# Outcome Studies

Patrick MM Bossuyt

Amsterdam UMC
University Medical Centers

# Financial disclosure

No actual or potential conflict of interest
in relation to this presentation,
or any of the products mentioned in it.

# What are outcome studies?

*“Outcome studies focus on*
*the end results of medical care:*

*the effect of the health care process on*
*the health and well-being*
*of patients and populations.”*

NIH GRANTS & FUNDING
NIH Central Resource for Grants and Funding Information

# Does Your Human Subjects Research Study Meet the NIH Definition of a Clinical Trial?

A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes. Learn more

**Answer the following four questions to determine if your study is a clinical trial:**

1. Does the study involve human participants?
2. Are the participants prospectively assigned to an intervention?
3. Is the study designed to evaluate the effect of the intervention on the participants?
4. Is the effect being evaluated a health-related biomedical or behavioral outcome?

**Your study is considered to meet the NIH definition of a clinical trial even if:**

- Your study uses healthy participants, or does not include a comparison group (e.g., placebo or control)
- Your study is only designed to assess the pharmacokinetics, safety, and/or maximum tolerated dose of an investigational drug
- Your study utilizes a behavioral intervention
- Your study uses an intervention for the purposes of understanding fundamental aspects of a phenomenon (See guidance and FAQs about Basic Experimental Studies with Humans (BESH)).

**Your study is NOT considered to meet the NIH definition of a clinical trial if:**

- Your study is intended solely to refine measures.
- Your study involves secondary research with biological specimens or health information.

https://grants.nih.gov/ct-decision

# Learning objectives

After this session, students should be able to explain

- some of the difficulties in imaging RCT
- more efficient designs for randomized trials in imaging
- how STARD 2015 can reduce waste in imaging research

# Outcomes studies: Outline

1. Clinical effectiveness
2. Imaging RCT: Challenges
3. Imaging RCT: Efficient Designs
4. Waste in Research and STARD reporting guidelines

# 1.
# Clinical Effectiveness

In imaging

# Clinical Effectiveness

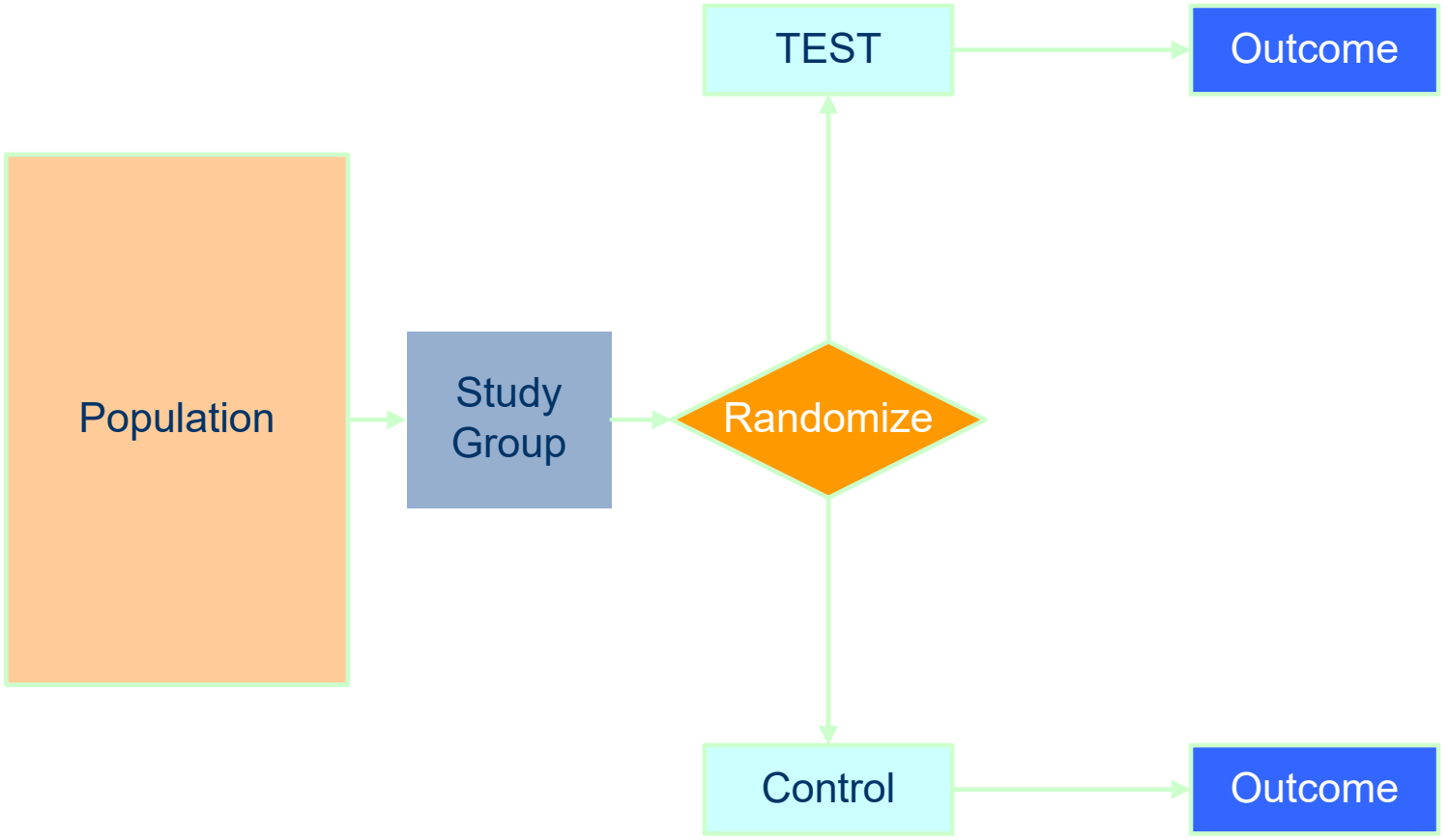
Change in patient outcomes

when implementing
a (different) healthcare intervention

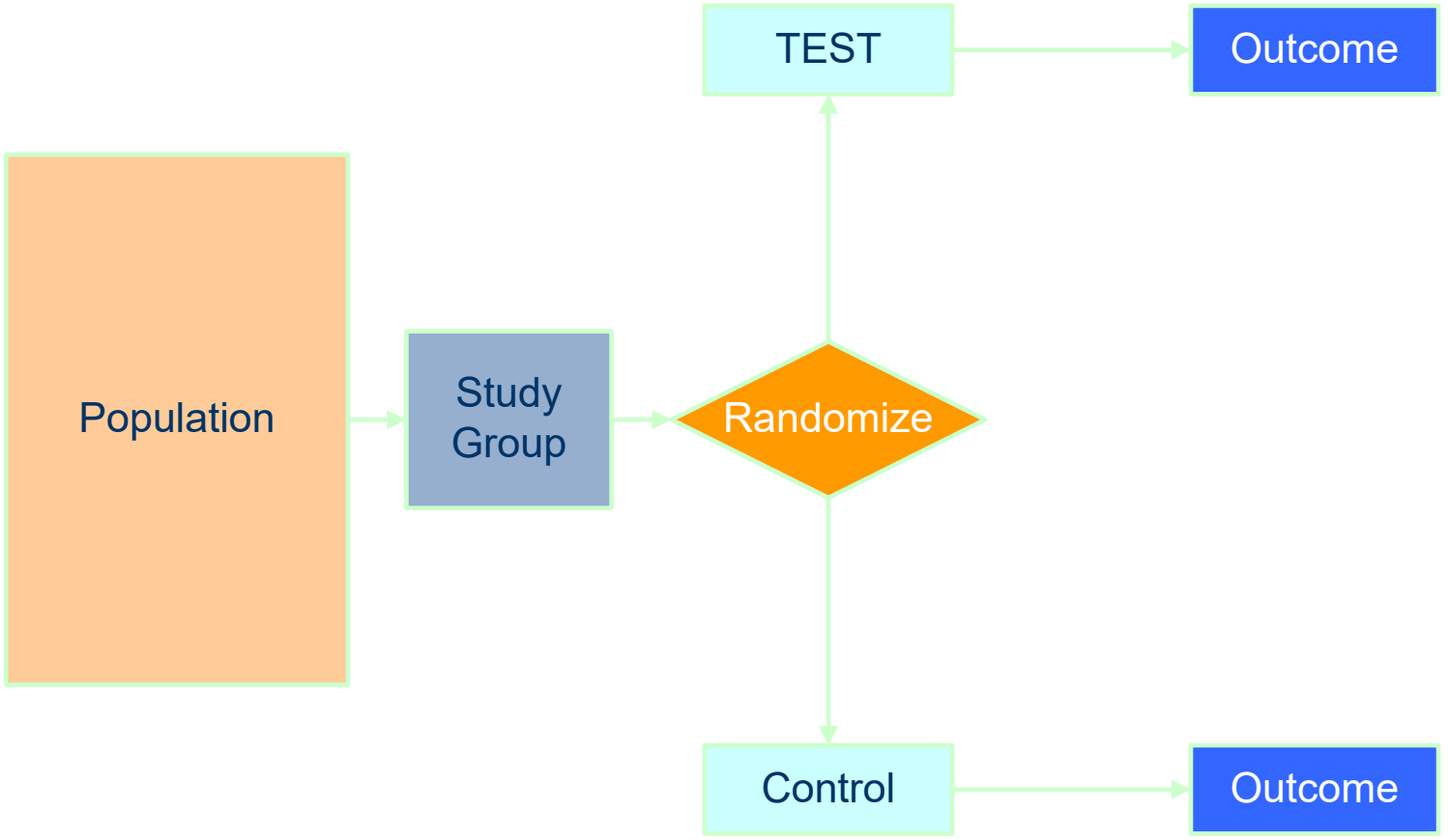
# Clinical Effectiveness

| | |
|---|---|
| **Health Outcome** | Health outcomes that matter to patients and society:<br>to prevent premature death,<br>to restore or maintain functional health. |
| **Probabilistic** | Not all outcomes will be observed in everyone tested;<br>evaluations will be made at the group level,<br>and expressed in terms of a distribution of outcomes. |
| **Comparative** | Effectiveness of testing is defined<br>relative to a comparator strategy:<br>current best standard practice. |

# Clinical Effectiveness



# Medical Test RCT



# 2.
# Imaging RCT: Challenges

Radiology

Marc C. J. M. Kock, MD, MSc
Miraude E. A. P. M. Adriaensen, MD, MSc[2]
Peter M. T. Pattynama, MD, PhD
Marc R. H. M. van Sambeek, MD, PhD
Hero van Urk, MD, PhD
Theo Stijnen, PhD
M. G. Myriam Hunink, MD, PhD

