WHO Vaccination Coverage Survey Briefing

Responding to Questions re: Steps 1-3

World Health Organization

Istanbul, December 2015

Questions from Tuesday

1. *Could you please describe the relationship between estimation and classification again?*

2. *Lot Quality Assurance Sampling (LQAS) is a method for rapid inexpensive survey to classify coverage – why aren't we listing that as an option here?*

3. *The case study assumes a design effect of 4. That seems too high. Do we see values that high in coverage surveys?*

Questions from Tuesday

- *Your slide on differences said: "The increase in coverage is estimated to be 4.0% [95% CI -0.1%-8.1%]. … indicating marginally strong evidence that Penta3 coverage is different…"*

  But if the CI for the difference includes zero, why are you concluding there is likely a difference?!?

Question 1.
Could you please describe the relationship between estimation and classification again?

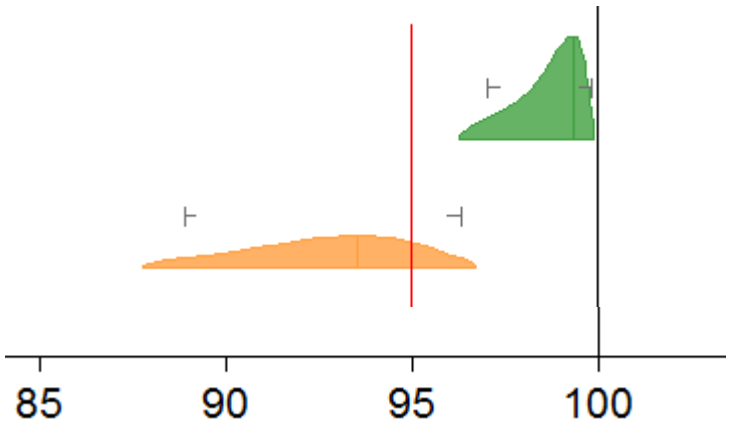
Estimate and Classify

- The 2015 manual (like the 2005 manual) recommends using survey results to estimate coverage
- The 2015 manual (like the 2005 manual) recommends calculating a point estimate and a 2-sided 95% confidence interval (CI)
- The methods in the 2015 manual are an improvement
  - Point estimate no longer assumes equal weights
  - Confidence interval method is designed for proportions
  - When sample sizes are small and coverage is near 0% or 100%, the CIs will not be symmetric

Estimate and Classify

- Unlike the 2005 manual, the 2015 update recommends also calculating 1-sided 95% upper and lower confidence bounds

- These bounds may be used to classify coverage



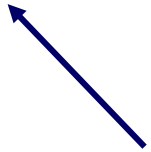
99.3 (96.3,99.9) [97.0,100) (0,99.9]

93.5 (87.8,96.7) [88.9,100) (0,96.3]

World Health Organization

Classification Conclusion

- **IF** we believe the survey is free of important biases:



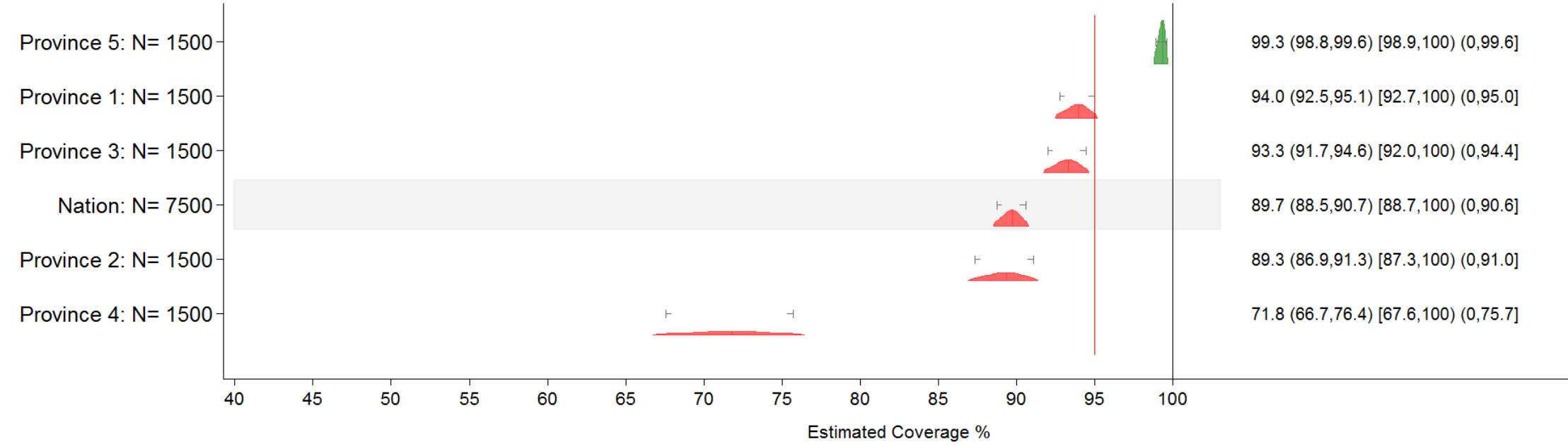
This is a BIG "If"

