Determining Limit of Detection of High Throughput Sequencing Diagnostics with MiFi®

Huizi Wang[1], Josh Habiger[1], Andres Espindola[2], Kitty Cardwell[2], Tyler Dang[3], Georgios Vidalakis [3] and Avijit Roy[4]

1. Department of Statistics, Stillwater, OK 74078
2. Institute of Biosecurity and Microbial Forensics, Oklahoma State University, Stillwater
3. Department of Microbiology and Plant Pathology, University of California, Riverside
4. APHIS PPQ, Riverdale, MD

Abstract

High throughput sequencing (HTS) technology can be applied to plant disease diagnostics. Microbe Finder (MiFi®) is an online platform for detection of plant pathogens in HTS data, eliminating pathogen isolation, bioinformatics, amplification. Diagnostic sensitivity, specificity and Limit of Detection (LOD) are crucial metrics of any diagnostic tool. We present how to calculate LOD for HTS diagnostics with a statistical inference model using pathogen-specific e-probe matches in metagenomic data. The LOD calculates the lowest levels of the target pathogen that can be reliably detected. Here we used a quadratic discriminant analysis to calculate the LOD of three citrus pathogens in metagenomic HTS data. The LOD assumes that positive samples have a higher e-probe 'hit x percent identity score' and a different Normal distribution than the negative control scores. LOD, formally defined as the estimated Bayes decision boundary, is computed using the mean and variance of the positive and negative groups.

The LOD of citrus leprosis virus C2, citrus tristeza virus, and citrus exocortis viroid were 4.7, 4.2, and 5.1 scores/10000 respectively, indicating when the chance of positive is 50/50. The LOD results were consistent with the RT-qPCR results, however MiFi® was found to be more sensitive. In this scenario, the model is trained on a viroid and two RNA viruses, but is assumed to be true for all taxonomic groups. The development of the probability model for citrus graft transmissible bacteria and a citrus specific oomycete (*Phytophthora spp*) is on-going.

Objectives

- Develop and validate a probability algorithm to generate a Limit of Detection (LOD).
- Test the algorithm with known positive and negative metagenomic sequence data of containing citrus and citrus pathogen nucleic acids.
- Determine if size of the pathogen, relative to the host, will provide an equivalent LOD across pathosystems.

Methods

- Infected and healthy citrus tissue was sequenced using HTS. Briefly, positive samples contained citrus leprosis virus C2, citrus tristeza virus and citrus exocortis viroid.
- HTS data obtained from the sequenced samples were analyzed using the MiDetect™ to retrieve hits and scores.
- Hit: Reads hitting with a selected e-probe sequence with a minimum e-value
- Score: Calculated for each hit based on percent identity and query coverage.
- Scores are generated for each e-probe sequence, which was added to retrieve the total score for each pathogen.
- The probability that a pathogen is positive/negative is calculated using the scores obtained for the pool of positive and negative samples for each virus.
- The formula used to calculate the LOD at which pathogen is present/absent is:

$$LOD = x = \frac{\left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right) - \sqrt{\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2\sigma_2^2} - \left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) \times 2log\frac{\sigma_2}{\sigma_1}}}{\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right)}$$

- The LOD is the Bayesian decision boundary.