**Published online**
10.1148/radiol.2372040616
**Radiology 2005**; 237:727–737

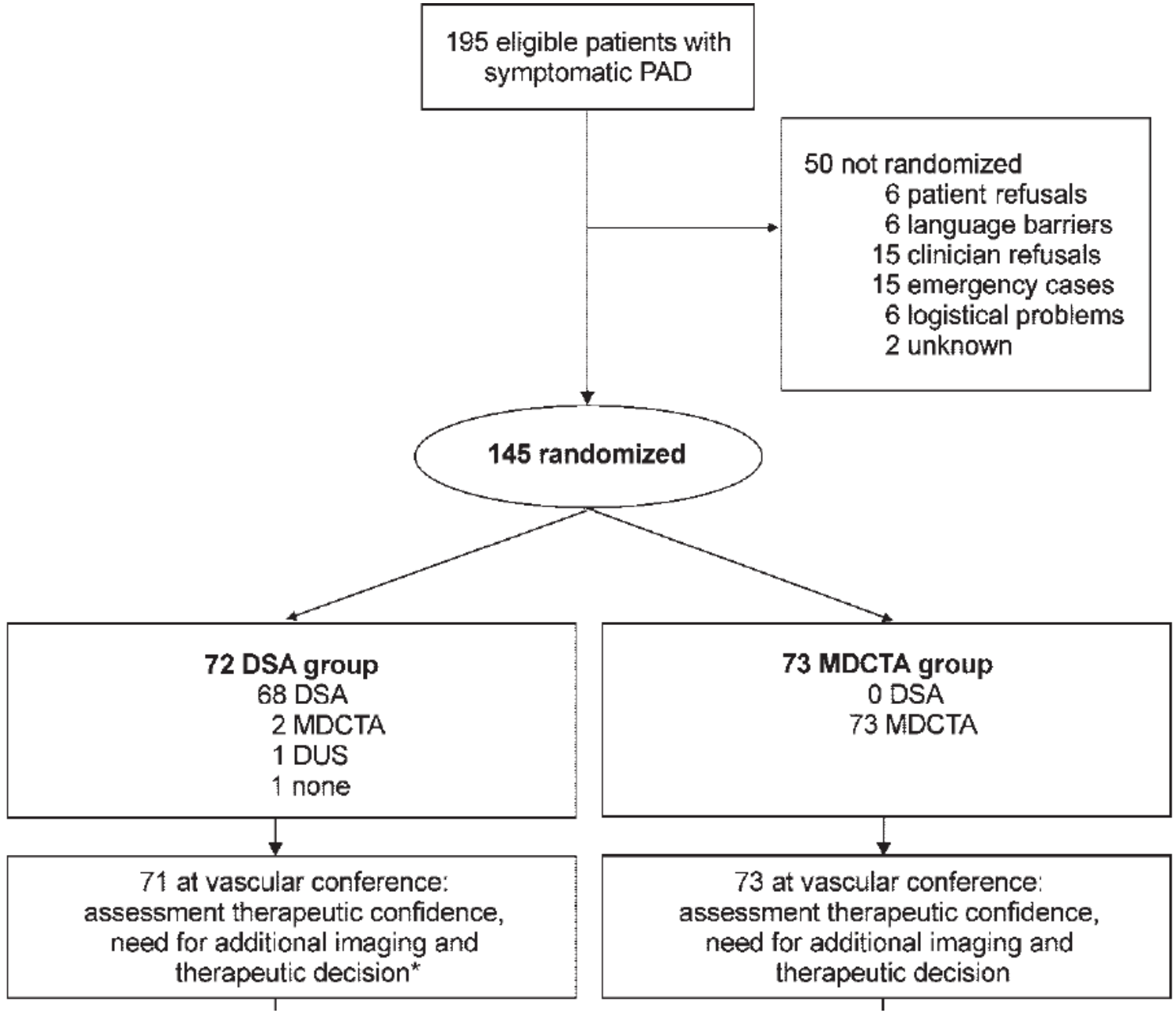
**Abbreviations:**
CI = confidence interval
DSA = digital subtraction angiography
EQ-5D = EuroQol-5D
PAD = peripheral arterial disease
SF-36 = Medical Outcomes Study 36-Item Short Form Health Survey

From the Program for the Assessment of Radiological Technology (M.C.J.M.K., M.E.A.P.M.A., M.G.M.H.) and the Departments of Radiology (M.C.J.M.K.

# DSA versus Multi–Detector Row CT Angiography in Peripheral Arterial Disease: Randomized Controlled Trial[1]

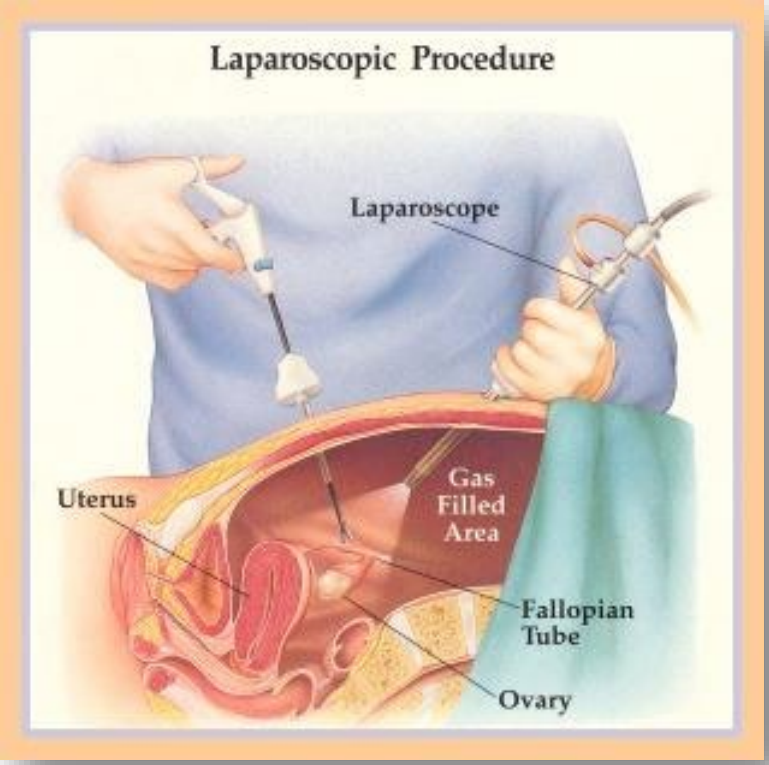
**PURPOSE:** To prospectively compare therapeutic confidence in, patient outcomes (in terms of quality of life) after, and the costs of digital subtraction angiography (DSA) with those of multi–detector row computed tomographic (CT) angiography as the initial diagnostic imaging test in patients with peripheral arterial disease (PAD).

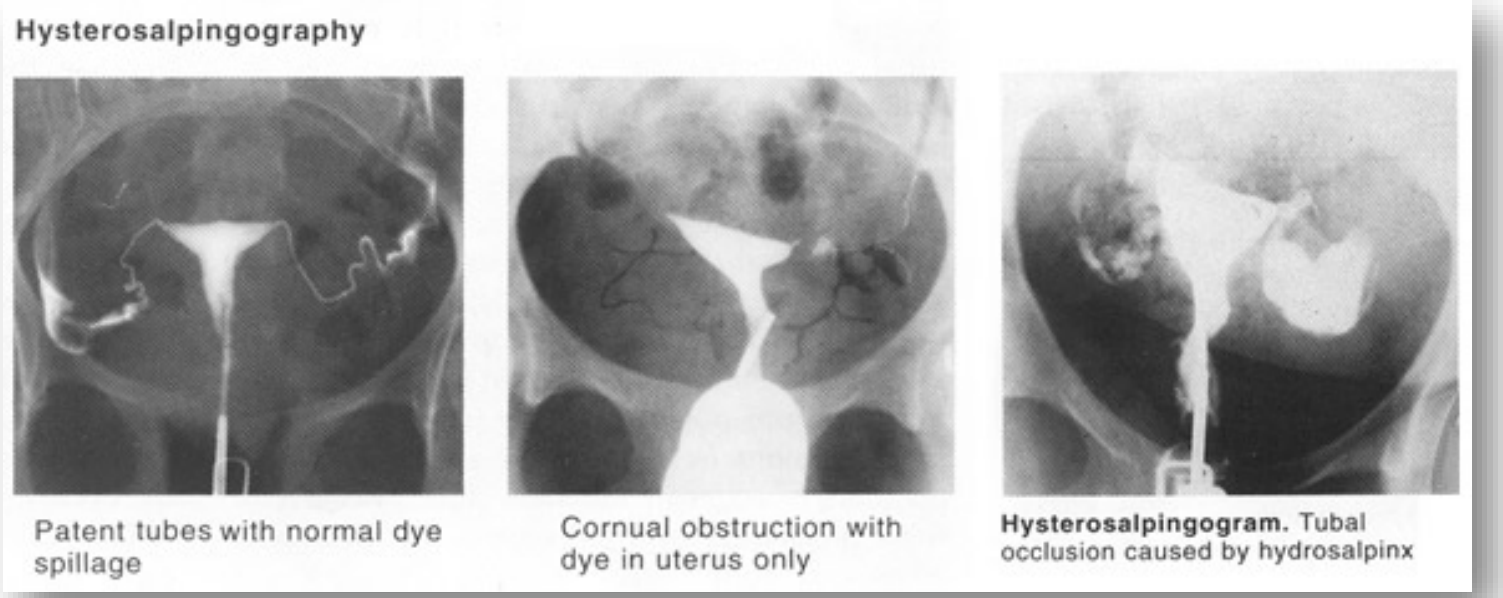
**MATERIALS AND METHODS:** Institutional medical ethics committee approval and patient informed consent were obtained. Between April 2000 and August 2001, patients with PAD were randomly assigned to undergo either DSA or multi–detector row CT angiography as the initial diagnostic imaging test. Outcomes were the therapeutic confidence assessed by physicians (on a scale from 0 to 10), the need for additional imaging, the health-related quality of life at 6-month follow-up, diagnostic and therapeutic costs, and the costs for a hospital stay. Costs were computed from a hospital perspective according to Dutch guidelines for cost calculations in health care. Mean outcomes were compared between groups with unpaired *t* testing and were adjusted for predictive baseline characteristics with multivariable regression analysis.


