

Pooled Sample Screening with Quality Control

Wesley O. Johnson
University of California, Davis

Timothy Hanson
Department of Mathematics and Statistics
University of New Mexico

Joseph L. Gastwirth
George Washington University
and
Ian A. Gardner
University of California, Davis

Background

- Screening for HIV in blood donors is done one unit at a time by ELISA test; costs about \$1 per unit
- Seropositive units are re-tested with Western-blot test (“Gold Standard”); costs about \$20 per unit
- Seronegative units are not re-tested
- False negative units are not discovered until HIV infection based on transfusion
- Prevalence of HIV among negatives and sensitivity of ELISA test used in the field are difficult to estimate with precision under this protocol cf. Johnson and Gastwirth (1991, 2000), Gastwirth, Johnson and Reneau (1991), Gastwirth and Johnson (1994), Johnson and Pearson (1999)

New Screening Protocol

- Blood is collected at I sites around the country
- As prevalence of HIV varies from region to region, the I sites may be broken down into a few regions
- n_i units sampled at site i for $i = 1, \dots, I$.
- These are pooled into m_i pools of size k
- Disease presence (absence) denoted by $D(\bar{D})$
- Screening test determines if pools are “reactive” (+), or not (-)

- If a pool tests (+), a “gold standard” (GS) test is applied
- If pool is GS negative, units in pool are considered to be \bar{D}
- If pool is GS (+), individual units are tested with the GS to determine the true status
- Negative pools may or may not be truly negative

Quality Control Stage

- Need statistical information about false negative pools
- Randomly sample units from (-) pools with probability f
- These selected FS negative units are then re-pooled into larger pools and re-tested with GS
- Individual units in positive pools are re-tested with GS

The Data

- Number of true (+), x_{itp} , false (+), x_{ifp} , and (-), x_{in} , first stage pools
- Number of diseased units in true (+) pools, x_{iD} , and number that are not, $x_{i\bar{D}}$.
- Number of units that were re-sampled from FS negative pools, $r_{il} \geq 0$ for pool l at site i , and out of these, the number that are actually D , s_{il} .

	D		
(+)	x_{itp}	x_{ifp}	
(-)	*	*	x_{in}
			m_i

$$x_{itp} \rightarrow (x_{iD}, x_{i\bar{D}}), \quad x_{in} \rightarrow (r_{il}, s_{il}), \quad l = 1, \dots, x_{in}$$

Two-Stage Pooling

- Prevalences may vary from site to site

$$pr(D|\text{site } i) = \pi_i, i = 1, \dots, I$$

- Test accuracies (sensitivity and specificity) remain constant

$$pr(+|\geq 1 D) = \eta, \quad pr(-|\text{no } D's) = \theta$$

- Sensitivity assumed to be the same regardless of # D 's in pool
- Prevalence expected to be low
- Pool size selected so that usually either 0 or 1 D 's in pool

Statistical Inference

- Build in “latent” data \rightarrow “nice” likelihood
- Bayesian inferences via Gibbs sampling
- Sensitivity/Specificity prior: η and $\theta \stackrel{\perp}{\sim}$ beta
- Prevalence prior: (parametric approach)

$$\pi_i \stackrel{\text{iid}}{\sim} \text{beta}(\alpha, \beta)$$

- Prevalence prior: (nonparametric approach)

$$\pi_i \stackrel{\text{iid}}{\sim} G, \quad G|M, G_{(\alpha, \beta)} \sim \text{DP}(MG_{(\alpha, \beta)}), \quad M \sim \Gamma$$

- $p(\alpha, \beta)$ proper prior cf. Hanson et.al.(2000) or

$$\propto (\alpha + \beta)^{-5/2}$$

Latent Data

- Number of D 's in negative group l from site i , $\{z_{il}\}$
- Total # D 's from (-) gps at site i , z_i .
- $z_{fn\{il\}} = I_{\{>0\}}(z_{li})$, indicates whether or not pool (li) is false negative or not; total number of false negative pools, $z_{fn\cdot} = \sum_{il} z_{fn\{li\}}$
- Augmented data likelihood:

$$Lik_{aug} = \prod_i \pi_i^{x_i + z_i} (1 - \pi_i)^{n_i - x_i - z_i}$$

$$\eta^{x_{tp\cdot}} (1 - \eta)^{z_{fn\cdot}} \theta^{x_{n\cdot} - z_{fn\cdot}} (1 - \theta)^{x_{fp\cdot}}$$

Full Conditionals

- Denote Z as the latent data, V as that observed data, and π as the collection $\{\pi_i\}$

- Sensitivity and Specificity:

$$\eta|Z, V \sim \text{beta}(a_\eta + x_{tp.}, b_\eta + z_{fn.}),$$

$$\theta|Z, V \sim \text{beta}(a_\theta + x_{n.} - z_{fn.}, b_\theta + x_{fp.})$$