Results

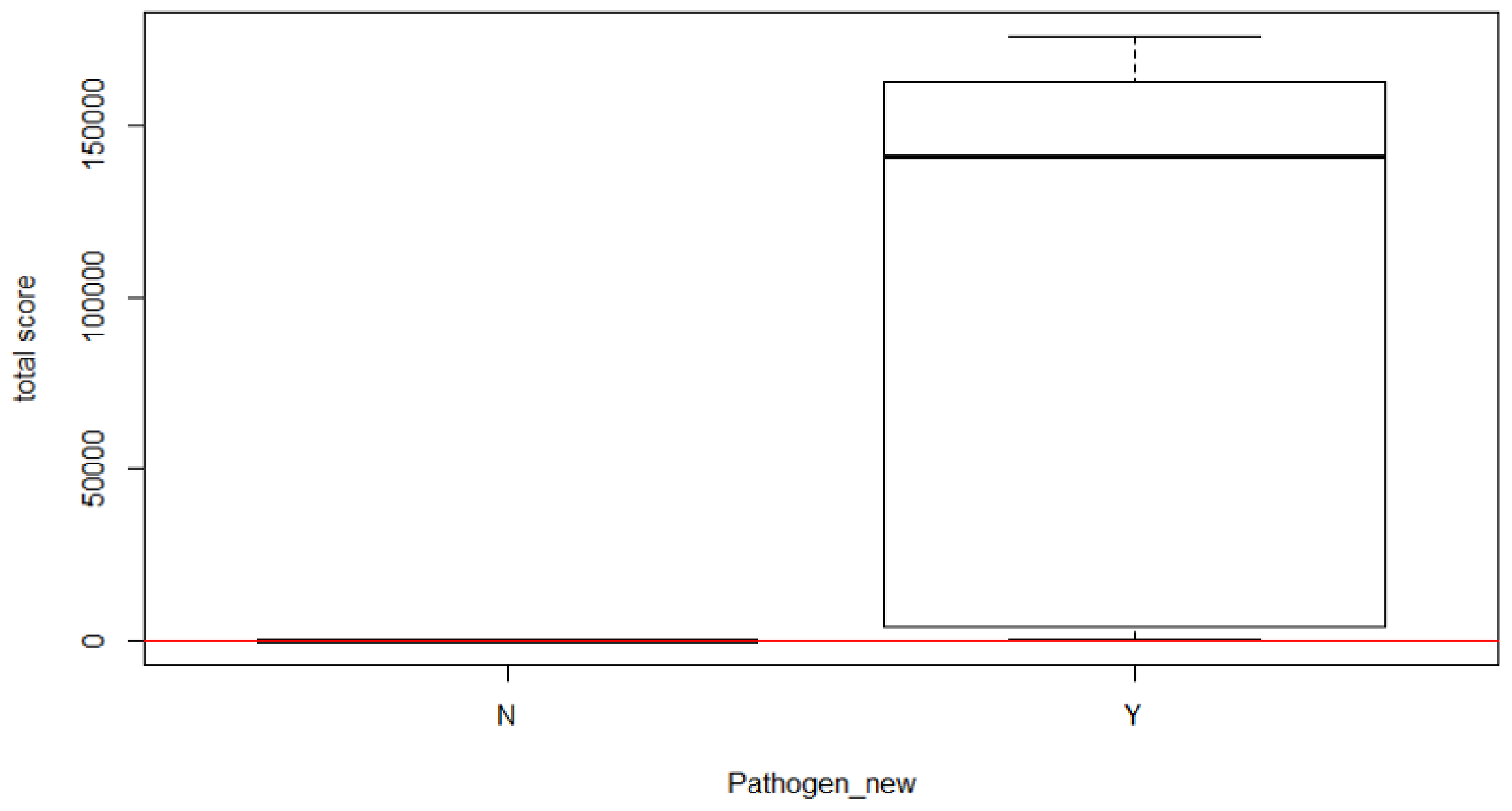

Figure 1. Boxplot of CTV: We have two different normal distributions with different variances. This suggests that a quadratic discriminant analysis is reasonable. The LOD of score/10000 is 4.2 (red line).

