Statistical aspects of surgery in clinical trials

Laurence Collette, PhD
Statistics Department, EORTC, Brussels (BE)
Employee of EORTC, non profit organization

No conflict of interest to declare
Surgery in EORTC clinical trials

EORTC protocols developed > 1980 (about 700 trials) & completed

- 64 trials with Surgery
  - 15 surgery before entry
  - 49 surgery inside trial
  - Assess surgery
    - Surgery vs Surgery N=6
    - Surgery vs. no-Surgery N=8
  - Assess other components
    - Same surgery for all pts N=33
    - S optional or for salvage N=2

- Feasibility study surgical approach (RP in cT3a PCA)
- Neo-adjuvant / Perioperative treatment (eg. Liver, rectum..)
- Surgery vs other approach (Radiofrequency ablation, radiotherapy ...)
- Organ sparing approaches (lung, kidney, larynx, sphyncter ...)
“Surgery” is a complex intervention

Main constituents of a surgical intervention

Adapted from Cook JA et al. Trials 2009
Elements to keep in mind in a surgery trial

• Need not only to assess outcomes of the procedure but also measure characteristics of the process.

• Some influential factors
  • Learning curve / experience of the surgeon and team
    • Required prior to doing the study or adjusted for in statistical analysis
    • Record surgeon ID, info about # of prior operations, sequence of operation..
  • Ascertainment bias (pathology, PROs …)
    • How to blind?
  • Performance bias (influence of postoperative management)
    • Record center-specific data, provide guidelines
  • Drop outs / refusals / inoperability after entry on study
    • Record adherence to procedure, specify how to handle in the data analysis
  • Acceptance to patients (esp. randomized trials)
    • Informed consent by a third party, not directly involved in any of the treatments being compared

Need a prospective plan (i.e. written protocol)
Elements to keep in mind in a surgery trial

- Need not only to assess outcomes of the procedure but also measure characteristics of the process.

- **Prospective plan (i.e. written protocol)**
  - Specify clear objectives, set a clear goal
  - Precisely define endpoints, specify primary
  - Specify which process parameters are controlled, which are only recorded for integration in the data analysis
  - Data collection plan
  - Endpoint assessment process
  - Data analysis plan
  - Quality assurance (training when required)
## IDEAL framework for surgical innovation (BMJ 2013)

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Patients</th>
<th>Study Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>IDEA:</strong></td>
<td>Proof of concept of new technique</td>
<td>Few patients, single center</td>
<td>First-in-man study, structured case reports</td>
</tr>
<tr>
<td>2a. <strong>DEVELOPMENT:</strong></td>
<td>Optimize technique and patient selection. Show safety and efficacy</td>
<td>10s of consecutive patients, select centers</td>
<td>Prospective development study</td>
</tr>
<tr>
<td>2b. <strong>EXPLORATION:</strong></td>
<td>Assess efficacy with more widespread use of the technique</td>
<td>100s of consecutive patients, multiple centers and surgeons</td>
<td>Prospective collaborative (controlled) observational study / feasibility randomized study</td>
</tr>
<tr>
<td>3. <strong>ASSESSMENT:</strong></td>
<td>Comparative effectiveness to other treatments</td>
<td>100s+ of patients, broad eligibility (“equipoise”), many centers</td>
<td>Randomized controlled “pragmatic” trial</td>
</tr>
<tr>
<td>4. <strong>LONG TERM STUDY:</strong></td>
<td>Long terms effects of the proecure, Quality assurance</td>
<td>100s+ of patients, standard practice</td>
<td>Observational study, comprehensive disease-based registry</td>
</tr>
</tbody>
</table>
“It is always too early [for rigorous evaluation] until suddenly it is too late” (Buxton’s law)

Surgery can be trialed but only with the right timing, once broadly adopted, it is generally too late

Overcoming barriers to randomization

<table>
<thead>
<tr>
<th>Box 1</th>
<th>Potential solutions to overcome common variations in surgical randomised controlled trials</th>
</tr>
</thead>
</table>

**Surgeon preferences**
- Maximise flexibility in the delivery of surgical interventions, beyond the key distinctive elements, to allow for variation in surgeon and centre practices
- Implement recruitment of participants by a third party
- Use broad patient eligibility criteria
- Undertake preliminary work to establish consensus regarding community uncertainty
- Adopt an expertise based trial design