195 eligible patients with symptomatic PAD
50 not randomized
6 patient refusals
6 language barriers
15 clinician refusals
15 emergency cases
6 logistical problems
2 unknown
145 randomized
72 DSA group
68 DSA
2 MDCTA
1 DUS
1 none
73 MDCTA group
0 DSA
73 MDCTA
71 at vascular conference: assessment therapeutic confidence, need for additional imaging and therapeutic decision*
73 at vascular conference: assessment therapeutic confidence, need for additional imaging and therapeutic decision

angiography group. There were 47 men in the DSA group and 58 men in the CT angiography group. Physician confidence in making a correct therapeutic choice was significantly higher at DSA (mean confidence score, 8.2) than at CT angiography (mean score, 7.2; $P < .001$). During 6-month follow-up, 14% less additional

# Tubal integrity testing



Human Reproduction doi:10.1093/humrep/dei478

# Routine use of hysterosalpingography prior to laparoscopy in the fertility workup: a multicentre randomized controlled trial

D.A.M.Perquin[1,4], P.J.Dörr[1], A.J.M.de Craen[2] and F.M.Helmerhorst[3]

[1]Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, The Hague, [2]Department of Gerontology and Geriatrics and [3]Department of Gynaecology, Division of Reproductive Medicine, Leiden University Medical Centre, Leiden, The Netherlands

[4]To whom correspondence should be addressed at: Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, PO Box 432, 2501 CK, The Hague, The Netherlands. E-mail: dperquin@knoware.nl

**BACKGROUND: A multicentre randomized controlled trial with or without hysterosalpingography (HSG) was conducted to assess the usefulness of HSG as a routine investigation in the fertility workup prior to laparoscopy and dye. METHODS: From 1 April 1997 to 1 April 2002, subfertile women were allocated by a computer–based 1 : 1 ratio randomization procedure, either for an HSG followed by laparoscopy and dye (the intervention group) of for laparoscopy and dye only (the control group) as a part of their fertility workup. Cumulative pregnancy rate (CPR) within 18 months after randomization was the primary outcome of interest. RESULTS: 344 women were randomized to the intervention group ($n$ = 169) and the control group ($n$ = 175). There was no significant difference in CPR at 18 months in the intervention group (49.1%) [95% confidence interval (CI) 41.6 to 56.6] and the control group (50.3%) (95% CI 42.8 to 57.8), a difference of –1.2% (95% CI –11.8% to 9.5%). CONCLUSION: The routine use of HSG at an early stage in the fertility workup prior to laparoscopy and dye does not influence CPR, compared with the routine use of laparoscopy and dye without HSG.**

*Key words*: hysterosalpingography/laparoscopy and dye/pregnancy rate/randomized controlled trial

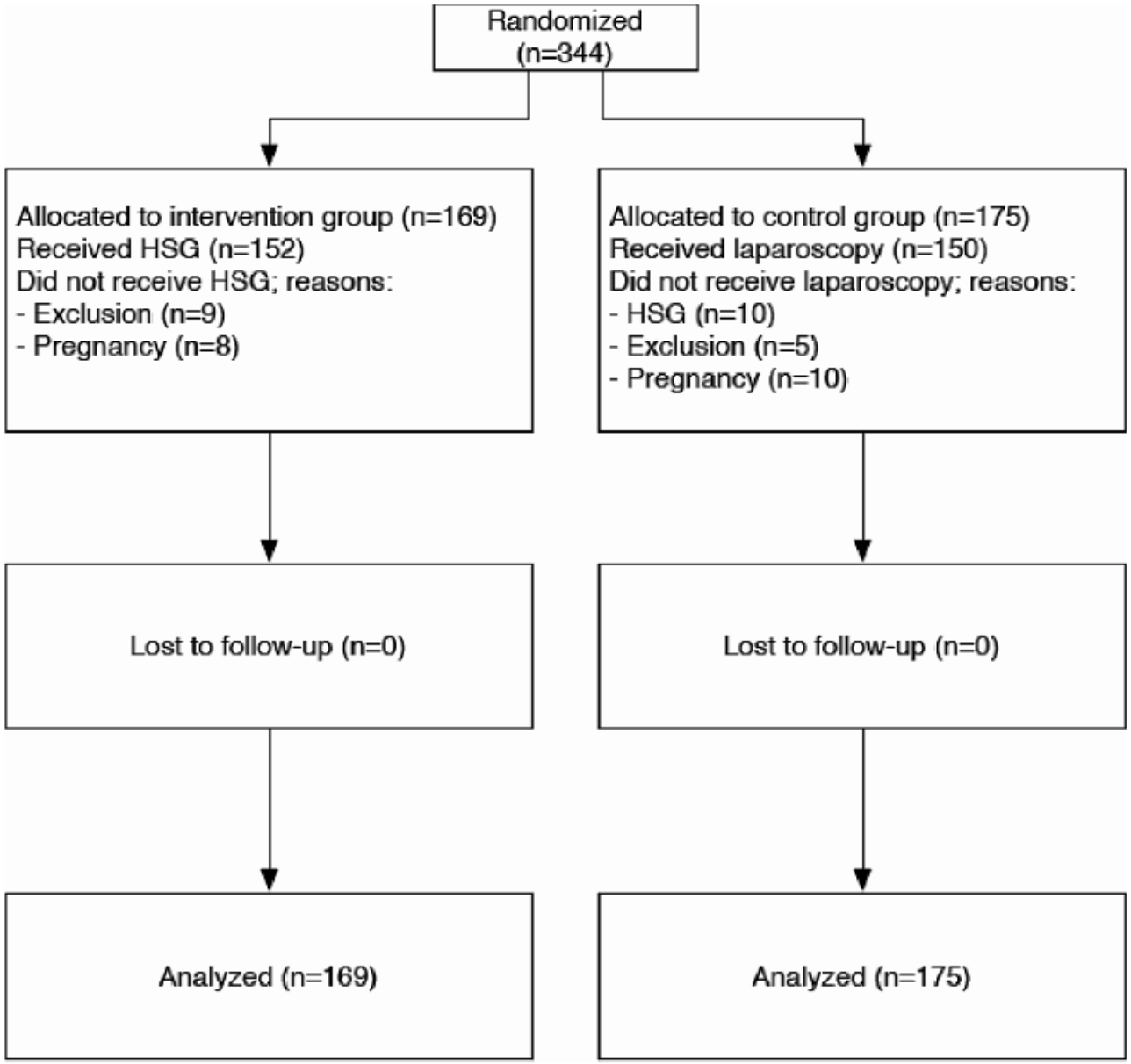
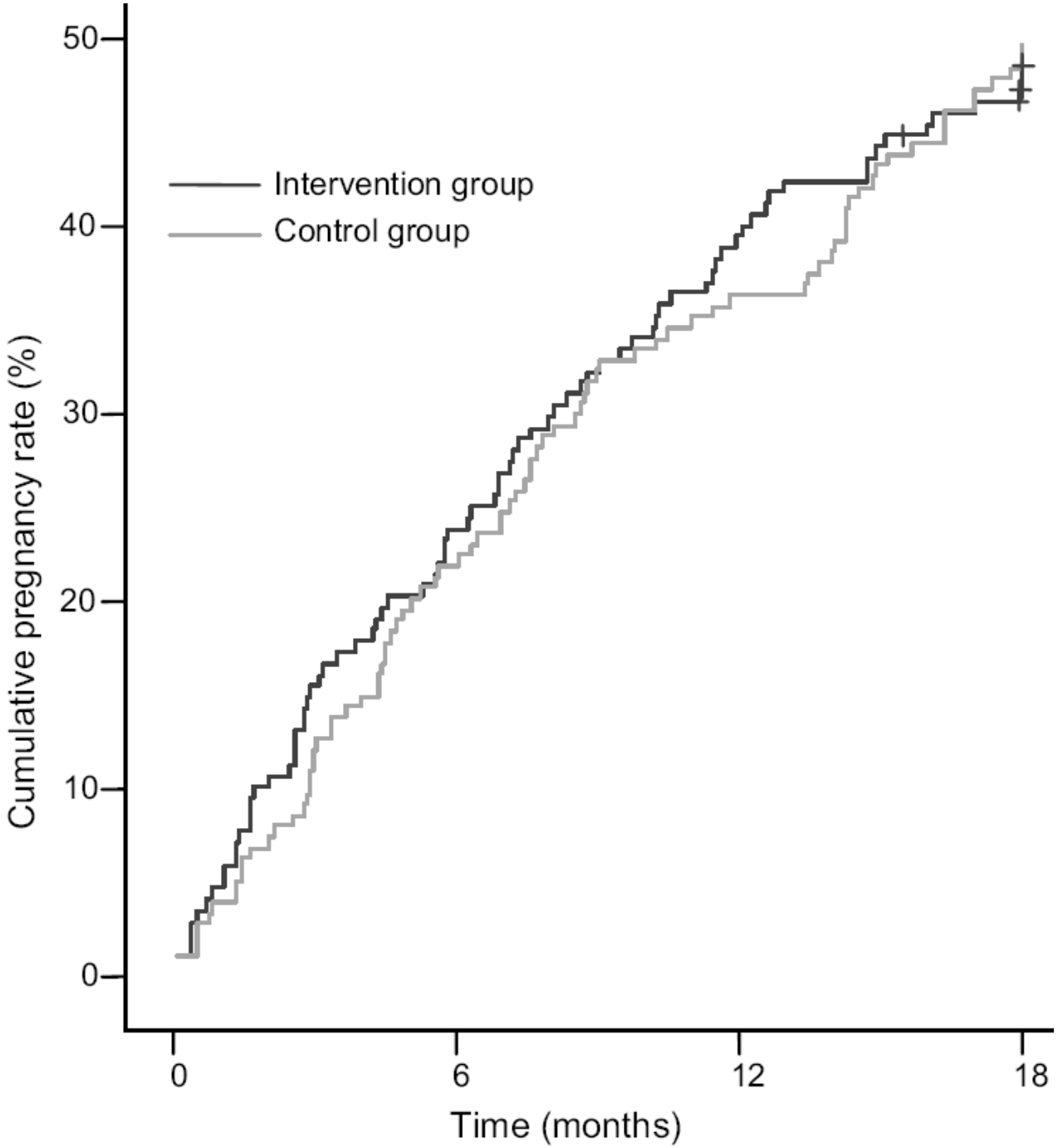



**Figure 1.** Flow chart of participants.




Hum Reprod. 2006 21:1227-31

Hum. Reprod. Advance Access published February 13, 2006 doi:10.1093/humrep/del478

# Routine use of hysterosalpingography prior to laparoscopy in the fertility workup: a multicentre randomized controlled trial

D.A.M.Perquin[1,4], P.J.Dörr[1], A.J.M.de Craen[2] and F.M.Helmerhorst[3]

[1]Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, The Hague, [2]Department of Gerontology and Geriatrics and [3]Department of Gynaecology, Division of Reproductive Medicine, Leiden University Medical Centre, Leiden, The Netherlands

[4]To whom correspondence should be addressed at: Department of Obstetrics and Gynaecology, Medical Centre Haaglanden, PO Box 432, 2501 CK, The Hague, The Netherlands. E-mail: dperquin@knoware.nl

**BACKGROUND: A multicentre randomized controlled trial with or without hysterosalpingography (HSG) was conducted to assess the usefulness of HSG as a routine investigation in the fertility workup prior to laparoscopy and dye. METHODS: From 1 April 1997 to 1 April 2002, subfertile women were allocated by a computer–based 1 : 1 ratio randomization procedure, either for an HSG followed by laparoscopy and dye (the intervention group) of for laparoscopy and dye only (the control group) as a part of their fertility workup. Cumulative pregnancy rate (CPR) within 18 months after randomization was the primary outcome of interest. RESULTS: 344 women were randomized to the intervention group ($n$ = 169) and the control group ($n$ = 175). There was no significant difference in CPR at 18 months in the intervention group (49.1%) [95% confidence interval (CI) 41.6 to 56.6] and the control group (50.3%) (95% CI 42.8 to 57.8), a difference of –1.2% (95% CI –11.8% to 9.5%). CONCLUSION: The routine use of HSG at an early stage in the fertility workup prior to laparoscopy and dye does not influence CPR, compared with the routine use of laparoscopy and dye without HSG.**

*Key words*: hysterosalpingography/laparoscopy and dye/pregnancy rate/randomized controlled trial

## Introduction

After history taking, physical examination, semen analysis and investigation of ovulation, assessment of tubal patency is the next step in the standard examination of the subfertile couple. Owing to the noninvasive nature and low cost, hysterosalpingography (HSG) is widely used as a first-line approach to assess the patency of the Fallopian tubes in routine fertility workup (Helmerhorst *et al*., 1995; Mol *et al*., 2001), although laparoscopy and dye is considered the gold standard (Rowe *et al*., 1993; Swart *et al*., 1995).