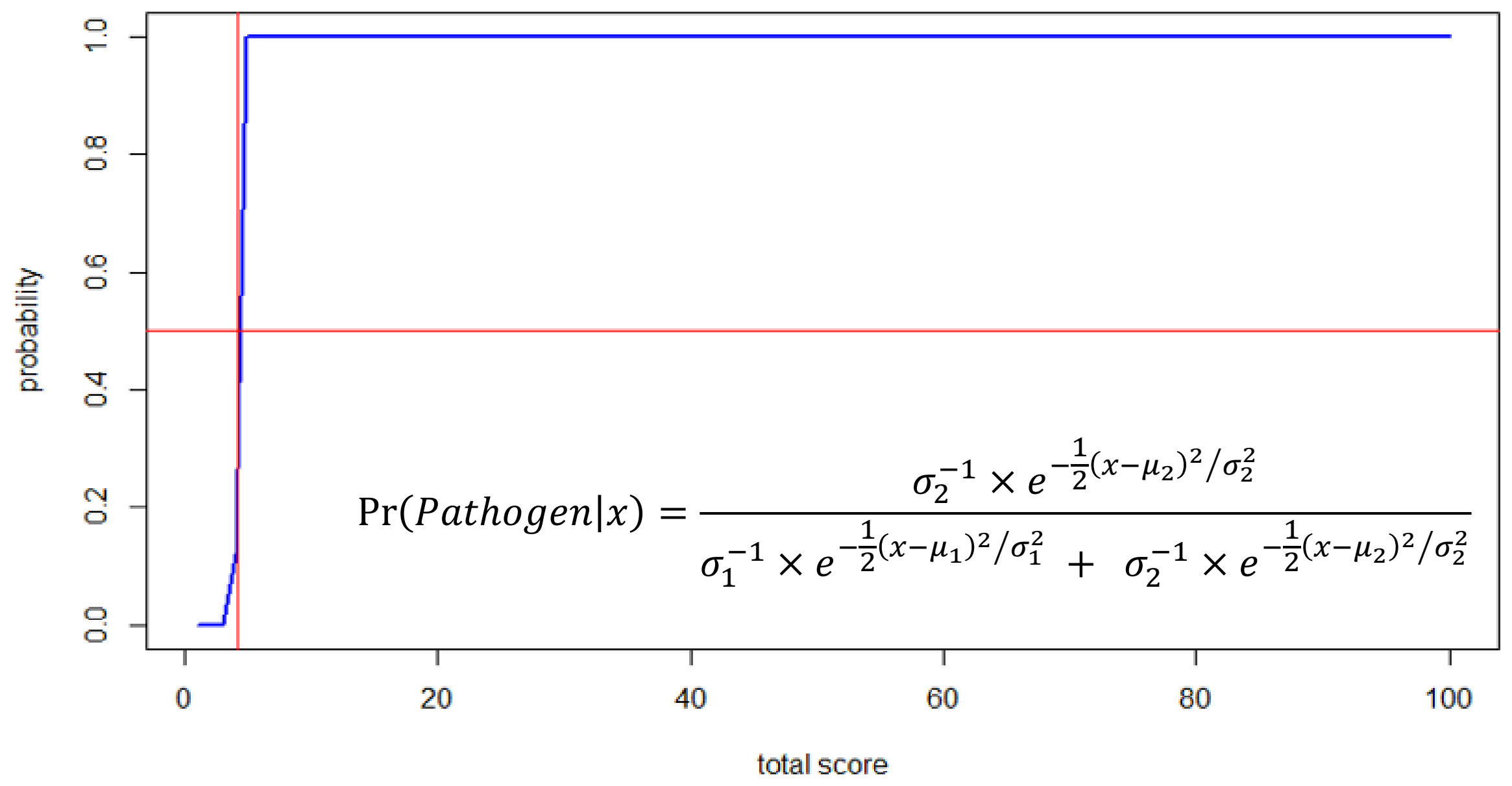

Figure 2. Plot the Probability of CTV for Given Total Score p(total score): The probability is on Y-axis. The horizontal line satisfies p(total score) = 0.5. The LOD is the solution to Pr(Pathogen|x) = 0.5. A stringency of 80% probability barely changes the score.