**Patient preferences**
- Undertake a qualitative evaluation of patients’ perspectives and experiences

**Quality control of the intervention**
- Use criteria for surgeon eligibility (for example, training and previous number of cases)
- Record an objective measure of quality (for example, lymph node yield for gastric cancer surgery)
- Record indicators of surgical decision making (for example, conversion from partial to total knee replacement, or from laparoscopic to open surgery)

*Cook JA et al.*
*IDEAL framework 3*
*BMJ 2013*
Trial endpoints

**Short-term endpoints**

Intra-operative variables
- Length of operation, blood loss

Postoperative morbidity
- Postoperative AE rates
- Duration of hospital stay

Short term oncologic outcome
- Completeness of excision
- Pathological response
- (organ function)

**Long-term endpoints**

- Long term AE
- Progression/disease-free survival
- Survival
- Patient reported outcomes
- (need for reoperation)
- Cost-effectiveness

**Assessment trials (IDEAL phase 3)**

**Exploration trials (IDEAL phase 2b)**
What are we talking about??

- ... 56 definitions and measurement of anastomotic leak after GI surgery in a review of 107 studies (Bruce et al. Br J Surg 2001)
- ... 10 different measures of mortality in studies of oesophgectomy (Blencove et al. Ann Surg 2012)
- ... in colorectal cancer surgery, 766 different clinical outcomes assessed, inconsistently measured and reported... (Whistance, Colorectal Dis, 2013)

Need a common terminology for reporting surgical outcomes and contextual factors (international standards)

COMET-initiative, Clavien-Dindo classification of surgical complications, CTCAE, ...
Who assesses the outcome matters

EORTC 40954 Neoadjuvant CT followed by surgery versus surgery alone in locally advanced gastric cancer

Comparison between **surgeon** and **pathologist**

<table>
<thead>
<tr>
<th>Arm</th>
<th>R-Type of resection</th>
<th>Surgeon</th>
<th>Pathologist</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCtx</td>
<td>R0-resection</td>
<td>90.0%</td>
<td>81.9%</td>
</tr>
<tr>
<td></td>
<td>R1-resection</td>
<td>5.7%</td>
<td>13.9%</td>
</tr>
<tr>
<td></td>
<td>R2-resection</td>
<td>2.9%</td>
<td>1.4%</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>1.4%</td>
<td></td>
</tr>
<tr>
<td>Surgery</td>
<td>R0-resection</td>
<td>92.6%</td>
<td>66.7%</td>
</tr>
<tr>
<td></td>
<td>R1-resection</td>
<td>0%</td>
<td>22.2%</td>
</tr>
<tr>
<td></td>
<td>R2-resection</td>
<td>7.4%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>
Who assesses the outcome matters

Impact of pathology review of stage and margin status of radical prostatectomy specimens (EORTC trial 22911)

Theodorus H. van der Kwast et al. Virchows Arch 2006

Table 5  Agreement between review and local pathology

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Overall agreement (%)</th>
<th>Simple kappa (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All centers (n=552)</td>
</tr>
<tr>
<td>Seminal vesicle invasion</td>
<td>92.8</td>
<td>0.83&lt;sup&gt;a&lt;/sup&gt; (0.77–0.89)</td>
</tr>
<tr>
<td>Extraprostatic extension</td>
<td>68.2</td>
<td>0.33&lt;sup&gt;a,b&lt;/sup&gt; (0.25–0.39)</td>
</tr>
<tr>
<td>Surgical margin status</td>
<td>69.4</td>
<td>0.45&lt;sup&gt;a,c&lt;/sup&gt; (0.37–0.53)</td>
</tr>
</tbody>
</table>

<sup>a</sup>The cases classified “unknown” or “not evaluable” are excluded.
<sup>b</sup>The categories “focal” and “extensive” are grouped together.
<sup>c</sup>The categories “apex positive,” “lateral positive” and “apex and lateral positive” are grouped together.
The case of pathological response

- “The primary outcome measure in definitive trials should be a clinical event relevant to the patient’, or an endpoint that measures directly how a patient feels, functions or survives”
  - “functions” refers to activities of daily living (not eg. organ function)
  - The above is FDA language
  - The word “clinical” is used so often that it has become meaningless

*Fleming and Powers, Stat Med 2012*

**pCR is a acceptable endpoint for definitive studies only if demonstrates surrogacy for final endpoint**
Is pCR relevant to the patient?