A reason for performing HSG instead of or prior to laparoscopy and dye cannot be found in the test characteristics of HSG. Comparing the accuracy of HSG with that of laparoscopy and dye in the diagnosis of tubal pathology, a meta-analysis demonstrated point estimates of 65% of sensitivity and 83% of specificity (Swart *et al*., 1995). Furthermore, considerable variability in the interpretation as well as clinical consequences of HSG abnormalities has been shown among practitioners (Mol *et al*., 1996; Glatstein *et al*., 1997). Advantages of HSG relative to laparoscopy are the short outpatient procedure and the enhancement of pregnancy with oil-soluble contrast medium (Johnson *et al*., 2005), although water-soluble media are mostly used (Glatstein *et al*., 1998). The therapeutic effect of tubal flushing with water-soluble media is, however, still unknown (National Institute for Clinical Excellence, 2004).

The relative merits of HSG and laparoscopy in screening for tubal factors have been discussed for more than 30 years, but so far no randomized controlled trial has been reported (Helmerhorst *et al*., 1995). To assess the value of HSG prior to laparoscopy and dye in a routine clinical setting, we performed a pragmatic multicentre randomized controlled trial comparing fertility workups with or without HSG. In a pragmatic trial, effectiveness of an intervention is assessed under usual circumstances, in contrast to efficacy trials in which the intervention is examined under ideal conditions (Haynes, 1999). Is the patient better off with or without the extra intervention (in this case, HSG)? We compared the two strategies, with pregnancy as a clinical endpoint, in terms of cumulative pregnancy rate (CPR).

## Subjects and methods

### *Patients and randomization procedure*

The study was performed in three teaching hospitals in The Netherlands. All newly referred and admitted subfertile women who visited the Department of Reproductive Medicine of Leiden University Medical Centre (April 1997 to April 2002), the Department of Obstetrics and Gynaecology of the Medical Centre Haaglanden, The Hague (April 1997 to April 2002) or the Department of Obstetrics and Gynaecology of the Groene Hart Hospital, Gouda, The Netherlands (April 1999 to April 2000) were eligible for inclusion in the trial.

Exclusion criteria were subfertility less than 1 year, woman older than 37 years at the time of first visit, anovulation despite clomiphene citrate or bromocriptine use, abnormal semen analysis according to World Health Organization (WHO) (World Health Organization, 1999) criteria or testing of tubal patency performed in the past. The institutional review boards of each of the three hospitals approved the study protocol. Women were asked to participate in the study by their treating gynaecologist at the time that HSG would normally be planned, and informed consent was obtained. The treating gynaecologist telephoned the secretariat of Medical Centre Haaglanden at The Hague to perform randomization. A computer-based 1 : 1 ratio randomization procedure was used to allocate the women into two groups. Randomization was stratified for each participating hospital. All women routinely received vaginal ultrasound before randomization. The intervention group underwent HSG first, and if the HSG showed normal uterine cavity and no tubal pathology and if the woman did not conceive within 6 months, a laparoscopy and dye followed after 6 months. When tubal pathology was assumed, laparoscopy was performed within 1–2 months after the HSG. The control group received a laparoscopy and dye immediately. If pathology of the uterine cavity was presumed by HSG or by vaginal ultrasound, hysteroscopy could be performed together with the laparoscopy. Moreover, a history of recurrent miscarriages or diethylstilboestrol (DES) exposure was an additional reason to perform a hysteroscopy during laparoscopy.

Because our trial was designed to determine the effectiveness of HSG in the routine fertility workup, we ensured that HSG and laparoscopy results were uniformly interpreted in all participating hospitals. At the same time, the study protocol intentionally allowed normal clinical freedom and a variety of choices and protocols after HSG and laparoscopy. Hence, the participating hospitals used their own protocol for therapeutic reproductive surgery and assisted reproductive treatments [e.g. intrauterine insemination (IUI) or IVF]. The primary analysis was conducted on an intention-to-treat basis. The primary outcome parameter in our study was occurrence of pregnancy within 18 months after randomization. The diagnosis of pregnancy was based on a positive urine or serum pregnancy test in association with the presence of an intrauterine gestation sac on ultrasound scan.

### *HSG and laparoscopy and dye*

All hysterosalpingographies were performed in the outpatient clinic of the department of radiology shortly after the menstrual period. A water-soluble contrast medium (Omnipaque 300®) was used. One photograph was taken of the phase when the cavity and tubes were just filled and one when there was overflow at both sides or when there was maximal filling of the tubes without overflow. After 30 min, a late film was made to detect contrast depots. Findings of tubal pathology at HSG were classified according to Mol *et al*. (2001), as normal, one-sided abnormality or two-sided abnormality. Additional intracavity abnormalities were scored separately. The results of HSG were interpreted in a weekly meeting by staff members specialized in reproductive medicine, who also decided whether laparoscopy and dye should be performed with or without delay.

Laparoscopy and dye was performed in the follicular phase and under general anaesthesia. After making pneumoperitoneum, a thorough inspection of the pelvis, internal genitalia, appendix and liver region was performed, followed by testing the patency of the Fallopian tubes using dye. A dilute solution of Methylene Blue was injected through the cervix. During laparoscopy, we determined adhesions, structural abnormalities of the uterus, endometriosis, periadnexal disease and Fallopian tube occlusion. Tubal pathology at laparoscopy was defined according to Mol *et al*. (2001), as normal, one-sided abnormality or two-sided abnormality. Furthermore, endometriosis detected at laparoscopy was classified according to the classification of the American Fertility Society (1985). Therapeutic reproductive surgery could be applied during laparoscopy, such as coagulation of endometriosis grade I/II, laparoscopic adhesiolysis or laparoscopic cystectomy.

### *Statistical methods*

Descriptive statistics were used to assess the similarity of the groups. Categorical data were assessed by the chi-square test and continuous variables by Student's $t$-test. CPRs were calculated using standard time-to-event analysis (Kaplan–Meier survival analysis). For comparison of the different CPR curves, the log-rank statistic was used. On the basis of local unpublished data of Leiden University Medical Centre, we calculated that for a subfertile couple the probability of getting pregnant after 1 year from intake, including artificial interference, is about 45%. With a smallest difference in CPR arbitrarily set at 10% (55% in the intervention group and 45% in the control group), an alpha error of 0.05 and a beta error of 0.20 (power of the study set at 80%), we calculated that at least 375 women should be included in each arm (a total of 750 women).

## Results

A total of 344 women were randomized, 169 to the intervention group and 175 to the control group. Follow-up either to pregnancy or for 18 months was complete for all subjects in both groups. Figure 1 shows the flow chart of participants. At the end of the study, HSG had been performed in 152 of the 169 (90%) women in the intervention group. In the control group, 10 of the 175 (6%) women had undergone an HSG. Laparoscopies had been performed on 94 of the 169 (56%) women in the intervention group and on 150 of the 175 (86%) women in the control group. To deal with this, our analysis was based on the groups as randomized, following the intention-to-treat principle.