- Then we can say:
  - *"We are 95% confident that coverage is ≥ LCB."*
  - *"We are 95% confident that coverage is ≤ UCB."*

When Does Classification Make Sense?

- In a survey with nested strata (e.g., provinces within a nation)
- We have to decide whether to do a large survey in every province or do a smaller survey in each province… accepting wider confidence intervals there…knowing that we will combine all the data to obtain narrow CI at the national level

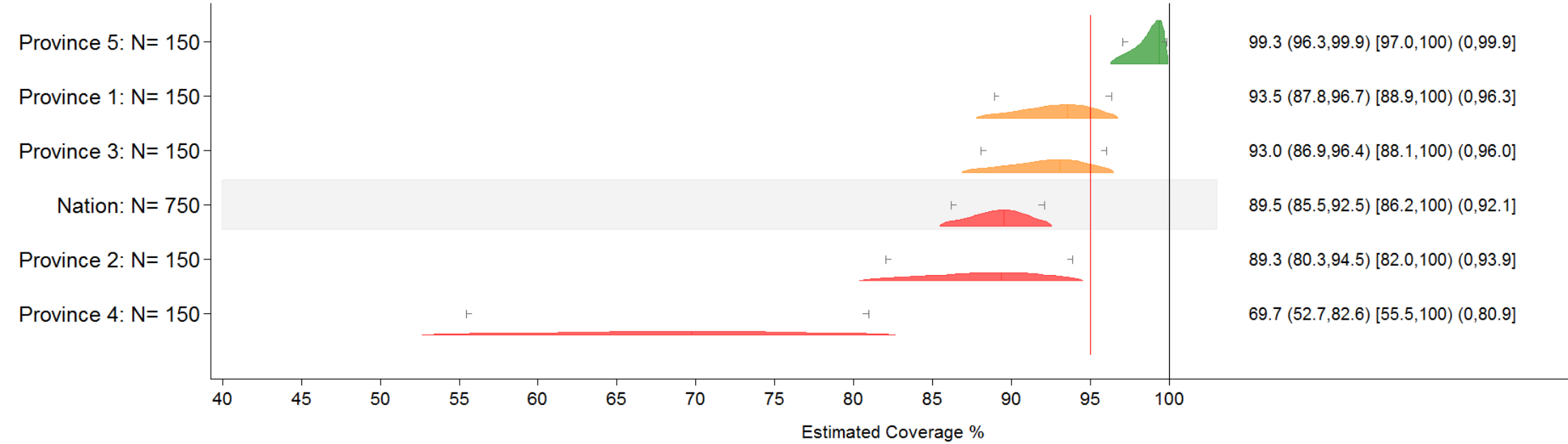
Big Survey
N=1,500 per Province



Text: Point Estimate (95% CI) [1-sided 95% LCB, 100) (0, 1-sided 95% UCB]