- Prevalences:

$$\pi_i|Z, V, \alpha, \beta \stackrel{\perp}{\sim} \text{beta}(\alpha + x_{iD} + z_{i.}, \beta + n_i - x_{iD} - z_{i.})$$

- Hyperparameters:

$$p(\alpha, \beta|Z, V, \pi) \propto \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^I \prod_i \pi_i^\alpha (1 - \pi_i)^\beta p(\alpha, \beta)$$

Full Conditional for Latent Data

- $z_{il}|V, \pi$ are \perp and depend on $u_{il} \equiv (r_{il}, s_{il})$ pair
- Let t_{ui} denote the number of pools at site i which have the same $u=(r,s)$ values

- Then

$$z_{ui.} \equiv \sum_{l:u_{li}=u} z_{il} \sim \text{Mult}(t_{ui}, q_{ui}),$$

- where q_{ui} is vector of probabilities

$$\text{pr}(z_{il} = j|u, \text{pool } l, \pi, \eta, \theta) = \frac{(1 - \eta)p(j)q(j; u)I\{s \leq j \leq k - (r - s)\}}{(1 - \eta) \sum_{j=\{1 \vee s\}}^{k-(r-s)} p(j)q(j; u) + \theta p(0)I\{s = 0\}},$$

- and

$$\frac{\theta p(0) I\{j = s = 0\}}{(1 - \eta) \sum_{j=\{1 \vee s\}}^{k-(r-s)} p(j) q(j; u) + \theta p(0) I\{s = 0\}}$$

- where

$$p(j) = \binom{k}{j} \pi_i^j (1 - \pi_i)^{k-j}, \quad q(j; u) = \binom{j}{s} \binom{k-j}{r-s} / \binom{k}{r}$$

- For the nonparametric (MDP) model, the full conditionals for π and for (α, β) are replaced by those given in Hanson et. al. (2000)
- (η, θ, π) easily sampled
- (α, β) sampled by Metropolis-Hastings or sample $\alpha|\beta$ and $\beta|\alpha$ via the Gilks and Wild (1992) adaptive rejection algorithm.

Illustrations

Johnson and Gastwirth (2000) Proposal

- Dual-group screening (frequentist inferences)
- Screening protocol similar to that proposed here
- Imperfect second stage test is allowed (sensitivity η_g , specificity θ_g)
- They allow for only one site and one prevalence
- They argued cost-effectiveness of dual-group screening
-



In this approach, a QC screening test is used to test pooled samples. This test will have its own sensitivity, η_g , and specificity θ_g . When a pool at the second stage tests positive, the pool is immediately re-tested with a GS test and is subsequently treated in the same way as the protocol we previously described in this paper. This results in similar data to the kind we obtained earlier. The new wrinkle is that second-stage pools that test negative are not re-tested, so it is never known precisely which units in these pools are truly \bar{D} . We thus have additional “latent” data to consider before we can make inferences. The new latent data is the information about which units in second-stage negative pools are D 's versus \bar{D} 's.

Cost Considerations In this section, we consider the expected cost of using the suggested

protocols from sections 2 and 4, and we compare with two other protocols, namely, no pooling at all, which is the current method that is used for HIV testing, and a protocol that was suggested by Gastwirth and Johnson (1994) and further considered by Johnson and Pearson (1999), where a QC stage is added to the usual no-pooling protocol.

*References

- Gastwirth, J.L. (1987). "The statistical precision of medical testing procedures: Applications to polygraph/AIDS antibodies test data," *Statistical Science*, 2, 213-38.
- Gastwirth, J.L. and Hammick, P.A. (1989). "Estimation of the prevalence of rare disease preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors," *Journal of Statistical Planning and Inference*, 22, 15-27.

Gastwirth, J.L., Johnson, W.O. and Reneau, D.M. (1991). "Bayesian analysis of screening data: Applications to AIDS in blood donors." *Canadian Journal of Statistics*, 19, 135-50.

Gastwirth, J.L. and Johnson, W.O. (1994). "Screening with cost-effective quality control: Potential applications to HIV and drug testing." *Journal of the American Statistical Assoc.*, 89, 972-81.

Geisser, Seymour and Johnson, Wesley O. (1992), "Optimal administration of dual screening tests for detecting a characteristic with special reference to low prevalence diseases," *Biometrics* **48**, 839-852.

Hammick, P.A. and Gastwirth, J.L. (1994). "Extending the applicability of estimation of the prevalence of sensitive characteristics by group testing to moderate prevalence populations," *International Statistical Review*, 62, 319-31.

Hughes-Oliver, J.M. and Swallow, W.H. (1994). "A two-stage adaptive group-testing procedure for estimating small proportions,"

Journal of the American Statistical Association, 89, 982-93

Johnson, W.O. and Gastwirth (1991). "Bayesian inference for medical screening tests: Approximations useful for the analysis of Acquired Immune Deficiency Syndrome," *Journal of the Royal Statistical Society, Ser. B, 53, 427-39.*