Supplementary Materials for “On some pitfalls of the log-linear modeling framework for capture-recapture studies in disease surveillance”

Simulation Settings

Data were generated from the population-level multinomial:
$(N_{111}, N_{110}, N_{101}, N_{100}, N_{011}, N_{010}, N_{001}, N_{000}) \sim \text{Multinominal}(N, p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000})$.
The true number of cases N is set to $5{,}000$ under both scenarios, and capture probabilities are computed based on parameters $(p_1, p_{2|1}, p_{2|\bar{1}}, p_{3|12}, p_{3|1\bar{2}}, p_{3|\bar{1}2}, p_{3|\bar{1}\bar{2}})$:

$$
\begin{aligned}
p_{111} &= p_1 p_{2|1} p_{3|12} \\
p_{110} &= p_1 p_{2|1} \left(1 - p_{3|12}\right) \\
p_{101} &= p_1 \left(1 - p_{2|1}\right) p_{3|1\bar{2}} \\
p_{100} &= p_1 \left(1 - p_{2|1}\right)\left(1 - p_{3|1\bar{2}}\right) \\
p_{011} &= (1 - p_1) p_{2|\bar{1}} p_{3|\bar{1}2} \\
p_{010} &= (1 - p_1) p_{2|\bar{1}} \left(1 - p_{3|\bar{1}2}\right) \\
p_{001} &= (1 - p_1)\left(1 - p_{2|\bar{1}}\right)\psi \\
p_{000} &= (1 - p_1)\left(1 - p_{2|\bar{1}}\right)(1 - \psi),
\end{aligned}
$$

where under three-stream cases, $\psi = p_{3|\bar{1}\bar{2}}$.

Scenario 1

We assume the probability of having capture history $(0,1,1)$ is equal to the probability of having capture history $(0,1,0)$, and that the association between the first data stream and the third data stream is not affected by whether cases are identified by the second data stream. Converting to mathematical expressions, these stipulations correspond to setting the testable assumption $p_{011} = p_{010}$ i.e., $E(N_{011}) = E(N_{010})$, and the untestable assumption $p_{3|12}/p_{3|\bar{1}2} = p_{3|1\bar{2}}/\psi$. True values of the parameters are: $p_1 = 0.3, p_{2|1} = 0.2, p_{2|\bar{1}} = 0.3, p_{3|12} = 0.8, p_{3|1\bar{2}} = 0.16, p_{3|\bar{1}2} = 0.5, \psi = 0.1$.

Scenario 2

We impose two testable assumptions $E(N_{111}) = E(N_{101})$ and $E(N_{110}) = E(N_{100})$, and one untestable assumption which states that the key parameter $\psi = p_{3|1\bar{2}}/0.8$. This untestable assumption implies that, among those not identified by the second stream, cases are more likely to be captured by the third stream if they are not captured by the first stream, i.e., the first stream and third stream are negatively correlated conditional on a lack of capture by the second stream.

Table S1: Possible log-linear models for two-stream toy example data in Table 1.

		Fitted cell counts [b]				MLE of key parameters [c]		Results from the toy example data presented in Table 1 [d]			
Model [a]	Predictors	$\hat{N}_{11}$	$\hat{N}_{10}$	$\hat{N}_{01}$	$\hat{N}_{00}$	$\hat{\psi}$	$\hat{\phi}$	$\hat{\psi}$	$\hat{\phi}$	$\hat{N}$	AIC
1	Intercept only	$\frac{n_c}{3}$	$\frac{n_c}{3}$	$\frac{n_c}{3}$	$\frac{n_c}{3}$	$\frac{1}{2}$	1	$\frac{1}{2}$	1	1333	142.6
2	X_1	$\frac{n_{1\cdot}}{2}$	$\frac{n_{1\cdot}}{2}$	n_{01}	n_{01}	$\frac{1}{2}$	1	$\frac{1}{2}$	1	1250	111.7
3	X_2	$\frac{n_{\cdot 1}}{2}$	n_{10}	$\frac{n_{\cdot 1}}{2}$	n_{10}	$\frac{n_{\cdot 1}}{2n_{10}+n_{\cdot 1}}$	1	$\frac{1}{3}$	1	1500	26.8
4	X_1X_2	n_{11}	$\frac{n_{10}+n_{01}}{2}$	$\frac{n_{10}+n_{01}}{2}$	$\frac{n_{10}+n_{01}}{2}$	$\frac{1}{2}$	$\frac{4n_{11}}{2n_{11}+n_{10}+n_{01}}$	$\frac{1}{2}$	0.8	1375	111.7
5	X_1, X_2	n_{11}	n_{10}	n_{01}	$\frac{n_{10}n_{01}}{n_{11}}$	$\frac{n_{11}}{n_{11}+n_{10}}$	1	$\frac{1}{3}$	1	1500	28.8
6	X_1, X_1X_2	n_{11}	n_{10}	n_{01}	n_{01}	$\frac{1}{2}$	$\frac{2n_{11}}{n_{1\cdot}}$	$\frac{1}{2}$	$\frac{2}{3}$	1250	28.8
7	X_2, X_1X_2	n_{11}	n_{10}	n_{01}	n_{10}	$\frac{n_{01}}{n_{01}+n_{10}}$	$\frac{n_{11}(n_{10}+n_{01})}{n_{01}n_{1\cdot}}$	$\frac{1}{3}$	1	1500	28.8