World Health Organization

Smaller Survey
N=150 per Province



Text: Point Estimate (95% CI) [1-sided 95% LCB, 100) (0, 1-sided 95% UCB]

World Health Organization

Recommendations

- Always report estimation results (point estimate & CI)
- Always report the measures you took to keep bias out of the survey
- Always report places where bias may have crept in to the survey
- If you want to classify and you want to be very likely to "pass" strata with coverage > some upper threshold and "fail" strata with coverage < some lower threshold, the annexes will help you pick a sample size to do that

Question 2.
Lot Quality Assurance Sampling (LQAS) is a method for rapid inexpensive survey to classify coverage – why aren't we listing that as an option here?

Why Not LQAS?

- LQAS uses a quota sample
  - Substitutes HH if no one at home
  - Keeps no record of how many substitutions
  - Probably biases coverage upward
- LQAS gives one decision rule, tuned for a pair of thresholds; our method can be used to classify against any threshold without modification
- Clustered LQAS has an assumed design effect built into the decision rule; our method uses the observed DEFF

Why not LQAS?

- Our method encourages graphic display of what we learned from the survey, i.e., how our confidence is distributed; LQAS is a black box…er, ball:

Question 3.
The case study assumes a design effect of 4. That seems too high. Do we see values that high in coverage surveys?

Q: Is DEFF = 4 realistic? If yes, why?

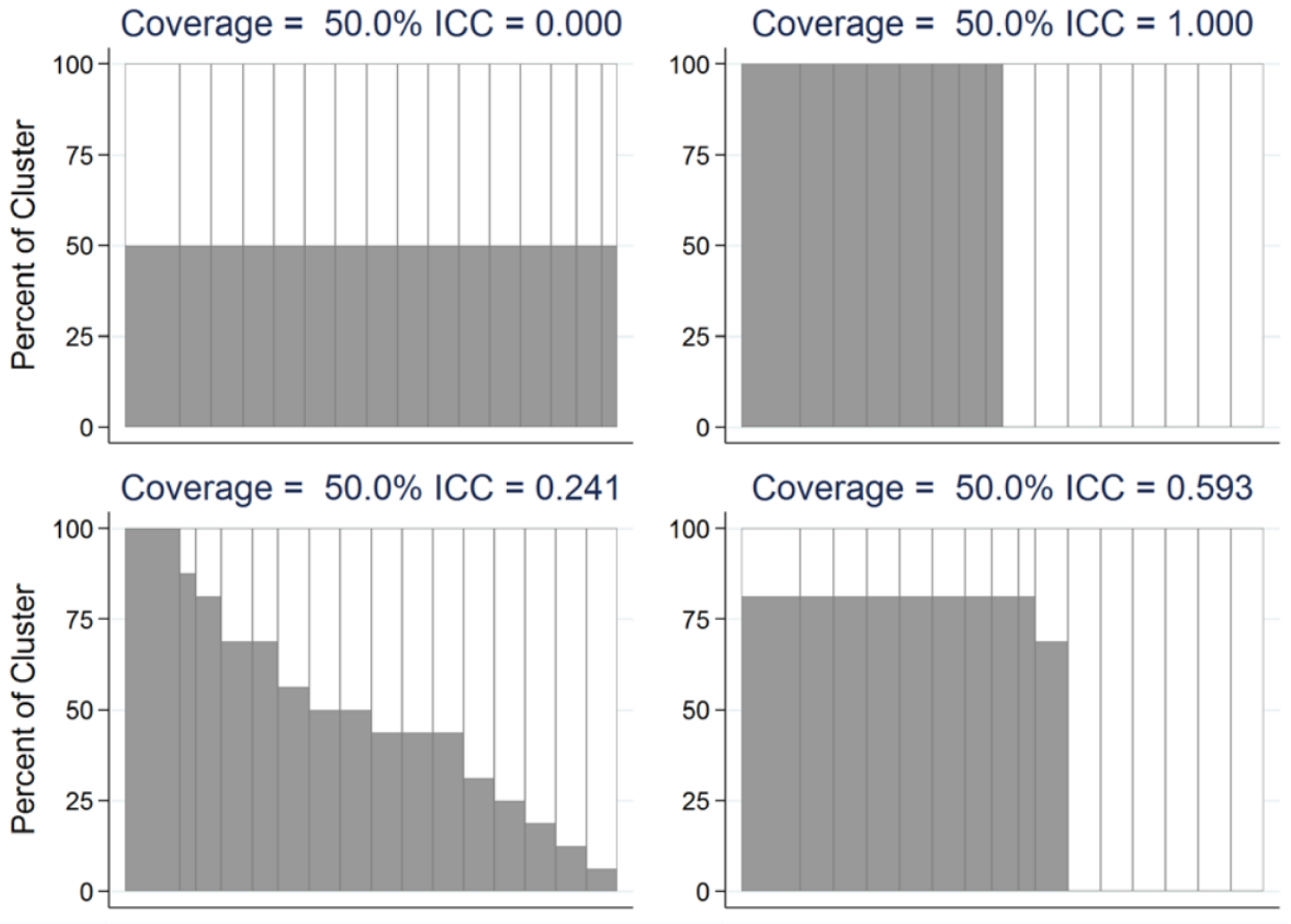
A: Yes, very realistic:

