Testing and Characterizing the Effectiveness of Cyber Event Detection Capabilities: A Diagnostic Approach

17 March 2016

Matthew Dinmore, PhD
JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Heidi Jensen
SURVICE ENGINEERING COMPANY

Donna Rickter
SURVICE ENGINEERING COMPANY

Agenda

- Introduction, Motivation & Background
- Analytical Approach
- Design
- Observations & Discussion

Introduction, Motivation & Background

Cyber Defense - DCO

Defensive Cyberspace Operations (DCO)

DCO are Cyber Operations intended to defend DOD or other friendly cyberspace. Specifically, they are passive and active cyberspace defense operations to preserve the ability to utilize friendly cyberspace capabilities and protect data, networks, net-centric capabilities, and other designated systems.

DCO also includes actively hunting for advanced internal threats that evade routine security measures.

JP 3-12, p. II-2

Cyber Defense - Detection

Event Detection and Characterization

Activities in cyberspace by a sophisticated adversary may be difficult to detect. Unlike adversary actions in the physical domains which may be detected by the presence of equipment or specific activity, adversary actions in cyberspace *may not be easily distinguishable from legitimate activity*. Capabilities for *detecting* and attributing activities in cyberspace are critical for enabling effective DCO and OCO.

JP 3-12, p. II-8

Detection Capabilities

- Consist of *People*, *Process*, *Technology* components:
  - *People*: operators, analysts, users
  - *Processes*: cyber defensive TTP
  - *Technologies*: cyber tools

Objective: Develop & *characterize* capabilities for detecting cyber adversaries as distinguished from legitimate users on the DoDIN

Characterization vs Test

- Characterization: focuses on capability performance with weak or absent criteria
- Test: focuses on capability performance against established criteria

Assertion: Characterization is more appropriate in most cases at this time

Analytical Approach

Binomial Approach

- Measure successful detection rate
  - Success/fail for each trial, characterize with CIs

	Attack Observables
Detected	25
Failed to Detect	15

Wilson Score CI:

$$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{p}+\frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)+\frac{1}{4n^2}z^2}\right]$$

Detection rate: 63% (95% CI: 47%, 76%)

- Can also measure time to detect, etc.
- Need to consider possibility of false positives
- Can't be treated independently

Machine Learning: Precision/Recall

- Often used for assessing intrusion detection systems, cyber ML models

	Attacker Present	Attacker Not Present
Detected	25 (TP)	5 (FP)
Failed to Detect	15 (FN)	25 (TN)

TP = True Positive FP = False Positive
FN = False Negative TN = True Negative

Precision:

$$\frac{TP}{TP+FP} = \frac{25}{25+5} = 0.83$$

Recall:

$$\frac{TP}{TP+FN} = \frac{25}{25+15} = 0.63$$

- Approach ignores handling of false cases
- Is sensitive to *prevalence*
- Authorized >>> unauthorized observables

Medical Diagnostic Testing

- Answers the question, “How effective is a test for detecting a disease (or ruling it out)?”
- The disease may be relatively rare in the population

	Disease Present	Disease Not Present
Detected	True Positive	False Positive
Failed to Detect	False Negative	True Negative

- Focus on *sensitivity* and *specificity*
  - *Sensitivity (S_n) (True Positive Rate)* – test is positive over all results where disease is present
  - *Specificity (S_p) (True Negative Rate)* – test is negative over all results where disease is not present

Test Data Diagnostic
Is Your Data Usable?

- Chi-Square or Fisher's Exact Test
  - "Is there a relationship between the groups in the columns and the rows?"
- Use as a diagnostic for data validity

	Attacker Present	Attacker Not Present
Detected	25	5
Failed to Detect	15	25

Fisher's Exact Test Result

- Significance level - α = .10
- Result: *p*-value = 0.000196

☑ Since the *p*-value < α, Fisher's suggests that there is a strong relationship between the variables, and that the arrangement of test results is not independent.

Sensitivity and Specificity

- Sn & Sp are related

	Attacker Present	Attacker Not Present
Detected	25	5
Failed to Detect	15	25

- **Sensitivity (S_n) = $\frac{TP}{TP+FN} = \frac{25}{25+15} = 0.63$**
  - Note: same as Recall
- **Specificity (S_p) = $\frac{TN}{FP+TN} = \frac{25}{5+25} = 0.83$**
- Sample size: $N_{AP} = z \times \frac{s_n - (1 - s_n)}{W}$

z = Z-score for selected significance

Sn = target sensitivity

W = CI width

Characterizing Effectiveness

Wilson Score CI:

$$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{p}+\frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\hat{p}\left(1-\hat{p}\right)+\frac{1}{4n^2}z^2}\right]$$

