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

Addressing methodological challenges in evaluating diagnostic tests: combining the "Grading of Recommendations Assessment, Development and Evaluation (GRADE)" approach and the RAND/UCLA Appropriateness Method to produce clinical recommendations

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

Background

Health care decisions
have to be made irrespective of evidence (un)availability and must take into account many factors beyond test performance and treatment effectiveness
(Trikalinos 2009)

Consequentialist approach regarding the value of diagnostic test (Bossuyt 2010)

The value of any medical test is ultimately measured by whether the information it provides affects patient-relevant outcomes
(Trikalinos 2009).
The most robust empirical demonstration of the clinical effectiveness of a medical test - often unattainable – is a randomized trial

- The Regional Health Agency of Emilia-Romagna Region, Italy, was commissioned by the Health Authority to develop and regularly update guidance on the use of FDG-PET in oncology
- 2 multidisciplinary panels (with a total of 39 people) were convened to develop criteria of appropriate use of FDG-PET in 5 types of cancer (breast, esophageal, lung, colorectal, head and neck).

Objectives

- to provide a critical account of our attempt to operationalize the staged evaluation of medical tests, put forward by Bossuyt and colleagues (Bossuyt 2006), and the logic and principles of the **GRADE** approach, developed by the GRADE Group for diagnostic tests (Schunemann 2008);
- to describe the main challenges experienced by a working panel engaged in developing appropriateness criteria on the use of **FDG-PET in oncology**.

Methods

1) We developed an analytic framework in order to ensure a transparent and reproducible *consequentialist* approach for evaluation of clinical effectiveness of FDG-PET in oncology;
2) we applied the method suggested by Bossuyt et al (Bossuyt 2006) in order to develop research questions by positioning and comparing FDG-PET against existing diagnostic pathway (replacement, add-on, triage);
3) we used the GRADE approach in order to manage absence of evidence on clinical outcomes;
4) we used the voting procedure of RAND/UCLA Method of Appropriateness in order to work with 2 large panel and to formally register agreement on Criteria of Appropriateness.

Results

Between November 2010 and June 2011 two panels (total of 39 experts) met to discuss and agree on appropriate use of FDG-PET in a total of 43 clinical indications in five cancers (breast, oesophageal, lung, head and neck, colon).
Two meetings took place for each of the 5 types of cancer for a total of 10 meetings.
Two documents on FDG-PET in breast cancer and esophageal cancer have already been published (Dossier 207/2011 and 209/2011) and three are in press.

1. The analytic framework to produce criteria of appropriateness for diagnostic tests

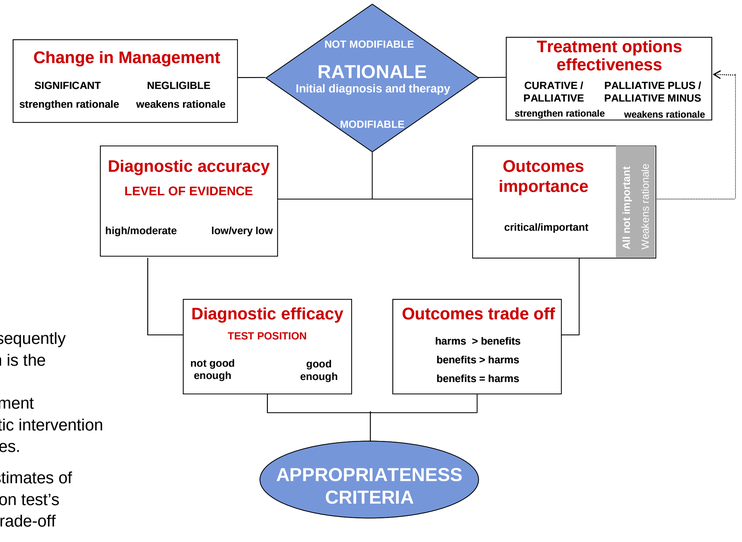
Definition of appropriateness:

- 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.

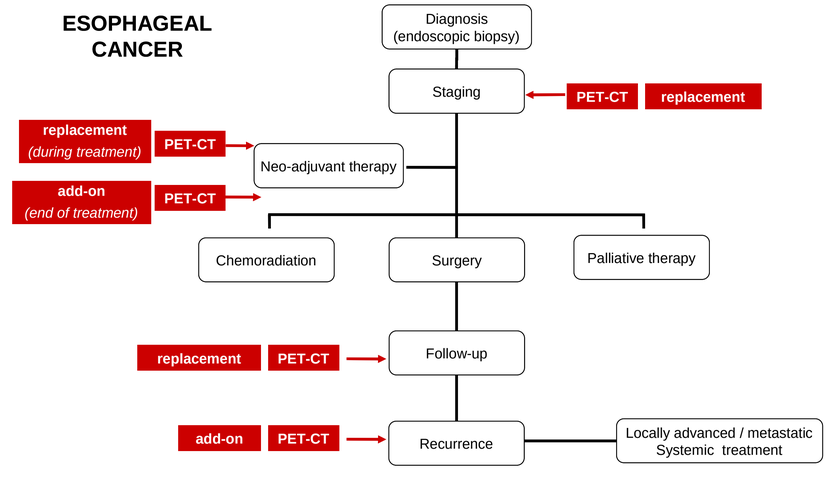
Analytic framework

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.