- 2012 Ethiopia EPI survey:
  - 31 / 182 (17%) coverage DEFFs were ≥ 4.0
  - (11 regions + national x 13 doses + fully vaccinated = 182 results)
- 2014 Kano, Nigeria EPI survey
  - many of the 585 coverage DEFFs were ≥ 4.0
  - (I didn't take time to count, but it looked like more than 10% of them)
- Why?!?

Reason 1: There are 2 “Design Effects”

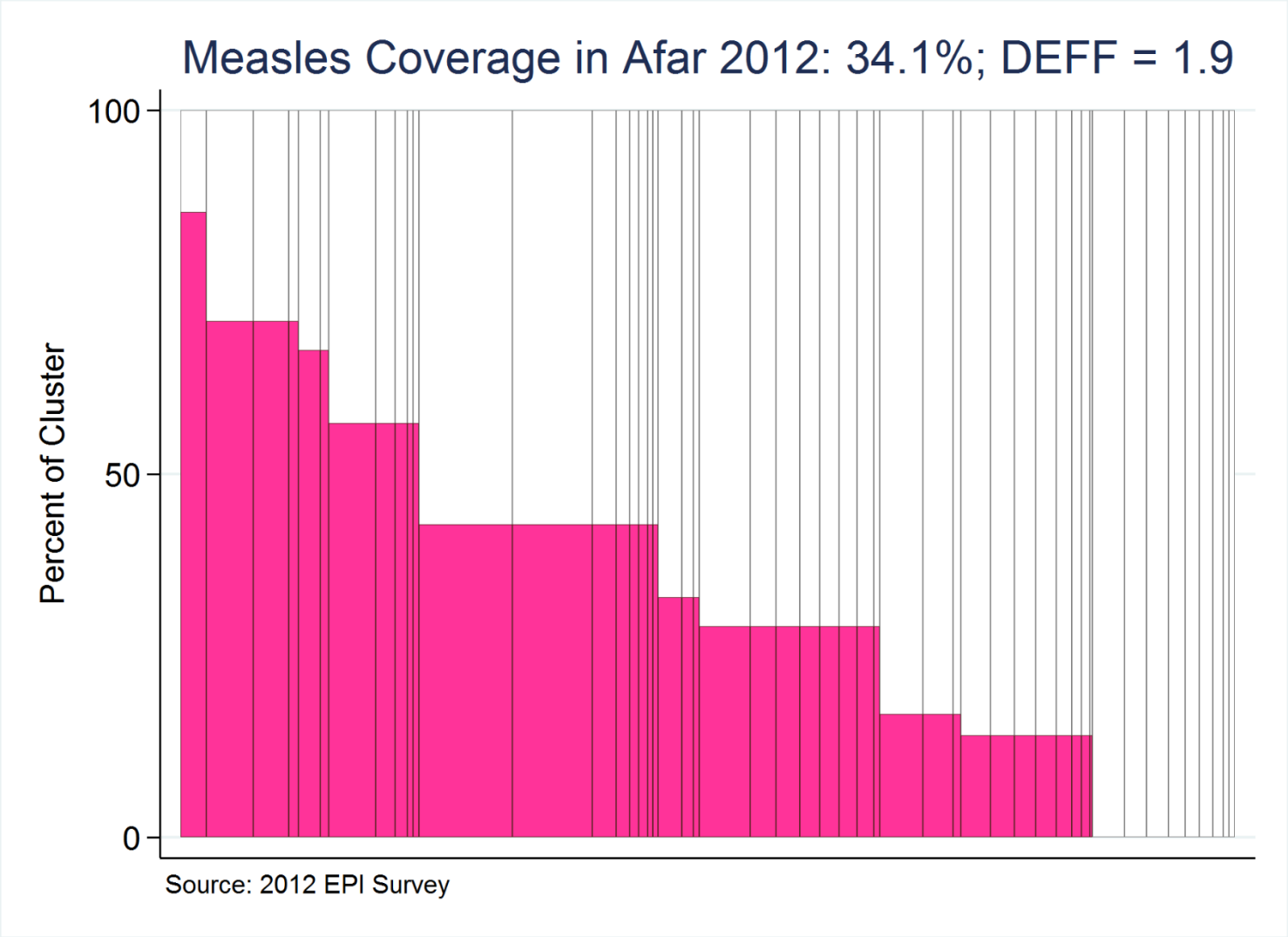
- Recall that $\mathrm{DEFF} \cong 1 + (m - 1)\rho$
  - where m is avg N / cluster
  - ρ is the intracluster correlation coefficient
- $\mathrm{DEFT} = \sqrt{DEFF}$ (This is what DHS reports.)
- Both DEFF and DEFT are called “the design effect”
- The WHO reference manual uses DEFF
- Maybe the audience members were thinking of DEFT ???