	Attacker Present	Attacker Not Present
Detected	25	5
Failed to Detect	15	25

- Sn: 63% (90% CI: 50%, 74%)
- Sp: 83% (90% CI: 70%, 92%)

Characterizing Effectiveness

- Characterize with a single measure: ***Diagnostic Odds Ratio (DOR)***

$$\text{DOR} = \frac{(\text{sensitivity}) * (\text{specificity})}{((1-\text{sensitivity}) * (1-\text{specificity}))}$$

$$= \frac{(0.625) * (0.833)}{((0.375) * (0.167))} = 8.333$$

** 8-time increase in the odds of detecting an attacker on your network if there is an attacker*

- CI for the log of the DOR:

$$\log(DOR) \pm Z' \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}}$$

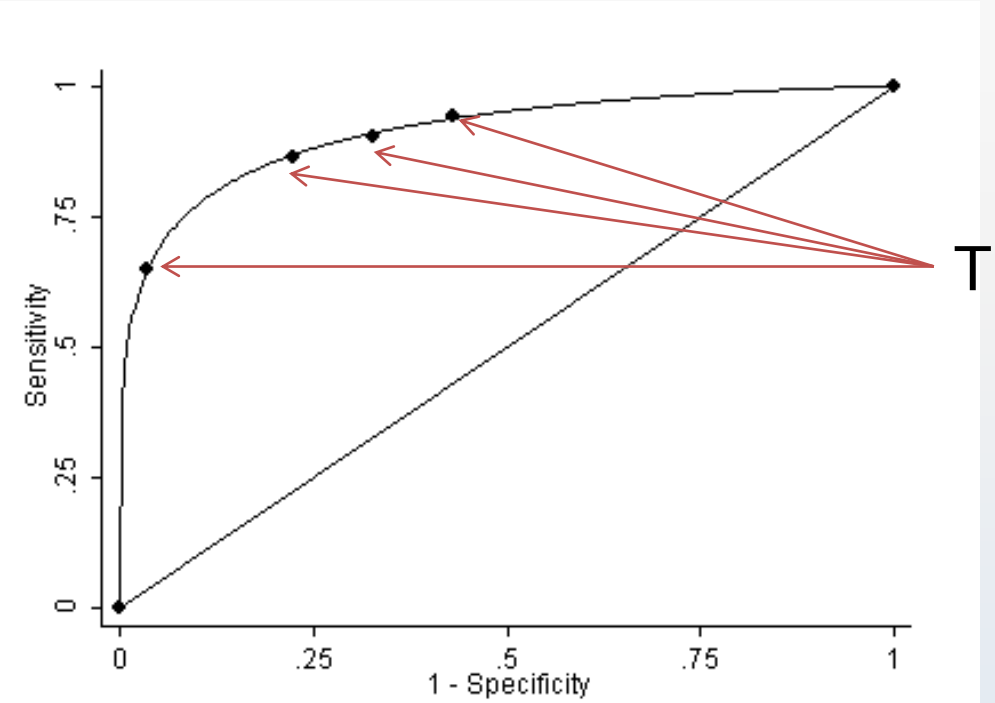
- 90% CI (DOR = 8.33): (2.78, 26.48)

Characterizing Effectiveness

- Diagnostic Odds Ratio (DOR):
  - Prevalence independent
  - Allows independent tests of effectiveness to be compared
  - Allows similar tests to be used for meta-analysis
  - Provides an explainable measure of discrimination
- Cons:
  - Obscures balance between Sn & Sp

Characterizing Effectiveness

- ROC Curve and Detection Threshold



Threshold points

- Operational usefulness of detection thresholds
- Area under the curve (AUC), discrimination: ability to separate malicious from benign activity

Test Design Considerations

Test Design Considerations

- Between vs within subjects
- Participants
- Choosing thresholds/objectives

- Red team injects & non-malicious failures

Test Trial Protocol

- Between subjects:
  - Need to randomize for experience/skill

- Within subjects:
  - Need to counterbalance

- Both
  - Watch for learning effect during test
  - Include plenty of hands-on time: a key element is understanding what's normal on a network

Participants

- Skills/experience assessment
- Service pipelines predominantly producing cyber ninjas vs teams
- Beware certifications, self-reported experience