2. Development of research questions: positioning and comparing FDG-PET against existing diagnostic pathway



3. GRADE APPROACH: methodological challenges

Rationale and effectiveness of treatment (source: experts and clinical practice guidelines)

Vote of outcome importance and trade-off between harms and benefits

According to GRADE outcome importance are voted by members of panel (Schunemann 2008). The final level of importance helps to explicit and resolve trade-off between harms and benefits.

CLINICAL QUESTION 3:
Role of FDG-PET in early response to preoperative chemoradiation of patients treated for locally advanced esophageal cancer

Outcomes importance

Consequences of TEST for		*Level of Importance* median*
Patients Responders	True Responders: Responders complete clinically effective preoperative treatment, which could improve survival but might carries some risk of postoperative mortality	6 (2-9)
	False Non Responders: Responders interrupt clinically effective treatment, which could have improved survival, and proceed directly to surgery	8 (2-9)
Patients Non-responders	True Non Responders: Non-responders interrupt ineffective treatment, which would not have improved survival, and proceed directly to surgery, with lower risks of postoperative mortality	7 (2-9)
	False Responders: Non responders complete ineffective preoperative treatment, with no possible gain in survival but with some risk of postoperative mortality	6 (2-9)

* not important (score 1-3), important (4-6), and critical (7-9) to a decision

Matrix of natural frequencies

Patients	According to PET	According to current practice
	N of patients out of 100 submitted to the exam	
True Responders	19 - 43	43
False Non-responders	24 - 0	0
True Non-responders	42	0
False Responders	15	57
	100	100

- *Consequences for a False Non Responders judged as most important*
- *Current practice avoids risk for False Non Responders*
- *Test not good enough for replacement*
- *Panel judgement:* ***inappropriate***

FDG-PET versus comparator by means of a "matrix of natural frequencies" (Gigerenzer 2007)

How to synthesise the analytic framework: the Voting Form

CLINICAL QUESTION 3: **Role of FDG-PET in early response to preoperative chemoradiation of patients treated for locally advanced esophageal cancer**

Rationale: As preoperative chemotherapy could increase the risk of postoperative mortality (ESMO 2010), a selection of respondents after the first cycles could spare non-respondents the risks of a futile full-length chemotherapy.

Effectiveness of treatment: in patients with locally advanced cancer, preoperative chemoradiation improves the 2-year survival by 13% (absolute difference) compared to surgical treatment only (Gebski 2007). On the other hand preoperative chemotherapy could increase the risk of postoperative mortality (ESMO 2010).

Research question: FDG-PET as replacement (new test)
Is FDG-PET accurate in evaluating the early response to preoperative chemoradiation of patients treated for locally advanced oesophageal cancer?

Pre-test probability: 43% of patients show an histopathological response to neoadjuvant chemotherapy (Ngamruengphong 2010, Lorenz 2007).

Diagnostic accuracy estimates: **Level of evidence: low**

FDG-PET	sensitivity (heterogeneous) range 44-100%	specificity: 74%
Comparator (current practice: all patients complete preoperative treatment)	sensitivity: 100%	specificity: 0%

Outcomes importance

Consequences of TEST for		*Level of Importance* (1-9)*
Patients Responders	True Responders: Responders complete clinically effective preoperative treatment, which could improve survival but might carries some risk of postoperative mortality	6 (2-9)
	False Non Responders: Responders interrupt clinically effective treatment, which could have improved survival, and proceed directly to surgery	8 (2-9)
Patients Non-responders	True Non Responders: Non-responders interrupt ineffective treatment, which would not have improved survival, and proceed directly to surgery, with lower risks of postoperative mortality	7 (2-9)
	False Responders: Non responders complete ineffective preoperative treatment, with no possible gain in survival but with some risk of postoperative mortality	6 (2-9)

* not important (score 1-3), important (4-6), and critical (7-9) to a decision

Matrix of natural frequencies

	Patients	According to PET	According to current practice
		N of patients out of 100 submitted to the exam	
Patients responders	True Responders	19 - 43	43
	False Non-responders	24 - 0	0
Patients non responders	True Non-responders	42	0
	False Responders	15	57
		100	100

CLINICAL QUESTION 3:

APPROPRIATENESS of FDG-PET 1-2-3 inappropriate 4-5-6 uncertain 7-8-9 appropriate	1	2	3	4	5	6	7	8	9
INDETERMINATE									

Voting scale according to **RAND** method

Research question:
FDG-PET use expressed according to postion with respect to standard practice (Bossuyt 2006)

Why Level of Evidence and not Quality of Evidence?

