How good need a new test be?

Graeme P Young
Outline

• What we want of a new test and concordance with guiding principles of the 2016 publication
• Specific areas:
  • Accuracy: sensitivity, specificity, test operating characteristics.
  • Scenarios for practical application
• What we do need to know
What we want of a new test

**Specified in 2016 recommendations**

1. It must be effective and at least as good as the best existing tests [1,3,4,7*].
2. The test must meet diagnostic laboratory requirements stand up to regulatory scrutiny [8*].
3. Testing process (sampling, logistics, risk) must be acceptable to the invitee [2*].
4. Confident that the lesions it detects are worth detecting (not just indolent) [5*].

**Proposed additional points of clarification**

1. The test endpoint (criterion values for positivity) should be flexible to allow adjustment to meet program goals as determined by policy and providers.
2. Accuracy (sensitivity and specificity) must be appropriate for the context of use.
3. Positivity rate in population must be feasible and it must be cost-effective.

* A balancing act with multiple dimensions
What accuracy is good enough?

1. **Detection** — determines mortality/incidence benefit.
   - a more sensitive test returns more **true-positives**.
2. **Burden associated with detection** — determines cost:
   - a more specific test returns fewer **false-positives**.

- So, we learn a lot by determining if a positive result is true or false
- Each is readily compared to a comparator test proved to be effective by RCT.
- In 2016 we concluded that:
  - “gFOBT are the *minimum* standard.” 29-47% sensitivity, 87-98% specificity.
  - “FIT are also acceptable” (clearly superior to gFOBT).
- We now need to make it clearer what this means!

* referred to as test operating characteristics  † cost in its broadest sense
Desirable values - sensitivity

<table>
<thead>
<tr>
<th>Result</th>
<th>Related accuracy characteristic</th>
<th>Desired achievement*: absolute and relative.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TRUE – Sensitivity (mortality and incidence impact).</td>
<td>Cancer: ≥ 70% (absolute). ≤ FIT (relative)</td>
</tr>
</tbody>
</table>

- FIT sensitivity for CRC (all stages; cutoff 100 ng Hb/mL): 75% (61-86%). (Imperiale et al., 2019).
- USPSTF suggests acceptable sensitivity to be at least 70%. (Bibbins-Domingo et al., 2016).

*once-off test
## Desirable values - sensitivity

<table>
<thead>
<tr>
<th>Result</th>
<th>Related accuracy characteristic</th>
<th>Desired achievement*</th>
<th>Advanced PICN†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TRUE – Sensitivity (mortality and incidence impact).</td>
<td>Cancer: ≥ 70% (absolute). ≥ FIT (relative)</td>
<td>≥ 25%; ≥ FIT ≥ 40%; ≥ Cologuard</td>
</tr>
</tbody>
</table>

- Desirable standards have not been set and it is unclear what proportion of advanced PICN will progress to cancer within a decade.
- A working standard might be a sensitivity for advanced colorectal neoplasia (ACN) of ≥ 40%.

*once-off test  †preinvasive colorectal neoplasm
Desirable values – specificity for advanced CRN

<table>
<thead>
<tr>
<th>Result</th>
<th>Related accuracy characteristic</th>
<th>Desired achievement*: absolute and <strong>relative.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>FALSE – Specificity. (utility, cost-effectiveness)</td>
<td>In range 92-98% in target population. Ideally ≥ 95%. ≥ FIT when set at equivalent sensitivity.</td>
</tr>
</tbody>
</table>

- FIT specificity for CRC (cutoff 100 ng Hb/mL): 95% (92-98%). (Imperiale et al., 2019).
- USPSTF suggest acceptable specificity to be at least 90%. (Bibbins-Domingo et al., 2016)

*once-off test
Resultant Scenarios

1. Superior true-positive rate (sensitivity)
   • Examine the false-positive rate and calculate colonoscopies per lesion detected (i.e. number needed to colonoscope, NNC) to judge efficiency.

2. Equivalent true-positive rate. Are there other benefits?
   • Better false-positive rate. (Specificity)
   • Lower costs,
   • Higher participation rate, etc
   • Use number NNC to make final decision. Prior consensus on what is required is crucial.

   • Noninferiority design becomes crucial.
Population positivity rate (PPR) as a driver

<table>
<thead>
<tr>
<th>Result</th>
<th>Related characteristic</th>
<th>Rate implication</th>
</tr>
</thead>
</table>
| Positive| TRUE + FALSE positives. Colonoscopy workload (feasibility) | • Match to resource capability.  
• At the chosen rate the new test must pose some accuracy, acceptability or cost advantage over the established test.  
• Rates > 5% are needed for detection of adv-PICN using FIT. |

1. Preliminary positivity criteria set in phase 1 based on likely sensitivity.  
2. PPR is then established in screening population.  
3. Criteria reset to give feasible PPR if necessary.  
4. Is the resultant sensitivity and specificity OK?  
5. Must also compare this new test with the comparator at equivalent PPR.
Why is a comparator test important?

• Neoplasms that bleed or that are detected by visualization (structural) means are worth detecting because early detection is proven to reduce mortality and/or incidence.
• Detection by some other method, e.g. genetic or molecular, might reflect a different biological behaviour where treatment might not be as effective, or, the lesion might be indolent.
• Showing that a new test can detect those we can already pick-up by established means provides confidence.
What we do need to know

Obligatory:

- It must be effective and at least as good as the best existing tests.
- The test must meet diagnostic laboratory and regulatory requirements.
- Testing process (sampling, logistics, risk) must be acceptable to the invitee.
- Any additional neoplasms detected must be worth detecting

Proposed clarifications and specifications:

- Accuracy claims (in intended-use population): Sensitivity >70-75% (for cancer) and specificity (for ACN) >90-95%.
- The test endpoint must be flexible to allow adjustment of operating characteristics so as to meet the goals of a screening program.
- If colonoscopy workload (PPN) drives program goals, choice of a flexible test’s endpoint that delivers such must also deliver acceptable accuracy.
SAVE THE DATE!