Reason 2: DEFF is high when coverage is spatially correlated



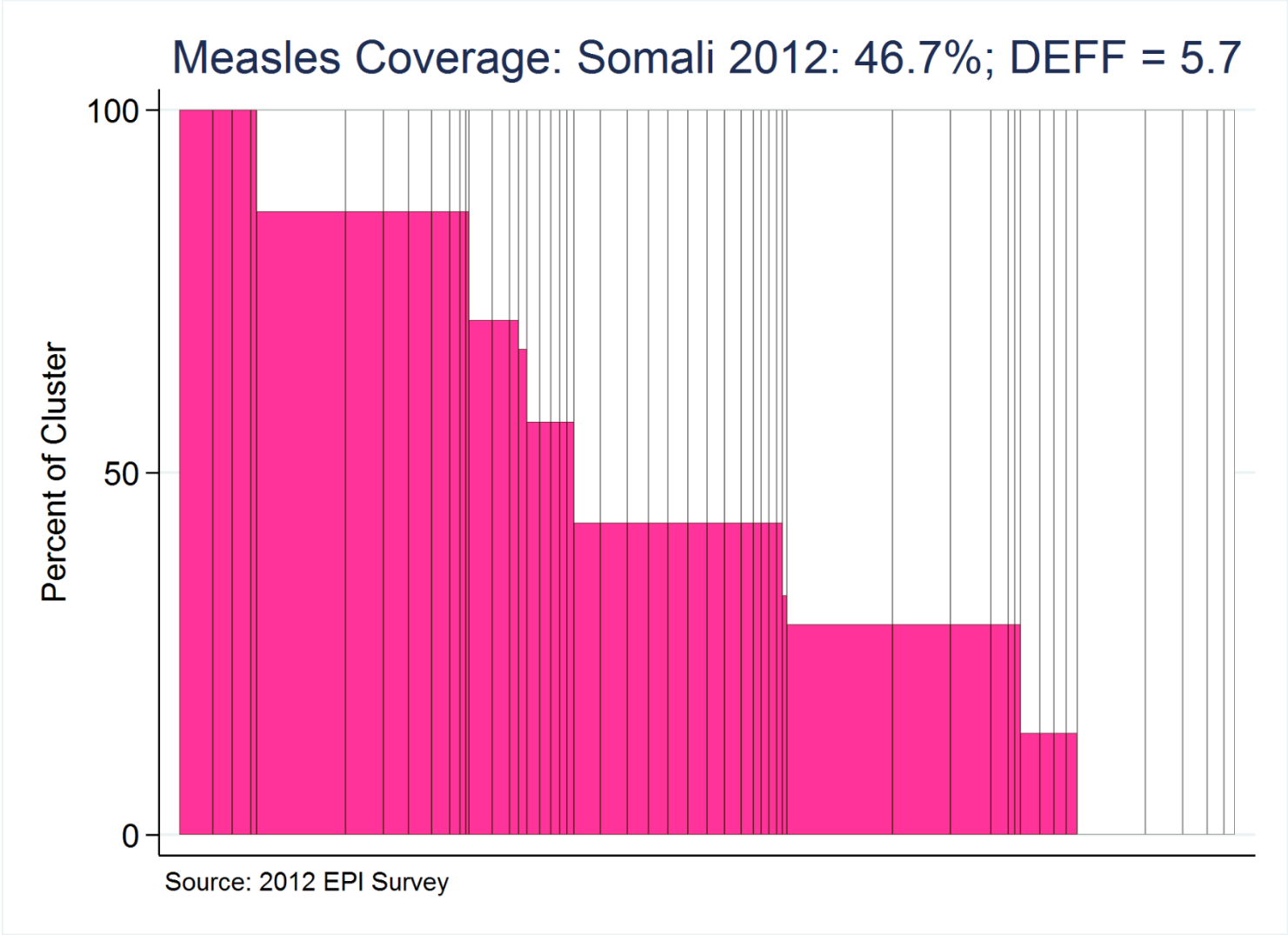
World Health Organization

Real Data with DEFF = 1.9



World Health Organization

Real Data with DEFF = 5.7



World Health Organization

Question 4.
Your slide on differences said: "The increase in coverage is estimated to be 4.0% [95% CI -0.1%-8.1%]. … indicating marginally strong evidence that Penta3 coverage is different…"

But if the CI for the difference includes zero, why are you concluding there is likely a difference?!?

How Should We Report Differences?

- First, please read section 6.4.6 on reasons why it may be a bad idea to compare coverage estimates from two surveys using a formal hypothesis test.

- Sections 6.4.7 describes what to report:
  - Estimated coverage in two groups (or surveys)
  - 95% CI for coverage in each
  - Estimated difference & 95% CI
  - Indicate that the CI for the difference is calculated using software that accounts properly for the complex sampling design
  - List the degrees of freedom available for the test
  - List the p-value and your conclusion in words

And if the CI includes zero?

- If the 95% CI includes zero then the p-value will be > 0.05 and we cannot conclude with 95% confidence that there is an underlying difference
- But the data may be suggestive of a difference…and it is fine to say that…
- In my example we can conclude with 94% confidence that there is an underlying difference…the test misses the magic p-value of 0.05 by less than 1%, so I chose to describe the results as showing “marginally strong evidence of a difference”

1-sided or 2-sided test?

- The manual recommends using a 2-sided test unless there is a strong programmatic reason to assume that coverage has increased (or decreased) over time
- If you can justify a strong reason for the 1-sided test, then state the reason, and state the p-value for that test
- I would also report the p-value for the 2-sided test
- If you want to report results of 1-sided tests, it is best to identify that plan before looking at the results (and to say so in the report)

How to describe results?

- If you report all the metrics suggested above then the reader can come to their own conclusion about how to label the difference (weak / moderate / strong evidence for a difference), so report the numbers and then report your interpretation (the Steering Committee's interpretation) of them in words

Questions?

- I'll be very happy to discuss any of these points further
- Talk to me here during the meeting, or send me a note

Dale.Rhoda@biostatglobal.com