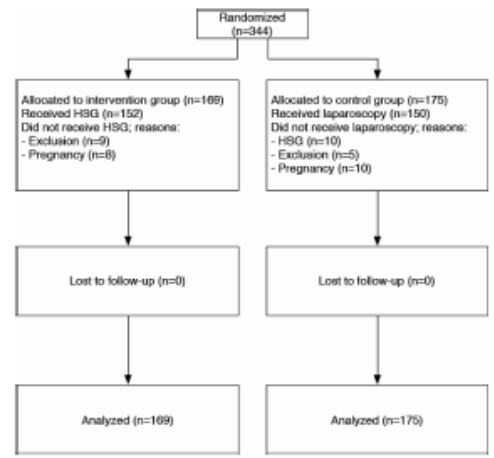
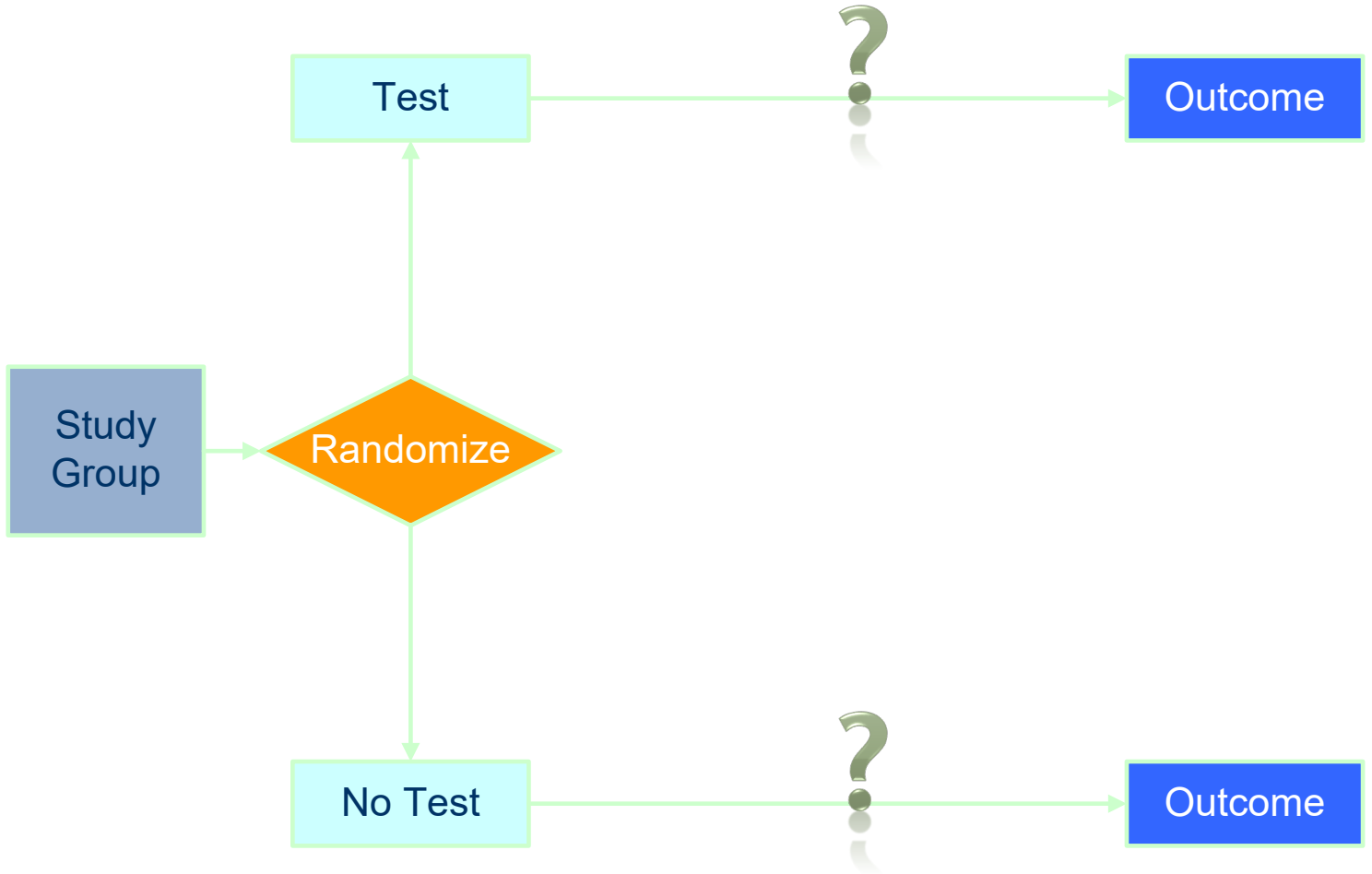
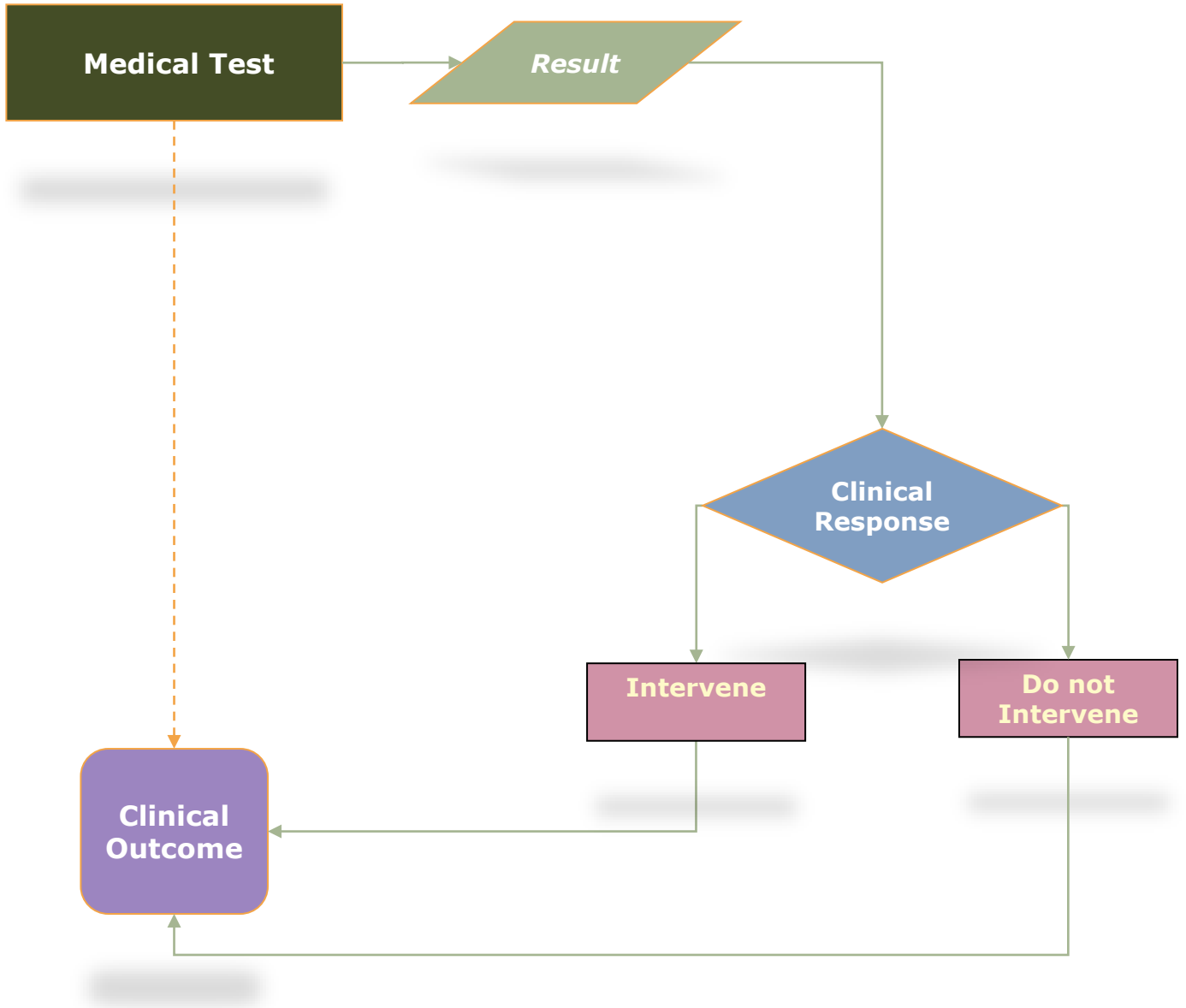



**Figure 1.** Flow chart of participants.

# RCT Medical Test






P.M. Bossuyt and K. McCaffery, MDM, 2009

# Hysteroscopy before in-vitro fertilisation (inSIGHT): a multicentre, randomised controlled trial

*Janine G Smit, Jenneke C Kasius, Marinus J C Eijkemans, Carolien A M Koks, Ronald van Golde, Annemiek W Nap, Gabrielle J Scheffer, Petra A P Manger, Annemieke Hoek, Benedictus C Schoot, Arne M van Heusden, Walter K H Kuchenbecker, Denise A M Perquin, Kathrin Fleischer, Eugenie M Kaaijk, Alexander Sluijmer, Jaap Friederich, Ramon H M Dykgraaf, Marcel van Hooff, Leonie A Louwe, Janet Kwee, Corry H de Koning, Ineke C A H Janssen, Femke Mol, Ben W J Mol, Frank J M Broekmans, Helen L Torrance*

*Lancet* 2016; 387: 2622–29

Published Online April 27, 2016 http://dx.doi.org/10.1016/S0140-6736(16)00231-2

See Comment page 2578

See Articles page 2614

**Department of Reproductive Medicine and Gynaecology** (J G Smit MD, J C Kasius PhD, Prof F J M Broekmans PhD, H L Torrance PhD) **and Julius Center for Health Sciences and Primary Care** (Prof M J C Eijkemans PhD), **University Medical Center Utrecht, Utrecht, Netherlands; Maxima Medical Center, Veldhoven, Netherlands** (C A M Koks PhD); **Maastricht University Medical Center, Maastricht, Netherlands** (R van Golde PhD); **Rijnstate Hospital, Arnhem, Netherlands** (A W Nap PhD); **Gelre Hospital, Apeldoorn, Netherlands** (G J Scheffer PhD); **Diakonessen Hospital Utrecht, Utrecht, Netherlands**

## Summary

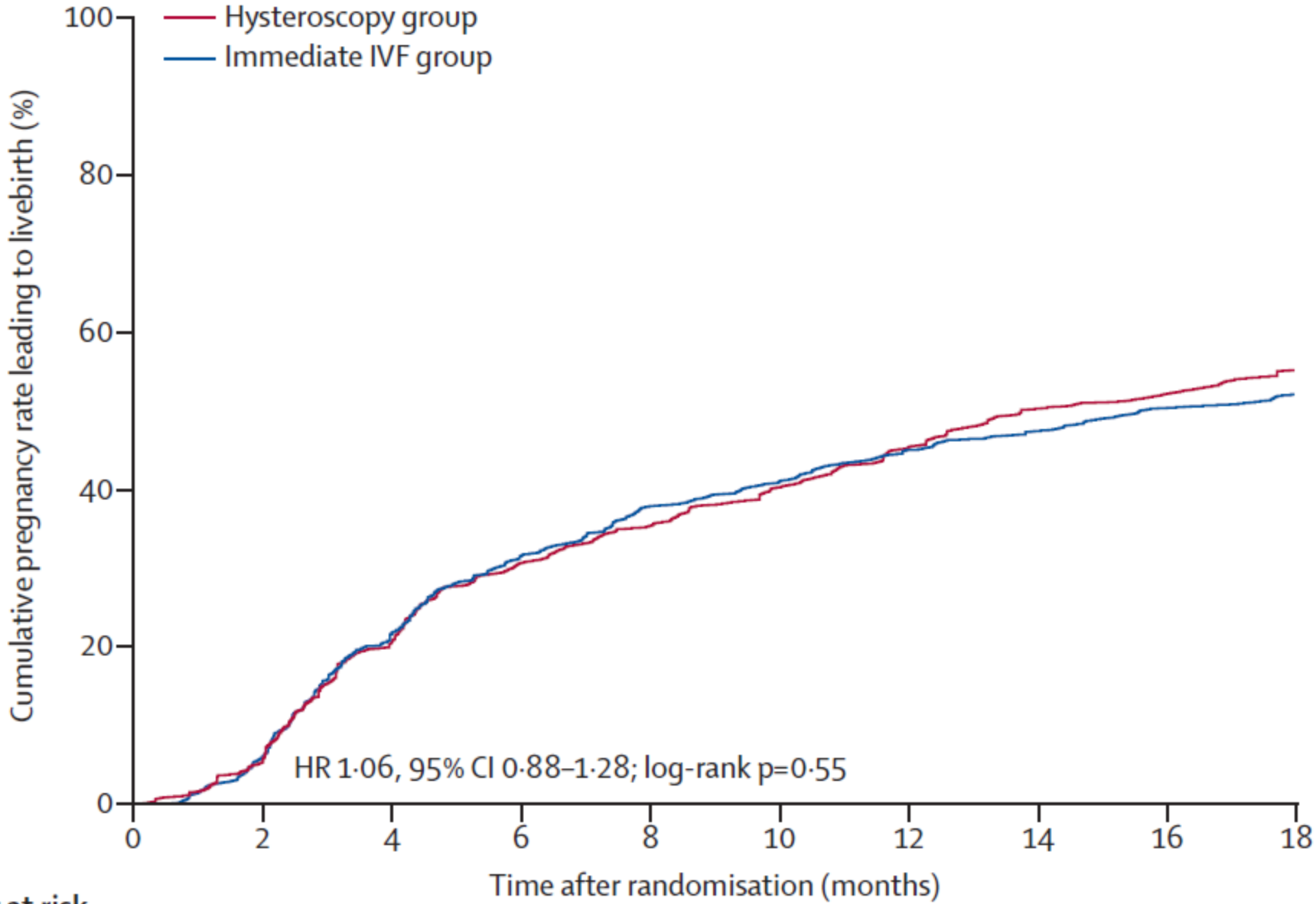
**Background** Hysteroscopy is often done in infertile women starting in-vitro fertilisation (IVF) to improve their chance of having a baby. However, no data are available from randomised controlled trials to support this practice. We aimed to assess whether routine hysteroscopy before the first IVF treatment cycle increases the rate of livebirths.