NOTE: pCR is binary (yes/no)

Does achieving pCR imply a benefit for the patients: yes, HOWEVER:

In the group of patients not achieving pCR there is still a whole range of different underlying tumor responses:

- almost pCR
- no change
- progressive disease

The pCR endpoint does not capture this heterogeneity, while these differences will likely impact the patients disease course (relevant to the patient).
Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis


Summary
Background Pathological complete response has been proposed as a surrogate endpoint for prediction of long-term clinical benefit, such as disease-free survival, event-free survival (EFS), and overall survival (OS). We had four key objectives: to establish the association between pathological complete response and EFS and OS, to establish the definition of pathological complete response that correlates best with long-term outcome, to identify the breast cancer subtypes in which pathological complete response is best correlated with long-term outcome, and to assess whether an increase in frequency of pathological complete response between treatment groups predicts improved EFS and OS.

Findings We obtained data from 12 identified international trials and 11955 patients

Interpretation Patients who attain pathological complete response defined as ypT0 ypN0 or ypT0/is ypN0 have improved survival. The prognostic value is greatest in aggressive tumour subtypes. Our pooled analysis could not validate pathological complete response as a surrogate endpoint for improved EFS and OS.

Funding US Food and Drug Administration.
pCR: prognostic for EFS/OS


pCR is ypT0/is ypN0
pCR: not a surrogate for EFS/OS
the magnitude of improvement in pCR rate did
not predict EFS or OS effect

R²=0.03 (95%CI: 0.00-0.25)
R²=0.24 (95%CI: 0.00-0.70)

Meta-analysis Results from the Collaborative Trials in Neoadjuvant Breast Cancer (CTNeoBC).
Cortazar et al. Lancet 2014
The treatment question

Effect of treatments on pCR

Long-term: EFS, OS

The extent to which pCR is
a causal intermediate step in getting to the clinical endpoint (YES)

Other (unobserved) treatment effects on long term endpoint
(e.g. effects on distant subclinical disease)

pCR dichotomous ("boundary effect"), other high-risk factors may be present despite pCR,
impact of adjuvant treatments etc..

Large effects on pCR needed to change EFS/OS
Recommendations on trial design features

• **Prospective plan with a specified research hypothesis** (all studies, not just randomized ones)
• **Selection criteria (on pts, surgeons, centers) must match objectives** (broader for pragmatic efficacy trials, to ensure generalizability of results)
• **Design** experiment to minimize risk of biases (central blind assessment of outcome, consent by third party ..)
• **Fully defined endpoints** as well as methods of measurement (who much change needed?) and process of assessment (who assesses?). Specify one as primary
• **Extensive record of relevant parameters using standardized tools** (patient and disease characteristics (to adjust case mix), surgeon learning curve, QA parameters (how the surgery was done, why it was changed, duration of surgery, blood loss..), pathology reports, data on postoperative management, complications and outcome)
Further reading on methodology

Phenotypic differences between male physicians, surgeons, and film stars: comparative study

Antoni Trilla, Marta Aymerich, Antonio M Lacy, Maria J Bertran

We finished our medical training at the University of Barcelona more than 25 years ago, and have enjoyed our work ever since. At medical school we noted certain differences between male trainees who selected either surgery or medicine as their speciality. The tallest and most handsome male students were more likely to go for surgery, and the shortest (and perhaps not so good looking) ones were more likely to become physicians (including doctors of internal medicine and its subspecialties).

Now, after all these years we hypothesise that, on average, surgeons are taller and better looking than physicians. We conducted a comparative study to test this hypothesis.

Methods

We selected a random sample of senior staff surgeons and physicians working at the University of Barcelona Hospital Clinic (a 700 bed public hospital), matched by age (52 ±7 years) and sex (all men), from the staff payroll of the surgical and medical departments. We contacted all eligible participants by email. If they agreed to participate, their height (in cm) was recorded and they were asked to submit a digital picture. Age (in years) was registered and checked against that participants (no further checking of this information was attempted). We decided to avoid (for the time being) male observers, because of potential bias. Observers used the “good looking score” to classify each participant. This score measures the degree of handsomeness on a seven point Likert scale (1, ugly; 7, very good looking).

We discarded the highest and lowest score (outliers) for each participant and used the six remaining scores for our study. Mean scores, differences in means with 95% confidence intervals, and standard deviations were used to compare the three groups. We used the standard t test to compare age and the non-parametric (Mann-Whitney U) test to compare height and mean looking scores.

