19TH COCHRANE COLLOQUIUM
VI INTERNATIONAL CONFERENCE ON PATIENT SAFETY
• X Annual Meeting of the Iberoamerican Cochrane Network • VIII Annual Meeting of the Iberoamerican Network-GPC
• XII LatinCLEN Congress (the Latin American branch of INCLEN-International Clinical Epidemiology Network)
Scientific evidence for healthcare quality and patient safety
19-22 OCTOBER 2011 MADRID

Investigating determinants of experts' judgement on appropriateness for diagnostic tests when using the "Grading of Recommendations Assessment, Development and Evaluation (GRADE)" method for presenting evidence

Luciana Ballini, Fabio Trimaglio, Susanna Maltoni, Luca Vignatelli, Antonella Negro, Roberto Grilli - Agenzia sanitaria e sociale regionale, Emilia-Romagna - Italy

Background

Within a research program on the development of criteria of appropriate use of FDG-PET in oncology, carried out by the regional Health Agency of Emilia-Romagna in Italy, we adopted:
- a "consequentialist" approach to test's evaluation (Bossuyt 2010)
- logic and principles of the **GRADE** approach, developed by the GRADE Group for diagnostic tests (Schunemann 2008) to manage absence of evidence on clinical outcomes.

Two multidisciplinary panes of 39 regional experts were convened to agree on criteria of appropriate use of FDG-PET.
An analytic framework (Figure 1 and 2) was provided to aid process of consensus and resolution of disagreement.
The RAND/UCLA Method of Appropriateness voting procedure was used to formally register agreement on criteria of appropriateness.

Definition of appropriateness :
- an initial diagnosis and therapeutic approach following the initial diagnosis;
- the capacity of the new test to modify the initial diagnosis;
- the subsequent change in the therapeutic approach;
- the clinical benefit expected from the change in the therapeutic approach endorsed by test results.

Analytic framework

A modifiable diagnosis and subsequently modifiable therapeutic approach is the rationale which is affected by
1. expected change in management
2. effectiveness of the therapeutic intervention
3. importance of clinical outcomes.

Level of evidence of required estimates of diagnostic accuracy depending on test's position are put in relation with trade-off between harms and benefits.