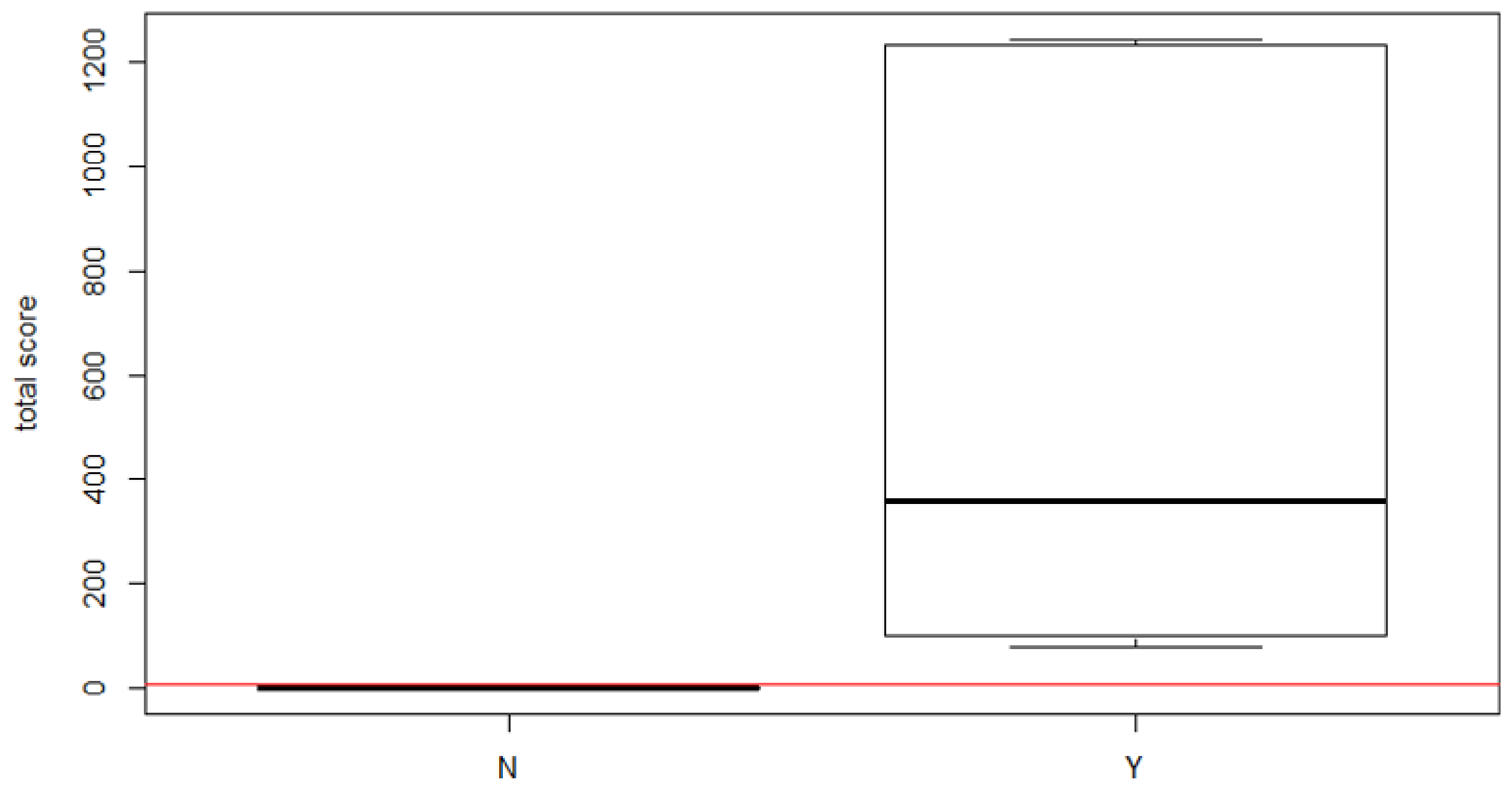

Figure 3. Boxplot of CiLV-C2: We have two different normal distributions with different variances. This suggests that a quadratic discriminant analysis is reasonable. The LOD of score/10000 is 4.7 (red line).