Results

We contacted 14 surgeons and 16 physicians (24 surgeons and 38 physicians were eligible). Only two surgeons and two physicians did not answer the questionnaire or send a picture (their out of office auto reply was switched on). Two additional physicians were dropped from the final analysis because of the poor quality (technical, of course) of their pictures. The final analysis therefore comprised 12 physicians and 12
# Results

<table>
<thead>
<tr>
<th></th>
<th>Surgeons</th>
<th>Physicians</th>
<th>Movie stars</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>51.1</td>
<td>50.6</td>
<td></td>
<td>0.76</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>179.4 (95%CI 175.1-184.0)</td>
<td>172.6 (95%CI 170.2-175.4)</td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Good looking score (1-7)</td>
<td>4.39</td>
<td>3.65</td>
<td>5.96</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>3.65</td>
<td>5.96</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.96</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Thank you very much for your attention

Now is time for questions
### Levels of evidence for efficacy

<table>
<thead>
<tr>
<th>Direct</th>
<th>A true clinical efficacy measure</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>A validated surrogate</td>
<td>PFS in adjuvant colorectal cancer, 5FU trt</td>
</tr>
<tr>
<td>Indirect</td>
<td>A non-validated surrogate, reasonably likely to predict clinical benefit</td>
<td>Large effect on PFS in some solid tumors</td>
</tr>
<tr>
<td>Indirect</td>
<td>A correlate measuring biological activity</td>
<td>PSA in prostate</td>
</tr>
</tbody>
</table>
## Comparison of important cancer approval endpoints

<table>
<thead>
<tr>
<th>Regulatory Evidence</th>
<th>Endpoint</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surrogate</strong></td>
<td>RR / CRR</td>
<td>1-arm possible, smaller, shorter, attributable to drug</td>
<td>No direct measure of benefit / no comprehensive measure of drug activity / only subset of benefiting pats.</td>
</tr>
<tr>
<td></td>
<td>DFS</td>
<td>Smaller, shorter</td>
<td>Not stat. validated as surrogate for OS / not precise, open to bias / many definitions</td>
</tr>
<tr>
<td></td>
<td>PFS</td>
<td>Smaller, shorter, SD included, crossover / other Tx not affecting, objective &amp; quantitative</td>
<td>Not stat. validated as surrogate for OS / not precise, open to bias /many definitions / frequent assessments / need to balance timing x arms</td>
</tr>
<tr>
<td><strong>Clinical benefit</strong></td>
<td>Symptoms</td>
<td>Patient perspective of direct clinical benefit</td>
<td>Blinding hard, missing data, clinically relevant effect, validated tools lacking</td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Direct measure of benefit, easy, precise</td>
<td>Large studies, crossover / follow-up Tx affects, non-cancer deaths</td>
</tr>
</tbody>
</table>
Summary

• Generalizability
• Randomized vs not randomized
• Endpoints
  • Selection
  • Evaluation
  • Definition
• pCR
Survival of patients with rectal carcinoma treated by TME and radiotherapy before during and after the Dutch TME trial

Positive impact of the trial through training of the surgeon on TME, standardization of RT and of pathology, and QA leading to national treatment guidelines
The surgeon as a prognostic factor variability among 13 consultant surgeons

- Curative resection (R0): 40 – 76%
- Anastomotic leakage: 0 – 25%
- Postoperative mortality: 8 – 30%
- Local recurrence: 0 – 21%
- Survival: 20 – 63%

Randomization: Surgeons can do it!

Long-term outcome of laparoscopic surgery for colorectal cancer: A cochrane systematic review of randomised controlled trials

Trials done in US, China, UK, The NL, Italy, Spain ...

>3500 patients randomized
Ethical issues in testing a theoretically superior treatment”

Belief in the new compared with the old seems to be deeply rooted in our culture, and so is the belief in technology, where even diseases gain prestige through technology

MEDICINE is an ART, isn’t it? It can’t be reduced to just logical reasoning
Fig 2 | Example of surgical innovation: laparoscopic procedure adoption. Reproduced from reference 8 with permission. Data are percentage of operations carried out using a laparoscopic approach in 1989-2003, from the Nationwide Inpatient Sample, a nationally representative annual sample of hospital admissions in the United States.