Figure 1. The analytic framework

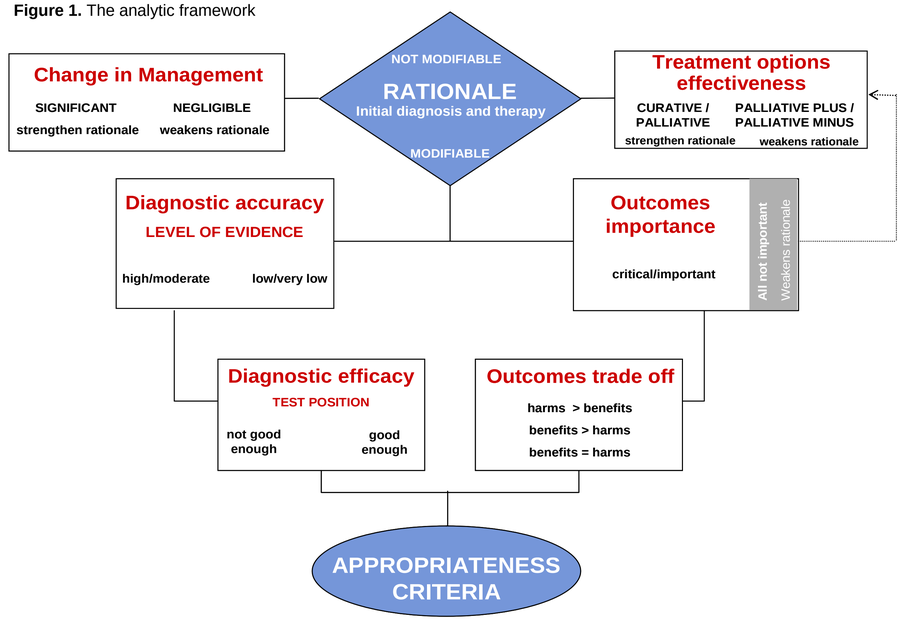
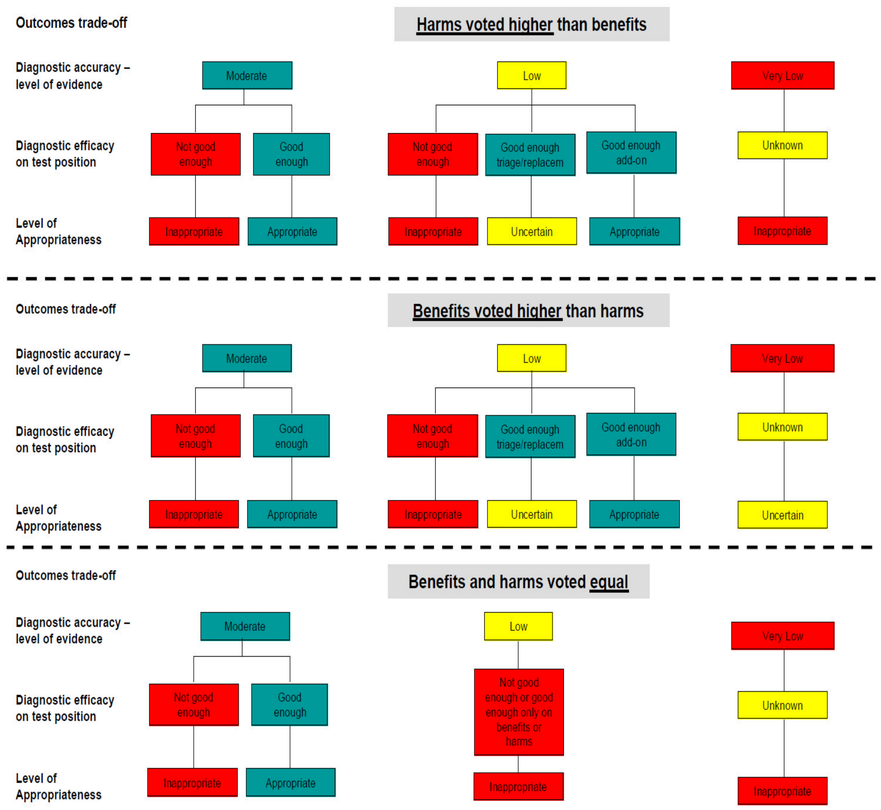


Figure 2. Decision-tree for trade-off between harms and benefits



Objectives

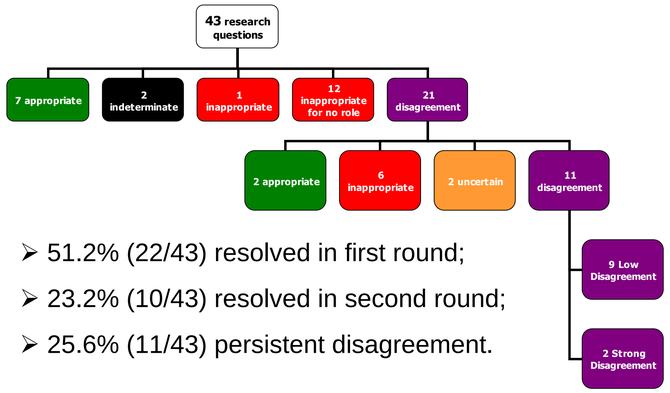
- to establish whether the panels fully applied the proposed analytic framework when expressing judgement on appropriateness;
- to investigate discrepancies;
- To investigate possible sources of persistent disagreement among panelists

Methods

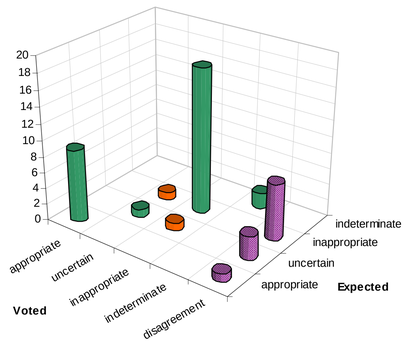
1. the application of the analytic framework was assessed by comparing expected and observed votes on appropriateness;
2. descriptive statistics of results were analysed to identify possible patterns of discrepancy
3. a focus on disagreement among panelists was carried out to see whether persisting disagreement was due to:
   a. lack of shift in opinion, detected by individuals' voting behaviour in first and second voting round;
   b. possible "corporative behavior";
   c. possible influence of new comers (panelist present only at second voting round);
   d. prevalence of strong disagreement in first voting round.