[a] The intercept (α) is included in all models

[b] $n_c = n_{11} + n_{10} + n_{01}$, $n_{1\cdot} = n_{11} + n_{10}$, $n_{\cdot 1} = n_{11} + n_{01}$; analytic results in columns 3 through 8 reproduced from Lyles et al. (2021)

[c] $\hat{\psi} = \frac{\hat{N}_{10}}{\hat{N}_{10}+\hat{N}_{00}}$ and $\hat{\phi} = \frac{\hat{N}_{11}(\hat{N}_{01}+\hat{N}_{00})}{(\hat{N}_{11}+\hat{N}_{10})\hat{N}_{01}}$

[d] $\hat{N} = n_c + \exp(\hat{\alpha})$, where $\hat{\alpha}$ is the estimated intercept from fitting the log-linear model based on the toy example data presented in Table 1

Table S2: Possible log-linear models for three-stream CRC data when applying the usual conventions.

Model [a]	Predictors	MLE of key parameter $\psi = p_{3\|\overline{12}}$ [b]	Fitted N_{000} [c]
1	X_1, X_2, X_3	$\frac{\widehat{N}_{111}^{\frac{3}{8}}\widehat{N}_{101}^{\frac{1}{4}}\widehat{N}_{011}^{\frac{1}{4}}\widehat{N}_{001}^{\frac{1}{8}}}{\widehat{N}_{111}^{\frac{3}{8}}\widehat{N}_{101}^{\frac{1}{4}}\widehat{N}_{011}^{\frac{1}{4}}\widehat{N}_{001}^{\frac{1}{8}} + \widehat{N}_{110}^{\frac{1}{4}}\widehat{N}_{100}^{\frac{3}{8}}\widehat{N}_{010}^{\frac{3}{8}}}$	$\frac{\widehat{N}_{100}^{\frac{1}{2}}\widehat{N}_{010}^{\frac{1}{2}}\widehat{N}_{001}^{\frac{1}{2}}}{\widehat{N}_{111}^{\frac{1}{2}}}$
2	X_1, X_2, X_3, X_1X_2	$\frac{\widehat{N}_{111}^{\frac{1}{3}}\widehat{N}_{101}^{\frac{1}{3}}\widehat{N}_{011}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{3}}\widehat{N}_{101}^{\frac{1}{3}}\widehat{N}_{011}^{\frac{1}{3}} + \widehat{N}_{110}^{\frac{1}{3}}\widehat{N}_{100}^{\frac{1}{3}}\widehat{N}_{010}^{\frac{1}{3}}}$	$\frac{\widehat{N}_{110}^{\frac{1}{3}}\widehat{N}_{100}^{\frac{1}{3}}\widehat{N}_{010}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{3}}\widehat{N}_{101}^{\frac{1}{3}}\widehat{N}_{011}^{\frac{1}{3}}}\widehat{N}_{001}$
3	X_1, X_2, X_3, X_1X_3	$\frac{\widehat{N}_{111}^{\frac{1}{6}}\widehat{N}_{110}^{\frac{1}{6}}\widehat{N}_{011}^{\frac{2}{3}}\widehat{N}_{001}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{6}}\widehat{N}_{110}^{\frac{1}{6}}\widehat{N}_{011}^{\frac{2}{3}}\widehat{N}_{001}^{\frac{1}{3}} + \widehat{N}_{101}^{\frac{1}{6}}\widehat{N}_{100}^{\frac{1}{6}}\widehat{N}_{010}}$	$\frac{\widehat{N}_{101}^{\frac{1}{3}}\widehat{N}_{100}^{\frac{1}{3}}\widehat{N}_{001}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{3}}\widehat{N}_{110}^{\frac{1}{3}}\widehat{N}_{011}^{\frac{1}{3}}}\widehat{N}_{010}$