**Methods** We did a pragmatic, multicentre, randomised controlled trial in seven university hospitals and 15 large general hospitals in the Netherlands. Women with a normal transvaginal ultrasound of the uterine cavity and no previous hysteroscopy who were scheduled for their first IVF treatment were randomly assigned (1:1) to either hysteroscopy with treatment of detected intracavitary abnormalities before starting IVF (hysteroscopy group) or immediate start of the IVF treatment (immediate IVF group). Randomisation was done with web-based concealed allocation and was stratified by centre with variable block sizes. Participants, doctors, and outcome assessors were not masked to the assigned group. The primary outcome was ongoing pregnancy (detection of a fetal heartbeat at >12 weeks of gestation) within 18 months of randomisation and resulting in livebirth. Analysis was by intention to treat. This trial is registered with ClinicalTrials.gov, number NCT01242852.

**Findings** Between May 25, 2011, and Aug 27, 2013, we randomly assigned 750 women to receive either hysteroscopy (n=373) or immediate IVF (n=377). 209 (57%) of 369 women eligible for assessment in the hysteroscopy group and 200 (54%) of 373 in the immediate IVF group had a livebirth from a pregnancy during the trial period (relative risk 1·06, 95% CI 0·93–1·20; p=0·41). One (<1%) woman in the hysteroscopy group developed endometritis after hysteroscopy.

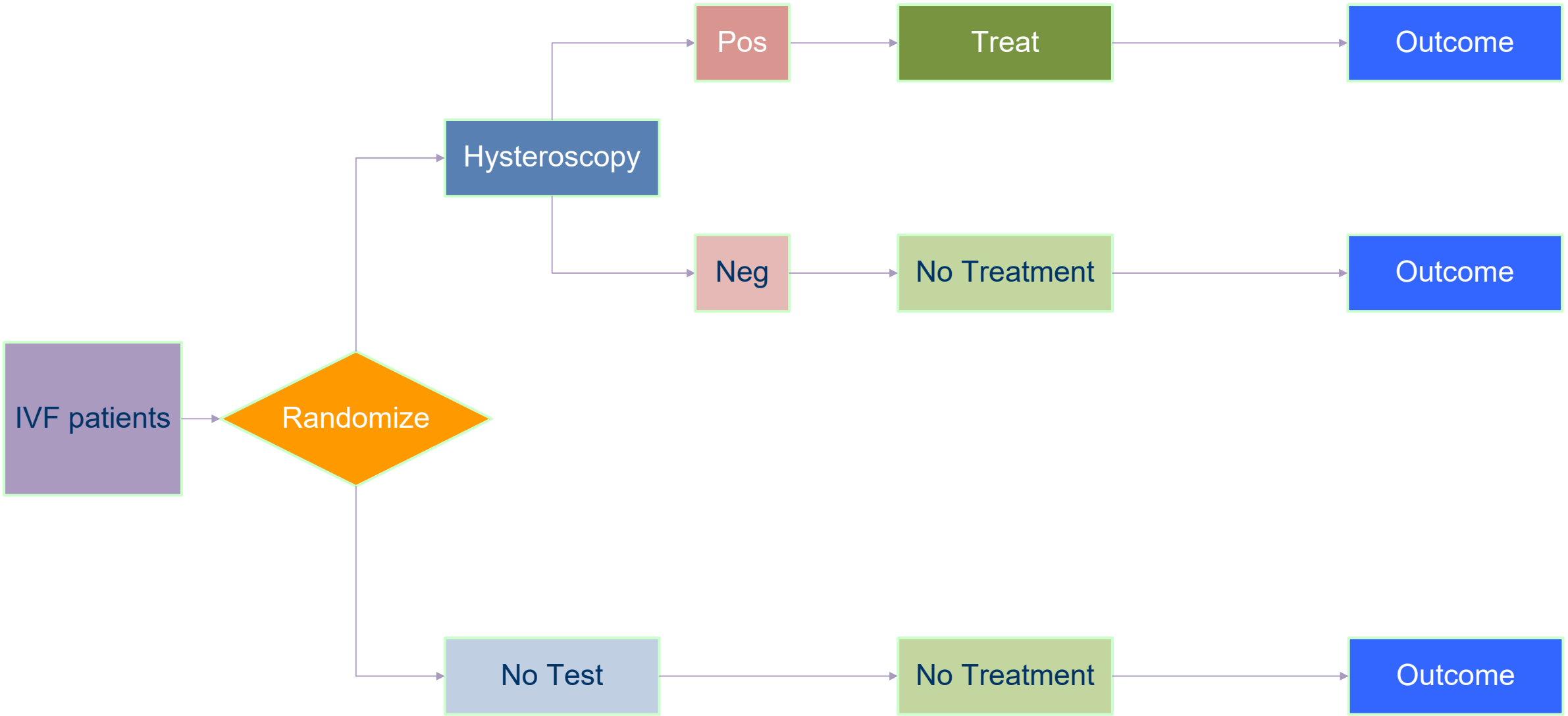
**Interpretation** Routine hysteroscopy does not improve livebirth rates in infertile women with a normal transvaginal ultrasound of the uterine cavity scheduled for a first IVF treatment. Women with a normal transvaginal ultrasound should not be offered routine hysteroscopy.

**Funding**: The Dutch Organisation for Health Research and Development (ZonMW).




**Number at risk**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Hysteroscopy group | 369 | 347 | 291 | 254 | 237 | 219 | 200 | 182 | 175 | 160 |
| Immediate IVF group | 373 | 350 | 291 | 254 | 232 | 219 | 204 | 195 | 184 | 173 |


IVF patients
Randomize
Hysteroscopy
Pos
Treat
Outcome
Neg
No Treatment
Outcome
No Test
No Treatment
Outcome

## Statistical analysis

To calculate the sample size needed we assumed that, compared with immediate IVF, hysteroscopy would increase the chance of a livebirth from 30% to 40%. To detect this difference, we needed to include 350 women per group (700 women overall) to provide 80% power at α 5%. Anticipating that 5% of the women in the intervention group would not undergo hysteroscopy, we established that the final sample size needed to be 370 women per study group (740 women overall).

THE LANCET

## Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

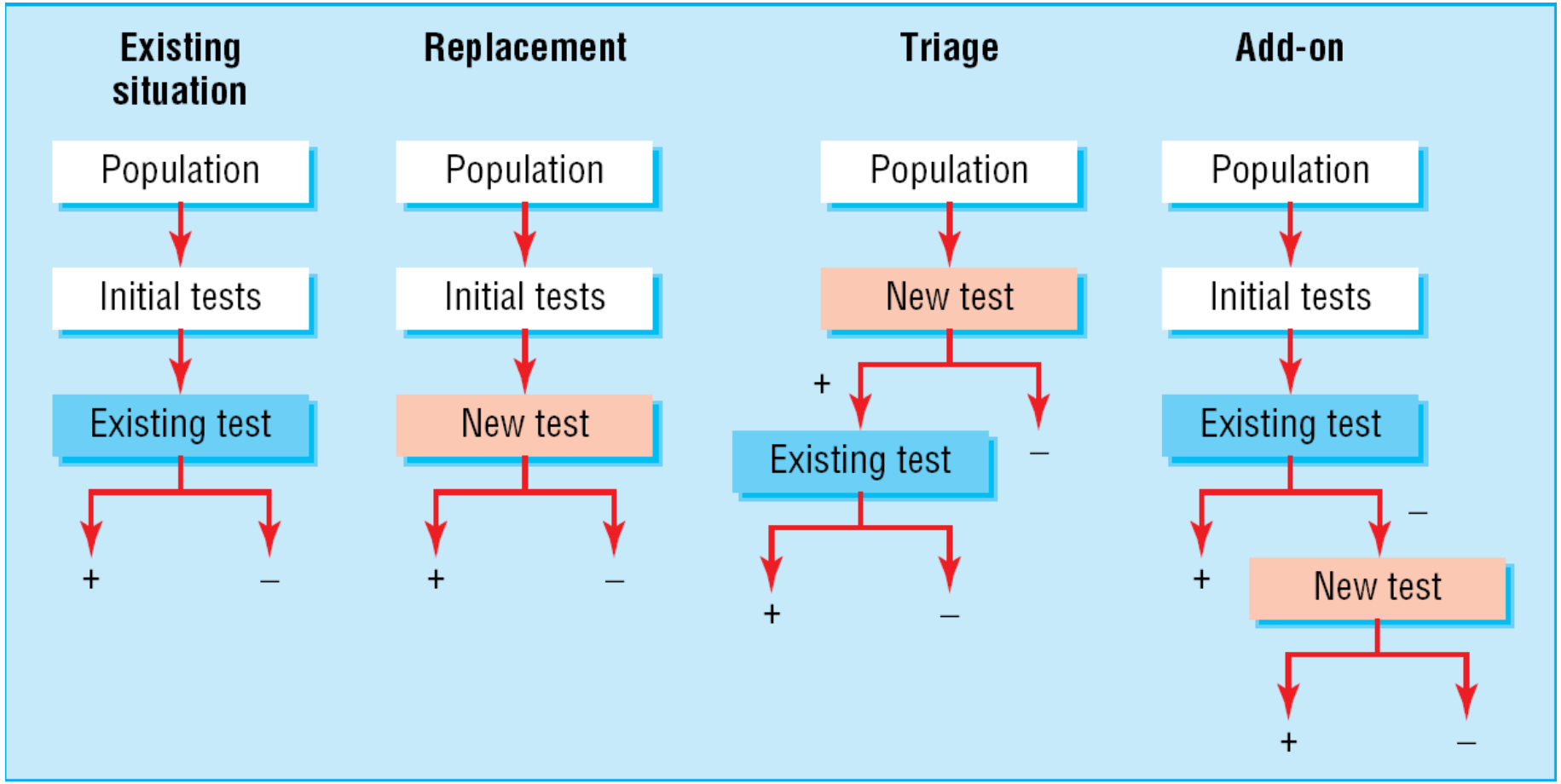
Supplement to: Smit JG, Kasius JC, Eijkemans MJC, et al. Hysteroscopy before in-vitro fertilisation (inSIGHT): a multicentre, randomised controlled trial. *Lancet* 2016; published online April 27. http://dx.doi.org/10.1016/S0140-6736(16)00231-2.