Results

Overall results of the process



- 51.2% (22/43) resolved in first round;
- 23.2% (10/43) resolved in second round;
- 25.6% (11/43) persistent disagreement.



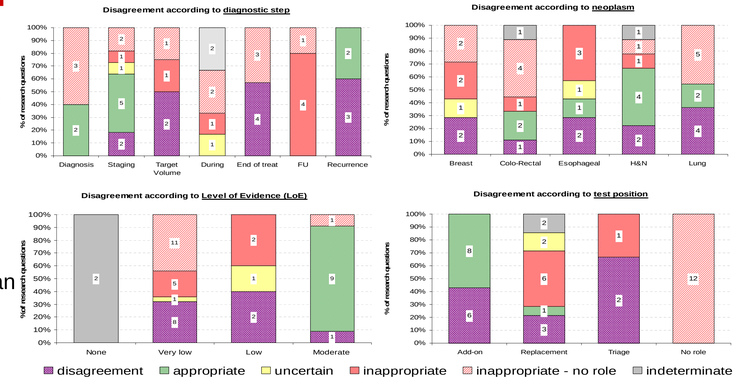
Expected level of appropriateness	Observed: appropriate	uncertain	inappropriate	indeterminate	disagreement	Total
appropriate	9				1	10
uncertain		1	1		3	5
inappropriate		1	18		7	26
indeterminate				2		2
Total	9	2	19	2	11	43

1. Did the analytic framework "work" for our experts?

Yes: Comparison between expected and observed appropriateness shows that in ≈70% of cases observed appropriateness coincides with expected appropriateness.

2. When didn't the analytic framework "work"?

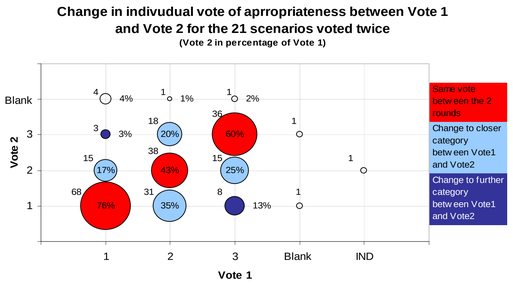
Except for two scenarios for whom there was an agreement on a results different from the expected one, the most part of the divergence between observed and expected is due to disagreement.

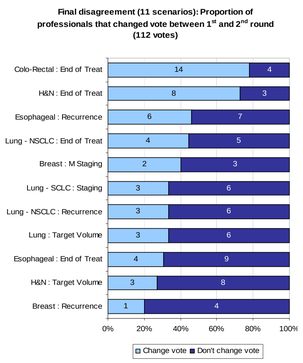


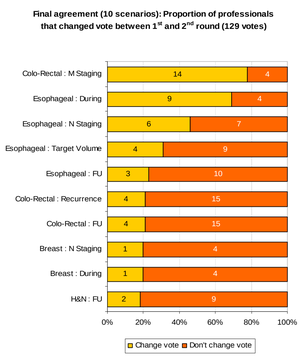
We looked at the distribution of disagreement with respect to neoplasm, diagnostic role of the test, position of the test and level of evidence, but we found no interesting associations.

3a. Was disagreement due to lack in shift of opinion?

No:
→ 47.6% (10/21) of disagreement was resolved at second round;
→ change in opinion at 2° round took place in 24%, 40% and 67% of panelist that had voted "appropriate", "inappropriate" and "uncertain", respectively
→ final disagreement and resolved disagreement show similar distribution of opinion shift "intensity".