4	X_1, X_2, X_3, X_2X_3	$\frac{\widehat{N}_{111}^{\frac{1}{6}}\widehat{N}_{110}^{\frac{1}{6}}\widehat{N}_{101}^{\frac{2}{3}}\widehat{N}_{001}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{6}}\widehat{N}_{110}^{\frac{1}{6}}\widehat{N}_{101}^{\frac{2}{3}}\widehat{N}_{001}^{\frac{1}{3}} + \widehat{N}_{100}\widehat{N}_{011}^{\frac{1}{6}}\widehat{N}_{010}^{\frac{1}{6}}}$	$\frac{\widehat{N}_{011}^{\frac{1}{3}}\widehat{N}_{010}^{\frac{1}{3}}\widehat{N}_{001}^{\frac{1}{3}}}{\widehat{N}_{111}^{\frac{1}{3}}\widehat{N}_{110}^{\frac{1}{3}}\widehat{N}_{101}^{\frac{1}{3}}}\widehat{N}_{100}$
5	$X_1, X_2, X_3, X_1X_2, X_1X_3$	$\frac{\widehat{N}_{011}}{\widehat{N}_{011} + \widehat{N}_{010}}$	$\frac{\widehat{N}_{010}\widehat{N}_{001}}{\widehat{N}_{011}}$
6	$X_1, X_2, X_3, X_1X_2, X_2X_3$	$\frac{\widehat{N}_{101}}{\widehat{N}_{101} + \widehat{N}_{100}}$	$\frac{\widehat{N}_{100}\widehat{N}_{001}}{\widehat{N}_{101}}$
7	$X_1, X_2, X_3, X_1X_3, X_2X_3$	$\frac{\widehat{N}_{110}\widehat{N}_{101}^{\frac{1}{4}}\widehat{N}_{011}^{\frac{1}{4}}\widehat{N}_{001}^{\frac{3}{4}}}{\widehat{N}_{110}\widehat{N}_{101}^{\frac{1}{4}}\widehat{N}_{011}^{\frac{1}{4}}\widehat{N}_{001}^{\frac{3}{4}} + \widehat{N}_{111}^{\frac{1}{4}}\widehat{N}_{100}\widehat{N}_{010}}$	$\frac{\widehat{N}_{100}\widehat{N}_{010}}{\widehat{N}_{110}}$
8	$X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$	$\frac{\widehat{N}_{110}\widehat{N}_{101}\widehat{N}_{011}}{\widehat{N}_{110}\widehat{N}_{101}\widehat{N}_{011} + \widehat{N}_{111}\widehat{N}_{100}\widehat{N}_{010}}$	$\frac{\widehat{N}_{111}\widehat{N}_{100}\widehat{N}_{010}\widehat{N}_{001}}{\widehat{N}_{110}\widehat{N}_{101}\widehat{N}_{011}}$

[a] The intercept (α) is included in all models

[b] $\widehat{N}_{ijk}$ denotes the fitted cell count with capture history (i, j, k) and is obtained by computing estimated $E(N_{ijk})$ from the fitted log-linear model, where $i, j, k \in \{0,1\}$; Under models 1- 4, and 7 (which are unsaturated models), fitted cell counts do not have closed form and can be computed by numerically maximize the Poisson log-likelihood; while fitted cell counts $\widehat{N}_{011} = n_{011}$ and $\widehat{N}_{010} = n_{010}$ under the model 5, and fitted cell counts $\widehat{N}_{101} = n_{101}$ and $\widehat{N}_{100} = n_{100}$ under the model 6; the saturated model 8 yields each of fitted cell counts equal to its corresponding observed cell count.

[c] Fitted N_{000} is computed as $\exp(\hat{\alpha})$, where $\hat{\alpha}$ is the MLE of α

Reference:

Lyles, Robert H, Amanda L Wilkinson, John M Williamson, Jiandong Chen, Allan W Taylor, Amara Jambai, Mohamed Jalloh, and Reinhard Kaiser. 2021. "Alternative Capture-Recapture Point and Interval Estimators Based on Two Surveillance Streams." In *Modern Statistical Methods for Health Research*, 43-81. Springer.