Thresholds/Objectives

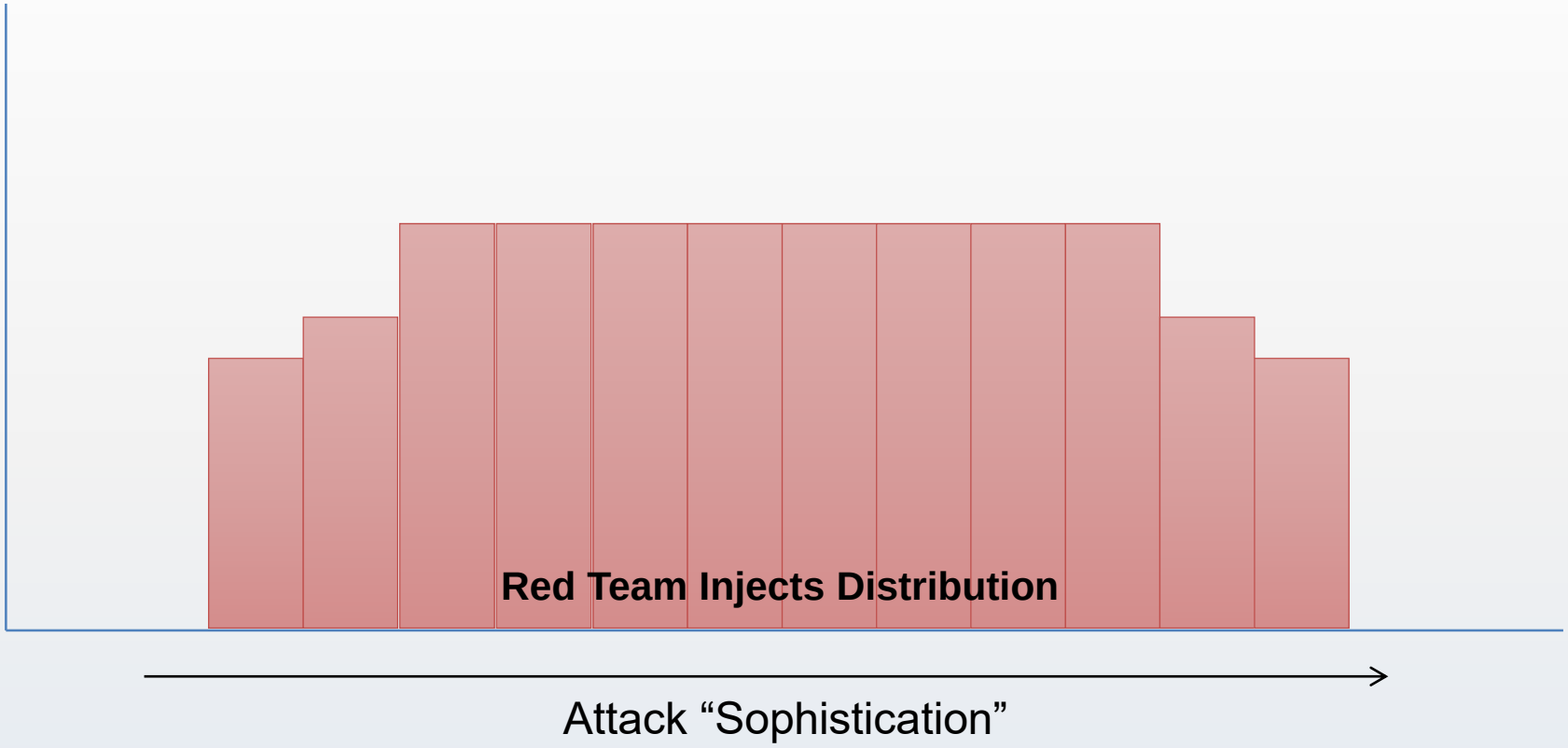
- Requirements (T/O)

- If you must choose… consider:
  - Negative diagnostic tests
  - DOR & Sp/Sn imbalance
  - Specify at cutoffs (if applicable)