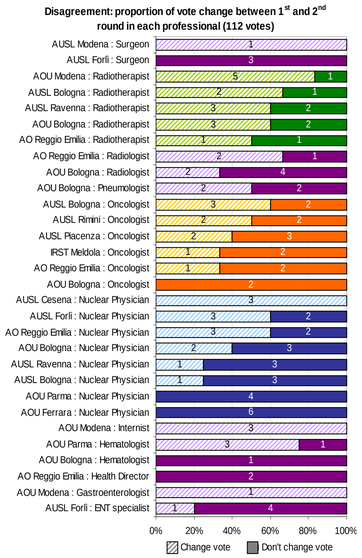


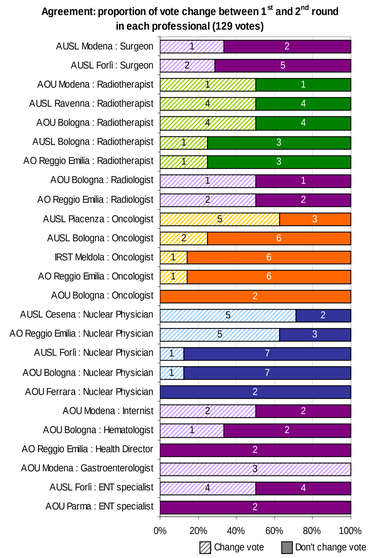




3b. Was disagreement due to possible corporative behavior?

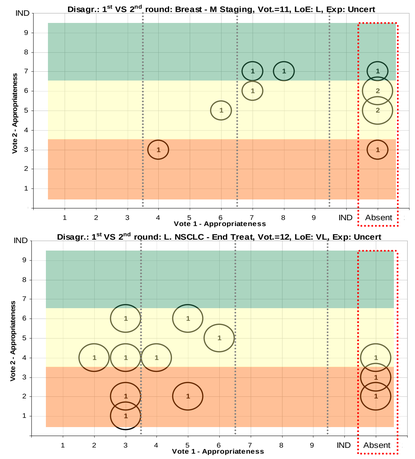
No: we didn't identify any particular corporative behavior.





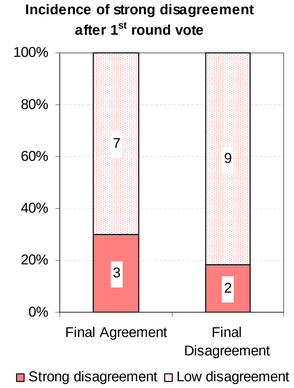
3c. Was disagreement determined by new comers (panelist present only at 2° round vote)?

No: new comers have never determined final disagreement (two examples below).



3d. Was final disagreement due to difference in the strength of disagreement after 1° round?

No: incidence of strong disagreement (as defined in RAND/UCLA methodology) is similar between resolved and unresolved scenarios.



Discussion and conclusion

The **analytic framework** used in the regional Health Agency of Emilia-Romagna research program was very complex. It entailed a multi-dimension definition of appropriateness of a diagnostic test, based on the test's capacity to modify the initial diagnosis and to induce a change in management resulting in clinical benefits. This approach was presented to, and accepted by our regional experts. However, by introducing a formal voting procedure, we sought confirmation that the approach would be applied in practice and not just accepted in principle. A 70% convergence between expected and observed appropriateness shows that the framework was in fact applied by panelists . Divergence from expected appropriateness was mainly due to persistent disagreement among panelists. We haven't identified "exogenous" causes for disagreement, such as rigidity in panelist opinion, corporative behavior, effects of new comers, and strength of initial disagreement. Persistent disagreement was probably due to elements inherent to the clinical scenarios (for example disagreement over treatment effectiveness or strength of rationale) but the small number of instances did not allow further investigations in this direction..

Corresponding author:
Luciana Ballini, ASSR, Emilia-Romagna Region, Italy
luballini@regione.emilia-romagna.it

References

Bossuyt Evidence-Based medical testing. Amsterdam, 2010. www.cvz.nl.
Schunemann et al BMJ 2008;336:1106-1110.
Ballini L, et al. Dossier 207/2011 (http://asr.regione.emilia-romagna.it/wcm/asr/collana_dossier/doss207.htm)
Ballini et al. Dossier 209/2011 (http://asr.regione.emilia-romagna.it/wcm/asr/collana_dossier/doss209.htm)