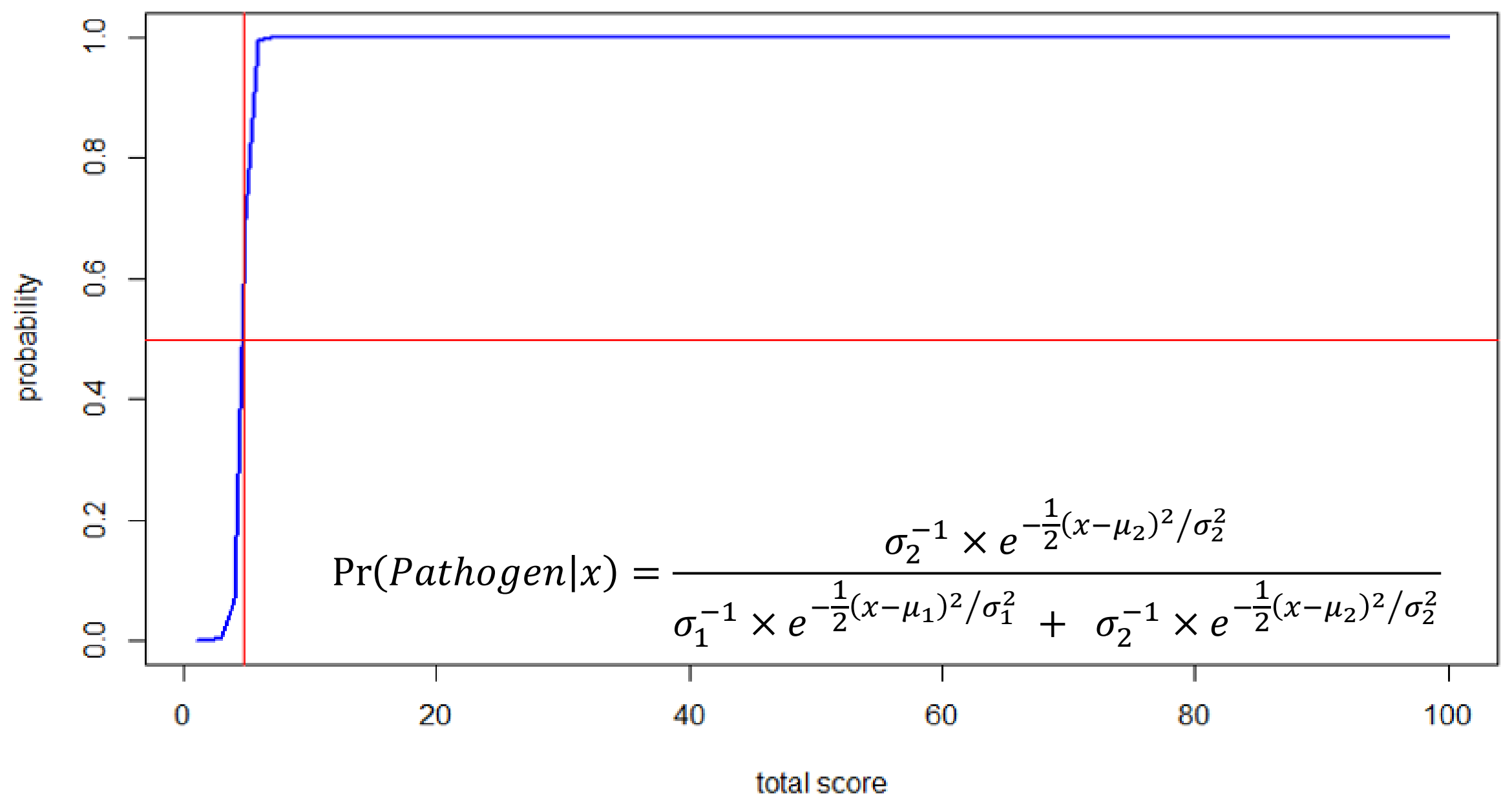

Figure 4. Plot the Probability of CiLV-C2 for Given Total Score p(total score): The probability is on Y-axis. The horizontal line satisfies p(total score) = 0.5. The LOD is the solution to Pr(Pathogen|x) = 0.5. A stringency of 80% probability barely changes the score.