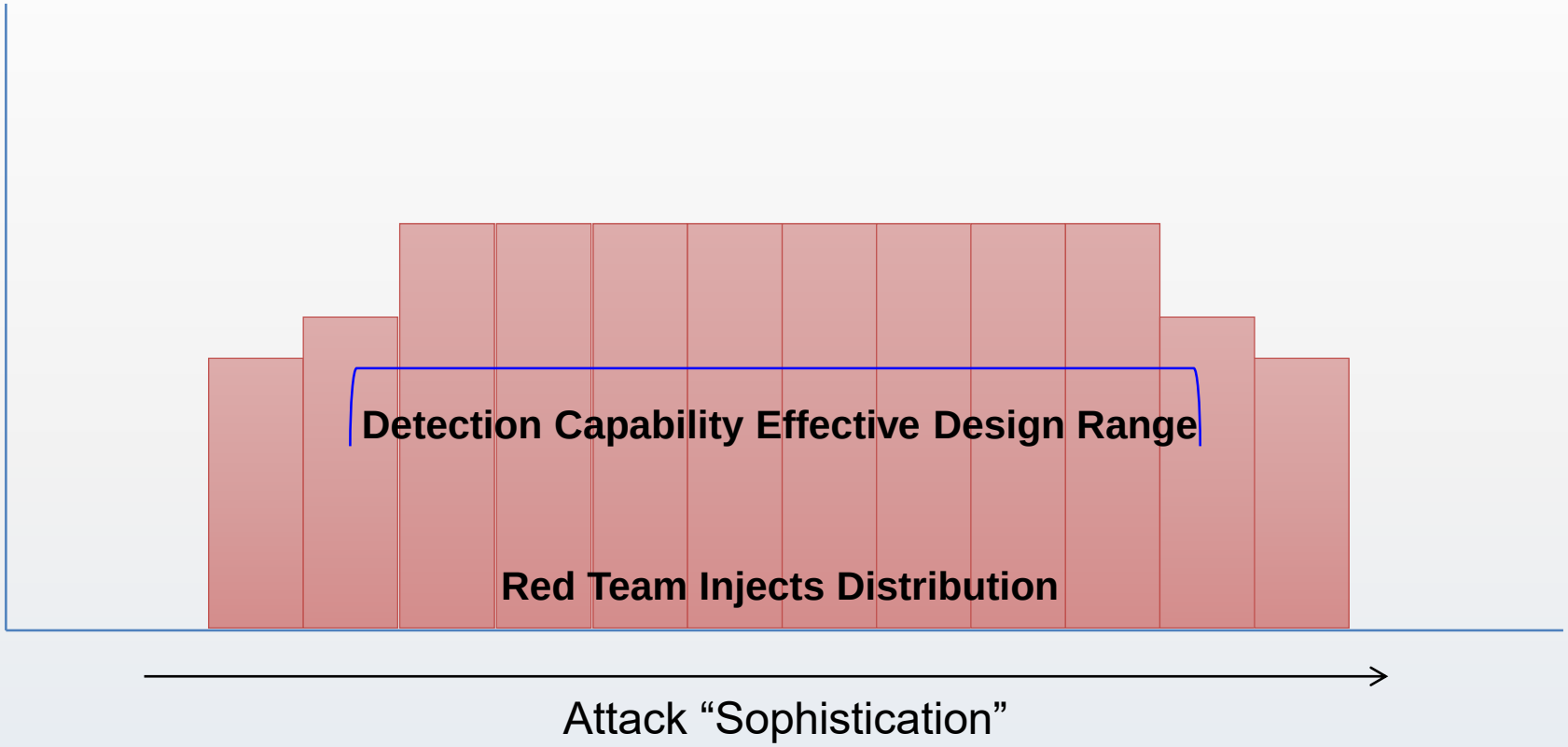
Red Team Injects

Attack “Sophistication”

Red Team Injects



Red Team Injects



Red Team Injects

- May need to create conditions for false positive detection
  - “Normal” may be too normal
  - Create failures that
    - Are somewhat common
    - Look like corresponding red team inject effects
  - *Critical* that white cell has/follows script

Discussion

Plain-Language Results

- Commanders/operators/analysts need to understand the results
- Detection tool X is 80-90% successful at detecting {range of attacks} with a confidence of 95%
- Detection tool X improves the odds of distinguishing a cyber adversary from a routine failure by 10x (the DOR)

Guidelines: STARD Process

- Standards for Reporting of Diagnostic Accuracy
- Provides:
  - Reporting checklist
  - Process
  - Examples

We need to move toward similar guidelines for evaluating cyber capabilities

Conclusions

- Need methods that
  - Assess cyber detection capabilities as a whole
  - Consider prevalence, threat, etc.
  - Enable results to be compared
  - Provide users an operationally-useful understanding of their effectiveness

Abbreviated References

1. Bossuyt, et. al. (2003). “The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration.” *Croation Medical Journal*. 44(5):639-650.
2. Chairman of the JCS. (2013). “Cyberspace Operations.” *US Department of Defense*, Joint Publication 3-12(R).
3. Glas, et. al. (2003). “The diagnostic odds ratio: a single indicator of test performance.” *Journal of Clinical Epidemiology*. 56: 1129-1135.

Questions?

Matthew Dinmore, PhD
Johns Hopkins University Applied Physics Laboratory

matthew.dinmore@jhuapl.edu

Heidi Jensen
Survice Engineering Company

heidi.jensen@survice.com

Donna Rickter
Survice Engineering Company

donna.rickter@survice.com