645579	0.783944592811167	0
1	1	1	0	-0.123552889963506	-0.130218038163589	-0.577753133826198	-0.332238185917959	-0.217341117281467	-0.833016348071396	0
1	0	0	0	0.357574385178293	0.105197115171494	-2.61036030488812	-0.277472193818539	-0.327454013517126	-0.85847056331113	0
1	0	0	1	0.20941829307091	-0.106978106522292	0.561899574499239	-0.277934784069657	0.660881928261369	0.362942206673324	1
1	0	0	1	0.732012086625168	2.65321211468648	-0.880023223101085	0.0557427236344665	-0.49197268881835	0.5915283896029	1
1	0	0	0	0.0295428982129894	0.37525786711139	0.618915663213347	-0.476296012755483	0.636912168096751	0.598075953777879	0
1	1	0	0	-0.427869655780887	0.67425456279963	1.75109146464357	-0.110762296710163	-0.430278629623353	0.358308413531631	1
1	0	0	0	0.2521303125358	1.33813795241465	-0.786853956719331	-0.272643036441877	0.290262595703825	-0.887888505123556	1
1	0	0	0	-0.188465483899609	-1.28400355779287	1.28751703957385	-0.195602803491056	0.109943454200402	-0.776925832498819	0
1	0	0	0	0.929880611488938	0.292029462181455	1.45522644075209	0.400692564900964	-0.411042087106034	0.580581910442561	0
1	0	0	1	-0.290062283837501	-0.229985154000554	-1.737741775429	-0.342668670695275	0.430097945127636	-0.0699737039394677	1
1	0	1	1	0.0788641831311624	0.323925059788923	-0.824834160030277	0.0779555193148553	0.502399182971567	0.267264610156417	0
1	0	0	1	-0.0262671331923498	0.442540185397164	-1.20274034154503	-0.205671174684539	-0.060573857696727	-0.667407729197294	0
1	1	0	0	0.094595146633636	0.598150906387246	-2.19966549796285	0.463577177142724	0.564325065677985	0.509126385673881	0
1	0	0	0	-0.339976631346426	0.0280302808712889	0.476521656061968	-0.148436368675902	0.516252241190523	0.703728739637882	0
1	1	0	1	-0.271862758502341	-0.3993880684538	1.80869196117512	-0.242478945758194	0.686807409115136	0.97503446880728	0
1	1	0	1	-0.864450806187898	1.14764551339763	0.602916716935016	-0.448130890261382	-0.327380828652531	-0.921251923777163	0
1	0	0	0	0.478510226827219	0.916032879109865	-1.71141955552891	-0.254946396918967	0.133525368385017	-0.419583762064576	0
1	0	0	1	-0.959802004917736	-1.89078126946231	-0.865999008133428	-0.450710180215538	0.0793874939903617	0.0737268179655075	0
1	0	0	0	-0.0408260564389988	-0.106983523584321	-0.120356474721234	0.0516209152992815	-0.654170870734379	0.936245980672538	1
1	0	0	1	0.197052790607555	2.88673851148656	1.30778689792227	-0.181985390139744	-0.0690821547061204	0.0739894388243556	0

Notations

- $X_{1,2,3,4}$ are independent categorical data
- $X_{5,\ldots,10}$ are independent continuous data
- y is the outcome
- sen denotes the sensitivity
- sp denotes the specificity
- N is the sample size
- ε is the random error follows $\mathcal{N}(0,1)$
- β is the true variables in $\mathbb{R}^{10}$
- Σ is the covariance matrix of trivariate normal distribution, defined as identity matrix I_3

- μ is the random effect in $\mathbb{R}_3$ space

Data generation process

First, let's define the true sensitivity and specificity as

$$sen = 0.6 \quad \text{and} \quad sp = 0.9$$

and $\beta = (-1.5, 0.1, -0.5, -0.3, 0.4, -0.2, -0.25, 0.35, -0.1, 0.5)$.

Also define $X_1 = \mathbb{1}_N$ as the intercept, and X_2, X_3, X_4 are generated with Bernoulli distribution (1) with probability $p = 0.1, 0.3, 0.5$ respectively.

$$f(X_i; p) = \begin{cases} p & \text{if } X_i = 1 \\ q = 1 - p & \text{if } X_i = 0 \end{cases} \tag{1}$$

then X_5, X_6, X_7 are generated from normal distributions $\mathcal{N}(0, 0.5), \mathcal{N}(0, 1), \mathcal{N}(0, 1.5)$ respectively. Lastly, X_8, X_9, X_{10} are generate from uniform distributions $\mathcal{U}(-0.5, 0.5), \mathcal{U}(-0.7, 0.7)$, $\mathcal{U}(-1, 1)$ respectively.

We generate the random effect μ using trivariate normal distribution

$$\mathcal{N}_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right), \quad \Sigma = I_3 \tag{2}$$

and with the settings, we can deduce the log-odds ratio with following formula

$$\log(\pi) = f(X\beta + \mu_1 + \varepsilon) \tag{3}$$

where f is the sigmoid function defined as

$$f(x) = \frac{e^x}{1 + e^x} \tag{4}$$

Now, we generate the outcomes y for each sample with Bernoulli distribution (1) where the log-odds ratio served as the probability p. Also, the sensitivity and specificity can be calculated with Binomial distribution with probability $sen + \mu_2$ and $sp + \mu_3$.