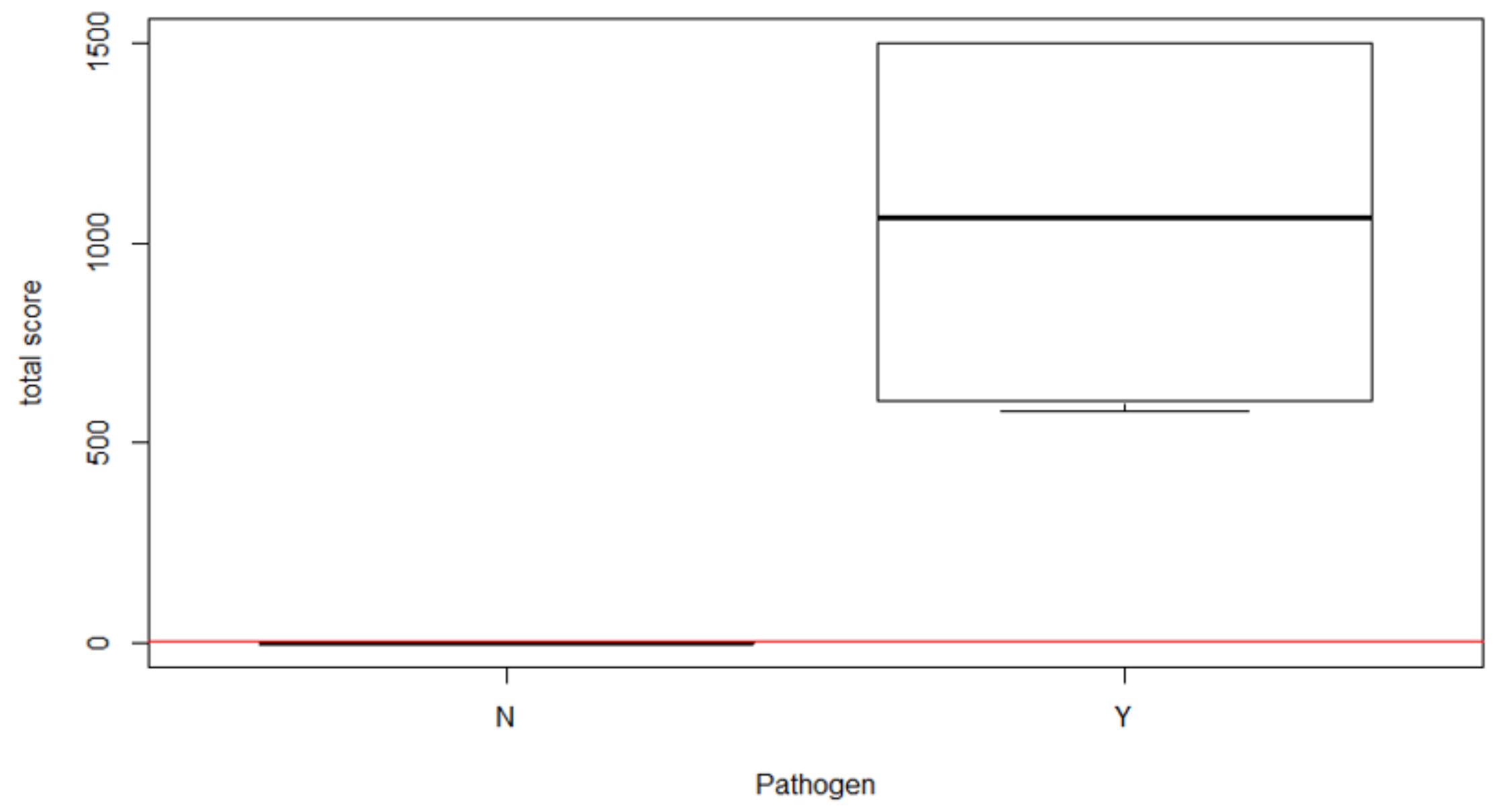

Figure 5. Boxplot of CEVd: We have two different normal distributions with different variances. This suggests that a quadratic discriminant analysis is reasonable. The LOD of score/10000 is 5.1 (red line).