| Treatment | Hysteroscopy | Immediate IVF |
|---|---|---|
| **Hysteroscopy*** | | |
| Number of include women | 369 | 373 |
| No hysteroscopy performed (see also figure 1) | 44 (12%) | 364 (98%) |
| Total number of hysteroscopies performed | 325 (88%) | 9 (2·4%) |
| Failed procedures | 29 (8·9%) | 0 (0%) |
| Completed procedures | 296 (91%) | 9 (100%) |
| Women with intracavitary abnormalities | 37 (13%) | 3 (33%) |
| Treated abnormalities | | |
| Polyps | 25 (68%) | 2 (67%) |
| Septate uterus | 1 (2·7%) | 0 |
| Adhesions | 3 (8·1%) | 0 |
| Myoma | 2 (5·4%) | 0 |
| Untreated abormalities | | |
| Polyp | 1 (2·7%) | 0 |
| Septate uterus | 5 (14%) | |
| Myoma | 2 (5·4%) | 0 |
| Untreatable abnormalities | | |
| Abnormal shape of tubal orifice | 2 (5·4%) | 1 (33%) |
| Bicornuate uterus | 1 (2·7%) | 0 |
| Polypoid endometrium | 1 (2·7%) | 0 |
| | | |
| **IVF/ICSI treatment cycles**** | | |
| **Fresh cycles** | | |
| Total number of fresh IVF/ICSI cycles | 707 | 692 |
| first cycle | 348 (49%) | 349 (50%) |
| second cycle | 214 (30%) | 205 (30%) |
| third cycle | 111 (16%) | 112 (16%) |
| fourth cycle | 26 (3·7%) | 21 (3·0%) |
| fifth cycle | 7 (1·0%) | 4 (0·6%) |
| sixth cycle | 1 (0·1%) | 1 (0·1%) |
| Downregulation protocol*** | | |
| GnRH-agonist | 562 (80%) | 577 (83%) |
| GnRH-antagonist | 127 (18%) | 96 (14%) |
| Mean (SD) time to start treatment - days**** | 67 (54) | 61 (46) |
| Mean (SD) starting dose gonadotrophins | 181 (75) | 186 (86) |
| Mean (SD) duration of stimulation - days | 12 (3·4) | 12 (3·6) |
| Cancelled cycles (including escape IUI) | 83 (12%) | 87 (13%) |
| Number of ovum pick-ups | 624 (88%) | 605 (87%) |
| Mean (SD) number of oocytes | 9·0 (5·3) | 8·6 (5·2) |
| Mean (SD) number of embryos | 4·3 (3·5) | 4·3 (3·6) |
| Number of embryo transfers | 554 (78%) | 535 (77%) |
| Mean (SD) number of embryos transferred | 1·3 (0·5) | 1·3 (0·5) |

| Treatment | Hysteroscopy | Immediate IVF |
|---|---|---|
| **Hysteroscopy*** | | |
| Number of include women | 369 | 373 |
| No hysteroscopy performed (see also figure 1) | 44 (12%) | 364 (98%) |
| Total number of hysteroscopies performed | 325 (88%) | 9 (2·4%) |
| Failed procedures | 29 (8·9%) | 0 (0%) |
| Completed procedures | 296 (91%) | 9 (100%) |
| Women with intracavitary abnormalities | 37 (13%) | 3 (33%) |
| Treated abnormalities | | |
| Polyps | 25 (68%) | 2 (67%) |
| Septate uterus | 1 (2·7%) | 0 |
| Adhesions | 3 (8·1%) | 0 |
| Myoma | 2 (5·4%) | 0 |
| Untreated abormalities | | |
| Polyp | 1 (2·7%) | 0 |
| Septate uterus | 5 (14%) | |
| Myoma | 2 (5·4%) | 0 |
| Untreatable abnormalities | | |
| Abnormal shape of tubal orifice | 2 (5·4%) | 1 (33%) |
| Bicornuate uterus | 1 (2·7%) | 0 |
| Polypoid endometrium | 1 (2·7%) | 0 |

# 3.
# Imaging RCT: Efficient Designs

# Roles of new test




Bossuyt et al; BMJ 2006

# Replacement: More efficient design

**Human Reproduction, Vol.37, No.5, pp. 969–979, 2022**
Advance Access Publication on February 27, 2022 https://doi.org/10.1093/humrep/deac034

human reproduction

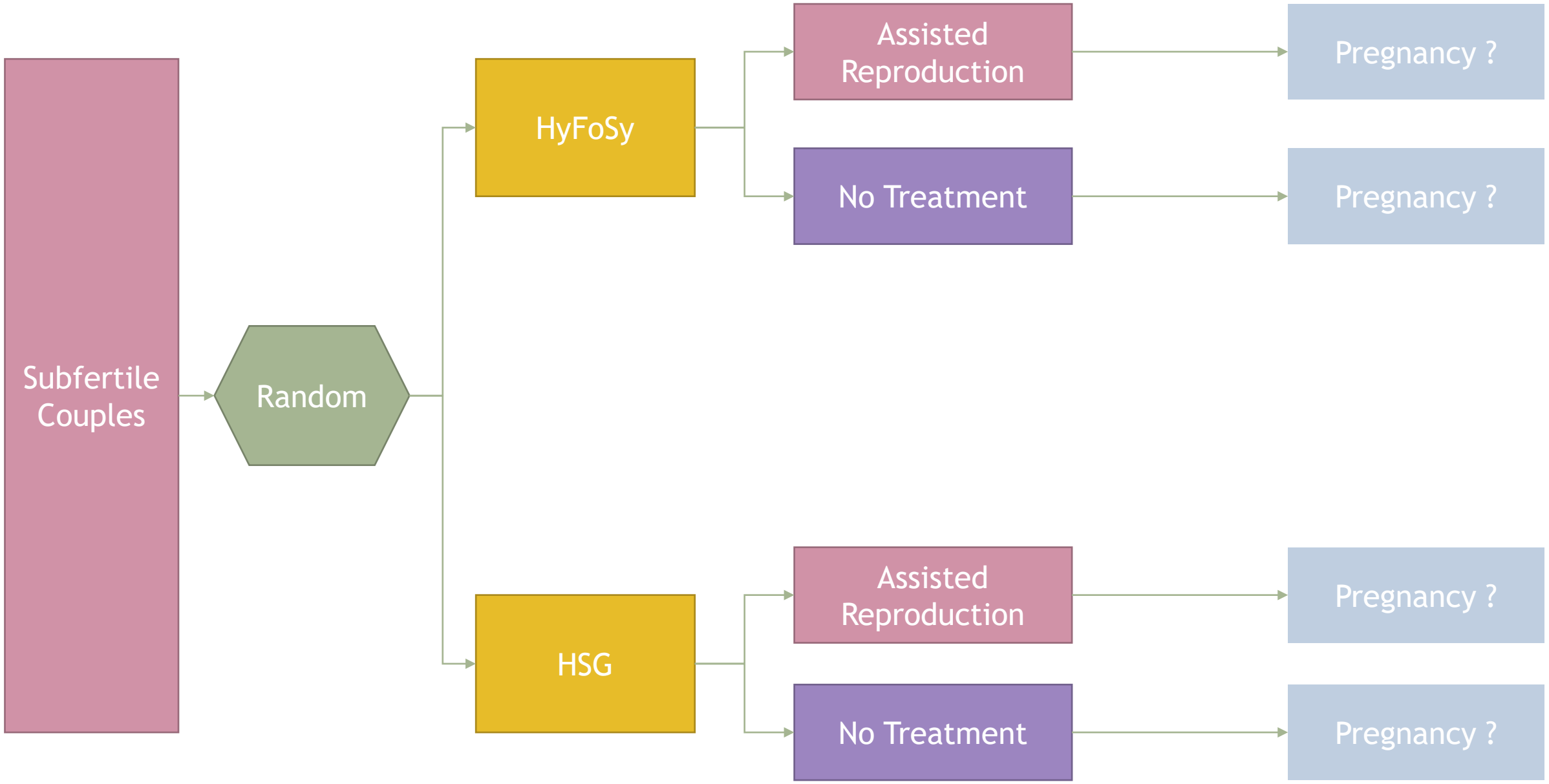
ORIGINAL ARTICLE *Infertility*

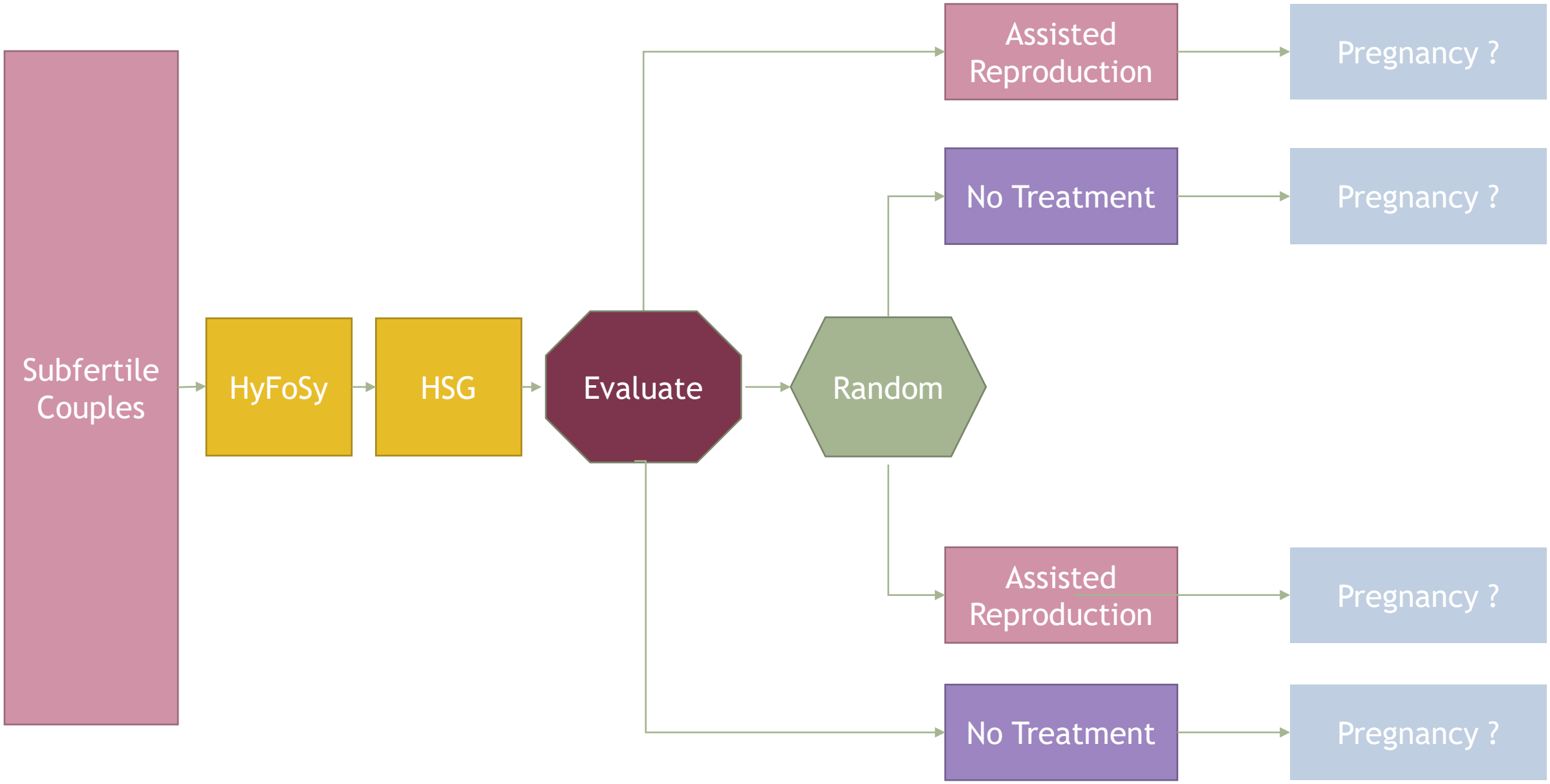
## Can hysterosalpingo-foam sonography replace hysterosalpingography as first-choice tubal patency test? A randomized non-inferiority trial