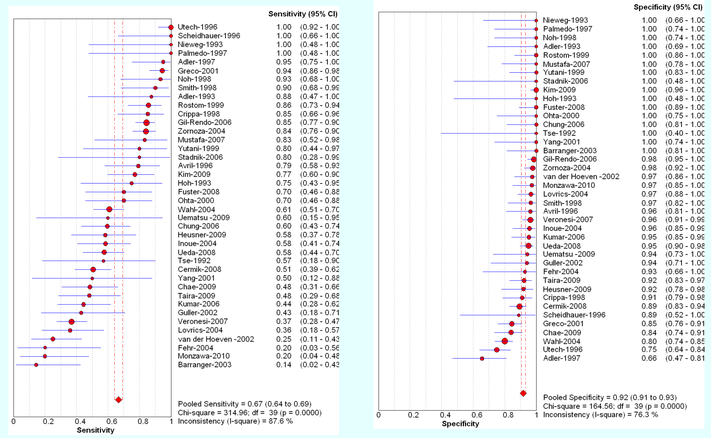
We adapted the **GRADE** scale for rating quality of evidence, using the following **levels of evidence:**

- ***High***: no risk of bias or important study limitations, consistent results from several studies and a large number of patients
- ***Moderate***: some study limitations, possible risk of bias, consistent results from several studies and a large number of patients
- ***Low***: presence of bias, inconsistency and heterogeneity of results for one estimate of diagnostic accuracy – either sensitivity or specificity – results coming from several studies and a large number of patients
- ***Very low***: presence of bias, sparse data or heterogeneity of results for both estimates of diagnostic accuracy.

Due to overall poor quality of literature, we decided not to address criteria of necessity (i.e. recommendations) and "settled" for criteria of appropriateness. This allowed differentiation of empirical findings which otherwise would have been all "flattened down" to a "Very Low" category of ***quality of evidence****, making it impossible to fulfill the mandate to discriminate appropriate from inappropriate for coverage purposes.*

How to manage heterogenity of diagnostic studies?

According to GRADE the evidence must be downgraded due to heterogeneity.
BUT: with 42 studies performed and 3.342 patients included could a judgement of ***high level of evidence of heterogeneous estimates*** be more trusty**?**
See the example below: FDG-PET for N staging in breast cancer



4. Criteria of appropriateness on FDG-PET in oncology – Emilia-Romagna Region, Italy

Phase / Cancer	Diagnosis	*N* Staging	*M* Staging	TV definition	During treatment response	End of treatment response	Follow-up	Diagnosis and staging of recurrence
Breast	inappropriate	inappropriate	disagreement	not assessed	uncertain	inappropriate	inappropriate	disagreement
Esophageal	not assessed	uncertain	appropriate	inappropriate	inappropriate	disagreement	inappropriate	disagreement
Colo-rectal	inappropriate	inappropriate	appropriate	inappropriate	indeterminate	inappropriate (rectal) disagreement (colon)	inappropriate	appropriate
Head & Neck	inappropriate appropriate (unknown primary H&N)	appropriate	appropriate	disagreement	Indeterminate	disagreement	inappropriate	appropriate
Lung	appropriate (SPN)	appropriate (NSCLC) disagreement (SCLC) inappropriate (BAC)		disagreement	inappropriate (NSCLC) inappropriate (SCLC)	Disagreement (NSCLC) inappropriate (SCLC)	inappropriate	disagreement

SPN: solitary pulmonary nodule; NSCLC: non-small cell lung cancer; SCLC: smal cell lung cancer; BAC: bronchoalveolar cancer

Full reports available at http://asr.regione.emilia-romagna.it

Discussion

The methodology proposed was very complex as 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 benefit, and involved clinical questions based on the comparison against existing diagnostic strategy (consequentialism vs essentialism).
This approach was presented to, and accepted by, all experts and each tumor resolved in two meetings.
The voting procedure of RAND/UCLA Method registered the level of agreement among panellists which was efficiently reached in 32 out of 43 clinical indications.

References

Bossuyt et al BMJ 2006;332:1089–92.
Bossuyt Evidence-Based medical testing. Amsterdam, 2010. www.cvz.nl.
Gigerenzer et al Psychological Science in the Public Interest 2007;8: 53-96.
Schunemann et al BMJ 2008;336:1106-1110.
Trikalinos et al Med Decis Making 2009;29:E22–E29.
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)

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