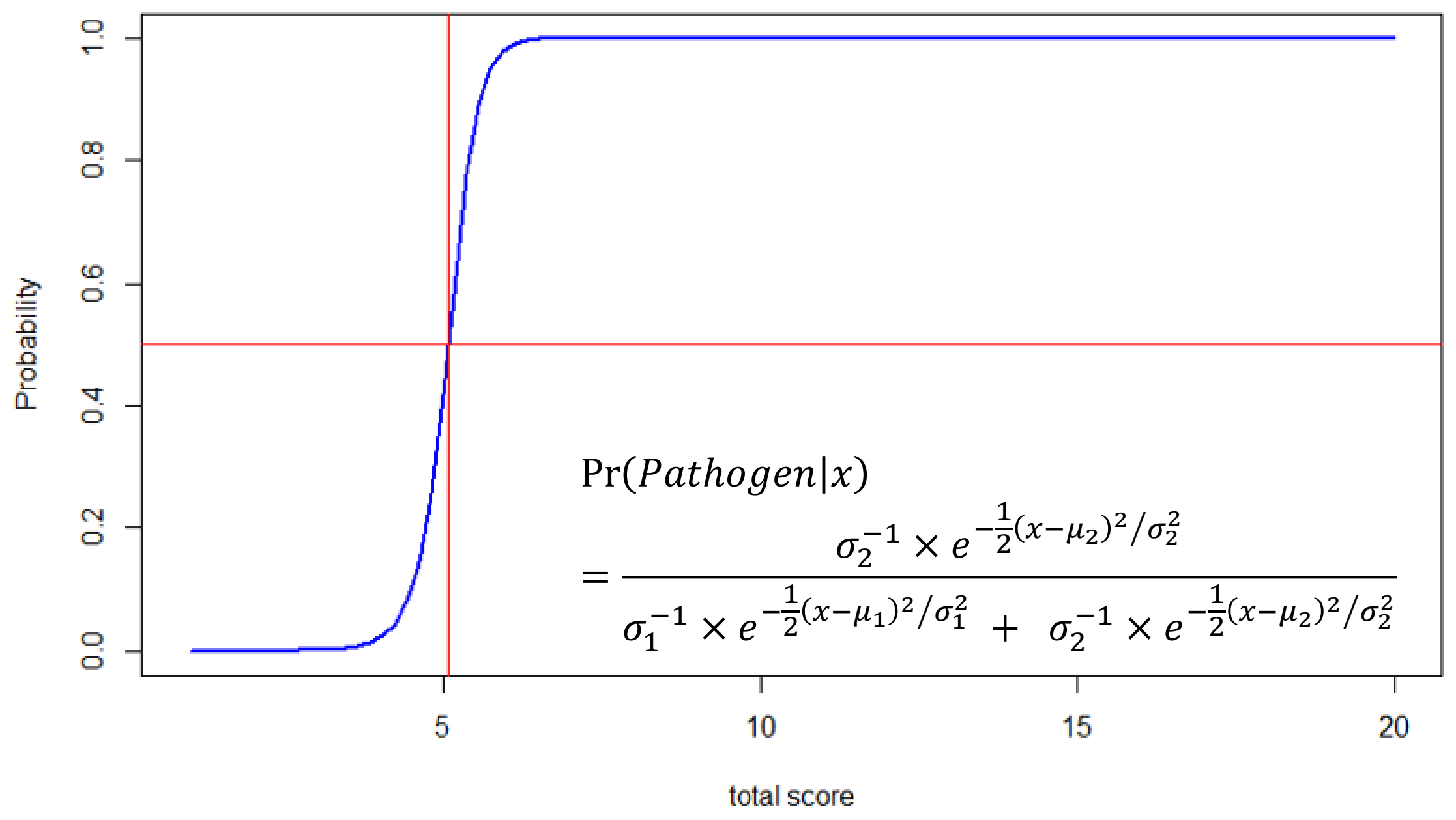

Figure 6. Plot the Probability of CEVd for Given Total Score p(total score): The probability is on Y-axis. The horizontal line satisfies p(total score) = 0.5. The LOD is the solution to Pr(Pathogen|x) = 0.5. A stringency of 80% probability barely changes the score.

Conclusions

- The LODs of citrus leprosis virus C2, citrus tristeza virus, and citrus exocortis viroid were 4.7, 4.2, and 5.1 scores/10000 respectively, indicating when the chance of positive is 50/50.
- The results are also consistent with and more sensitive than the PCR results.
- More known positive and negative samples will make these Bayesian models more robust.

Literature

1. Cardwell, Kitty. et al. Principles of Diagnostic Assay Validation for Plant Pathogens: A Basic Review of Concepts. Plant Health Progress 19, 272-278 (2018).
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
3. Stobbe, Anthony H., et al. "E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics." Journal of microbiological methods 94.3 (2013): 356-366.
4. Espindola, Andres, et al. "A new approach for detecting fungal and oomycete plant pathogens in next generation sequencing metagenome data utilizing electronic probes." International journal of data mining and bioinformatics 12.2 (2015): 115-128.