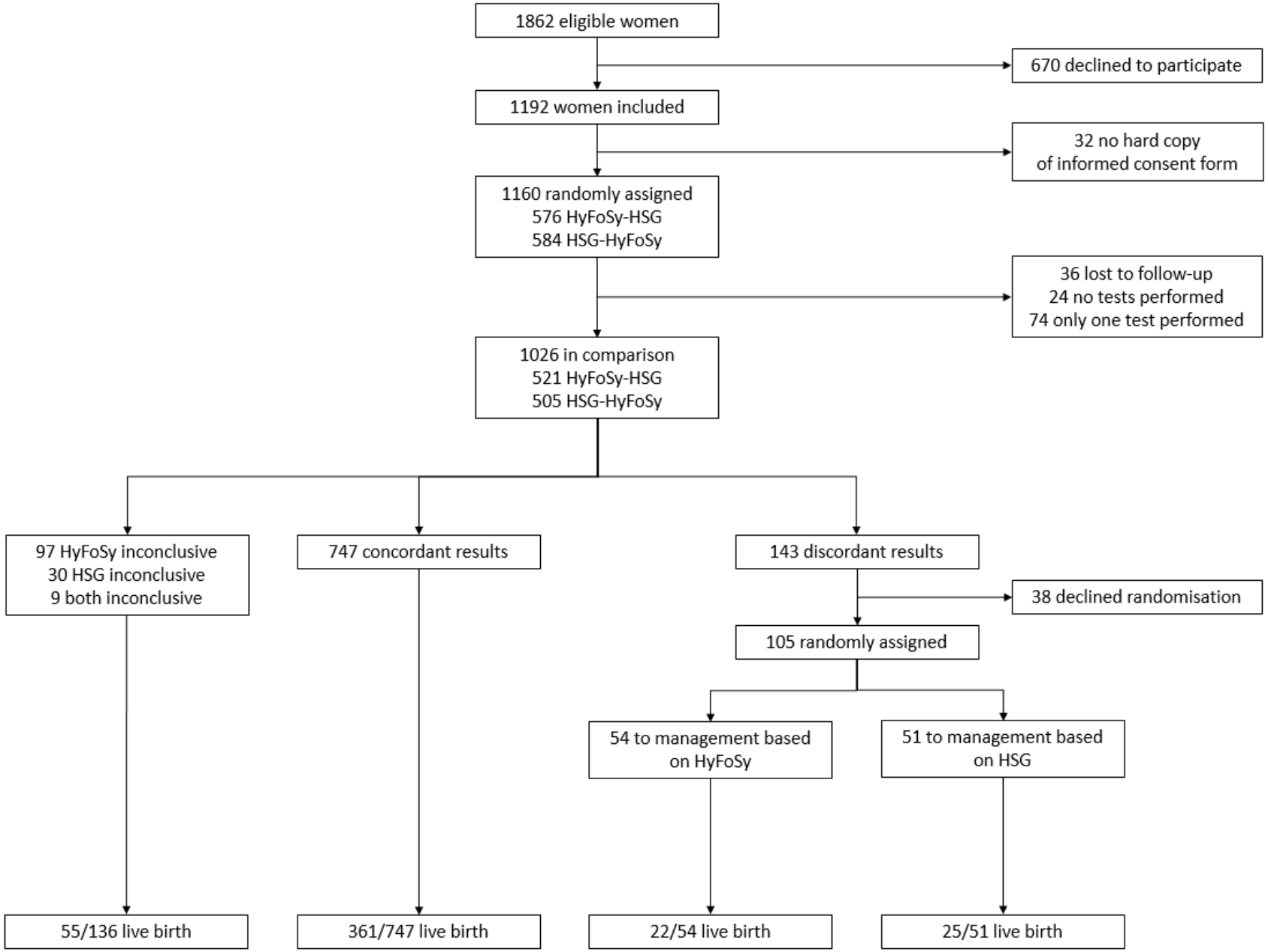
Nienke van Welie [1,*], Joukje van Rijswijk [1], Kim Dreyer[1], Machiel H.A. van Hooff[2], Jan Peter de Bruin[3], Harold R. Verhoeve[4], Femke Mol[5], Wilhelmina M. van Baal[6], Maaike A.F. Traas[7], Arno M. van Peperstraten[8,9], Arentje P. Manger[10], Judith Gianotten[11], Cornelia H. de Koning[12], Aafke M.H. Koning[13], Neriman Bayram[14], David P. van der Ham[15], Francisca P.J.M. Vrouenraets[16], Michaela Kalafusova[17], Bob I.G. van de Laar[18], Jeroen Kaijser[19], Arjon F. Lambeek[20], Wouter J. Meijer[21], Frank J.M. Broekmans[9], Olivier Valkenburg[22], Lucy F. van der Voet[23], Jeroen van Disseldorp[24], Marieke J. Lambers[25], Rachel Tros[26], Cornelis B. Lambalk [1], Jaap Stoker[27], Madelon van Wely[5,28], Patrick M.M. Bossuyt [28], Ben Willem J. Mol[29,30], and Velja Mijatovic[1]

# FOAM Study

- P: Subfertile Couples
- I: HyFoSy
- C: HSG
- T: Pregnancy at 12 months


Subfertile Couples
Random
HyFoSy
HSG
Assisted Reproduction
No Treatment
Assisted Reproduction
No Treatment
Pregnancy ?
Pregnancy ?
Pregnancy ?
Pregnancy ?


Subfertile Couples
HyFoSy
HSG
Evaluate
Random
Assisted Reproduction
No Treatment
Assisted Reproduction
No Treatment
Pregnancy ?
Pregnancy ?
Pregnancy ?
Pregnancy ?


1862 eligible women
670 declined to participate
1192 women included
32 no hard copy of informed consent form
1160 randomly assigned
576 HyFoSy-HSG
584 HSG-HyFoSy
36 lost to follow-up
24 no tests performed
74 only one test performed
1026 in comparison
521 HyFoSy-HSG
505 HSG-HyFoSy
97 HyFoSy inconclusive
30 HSG inconclusive
9 both inconclusive
747 concordant results
143 discordant results
38 declined randomisation
105 randomly assigned
54 to management based on HyFoSy
51 to management based on HSG
55/136 live birth
361/747 live birth
22/54 live birth
25/51 live birth

# Concordance HyFoSy/HSG

**Table II** **Comparison between hysterosalpingo-foam sonography (HyFoSy) result and hysterosalpingography (HSG) result (n = 1026).**

| HyFoSy \ HSG | Normal | One-sided tubal pathology | Double-sided tubal pathology | Inconclusive | Total |
|---|---|---|---|---|---|
| Normal | 702 (68%) | 52 (5%) | 10 (1%) | *27 (3%)* | 791 (77%) |
| One-sided tubal pathology | 46 (4%) | 35 (3%) | 7 (1%) | *2 (0%)* | 90 (9%) |
| Double-sided tubal pathology | 19 (2%) | 9 (1%) | 10 (1%) | *1 (0%)* | 39 (4%) |
| Inconclusive | *88 (9%)* | *8 (1%)* | *1 (0%)* | *9 (1%)* | 106 (10%) |
| Total | 855 (83%) | 104 (10%) | 28 (3%) | 39 (4%) | 1,026 (100%) |

The completed tests are indicated by the dashed line. Concordance between HyFoSy and HSG is shown in the diagonal blue boxes; discordance between HyFoSy and HSG is illustrated in red; inconclusive is illustrated in italic.

Human Reproduction, Vol.37, No.5, pp. 969–979, 2022

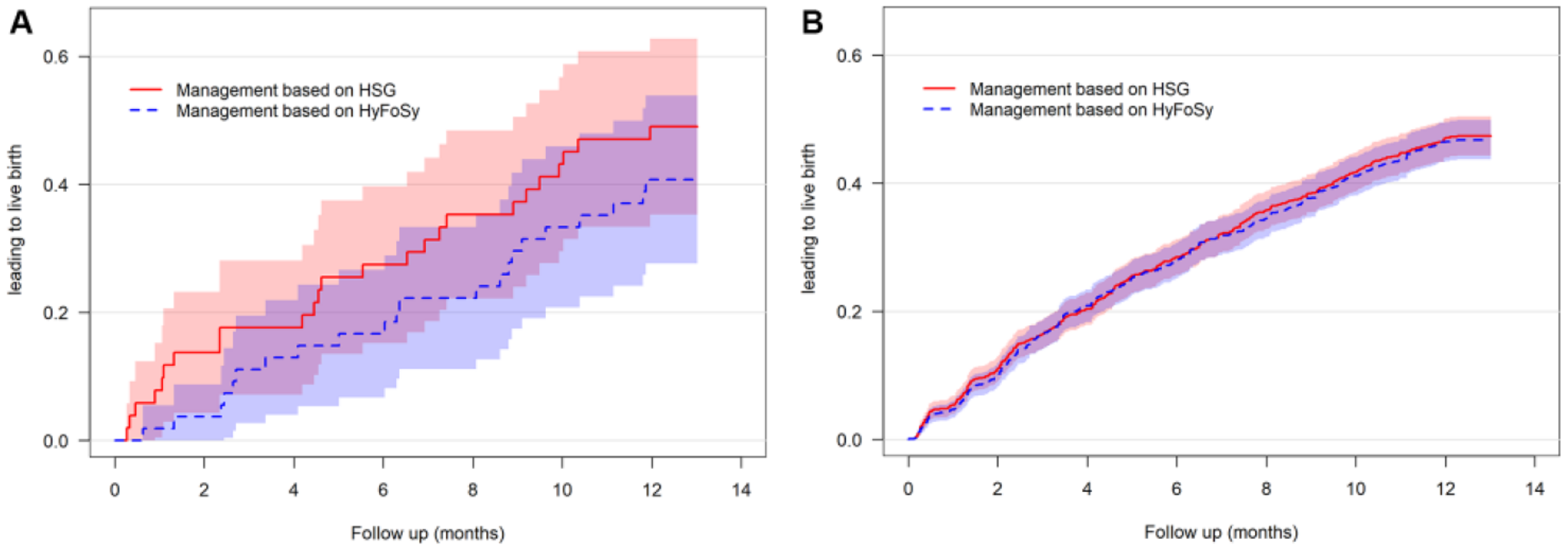



**Figure 3. Time to ongoing pregnancy leading to live birth for management based on hysterosalpingo-foam sonography (HyFoSy) compared to hysterosalpingography (HSG).** (**A**) Among discordant women (n = 105). (**B**) Among all women (N = 1026).

# FOAM Study

- P: Patients with peripancreatic carcinoma scheduled for surgery after radiologic staging
- I: Laparoscopic Staging
- C: No Laparoscopic Staging
- O: Hospital-free Survival

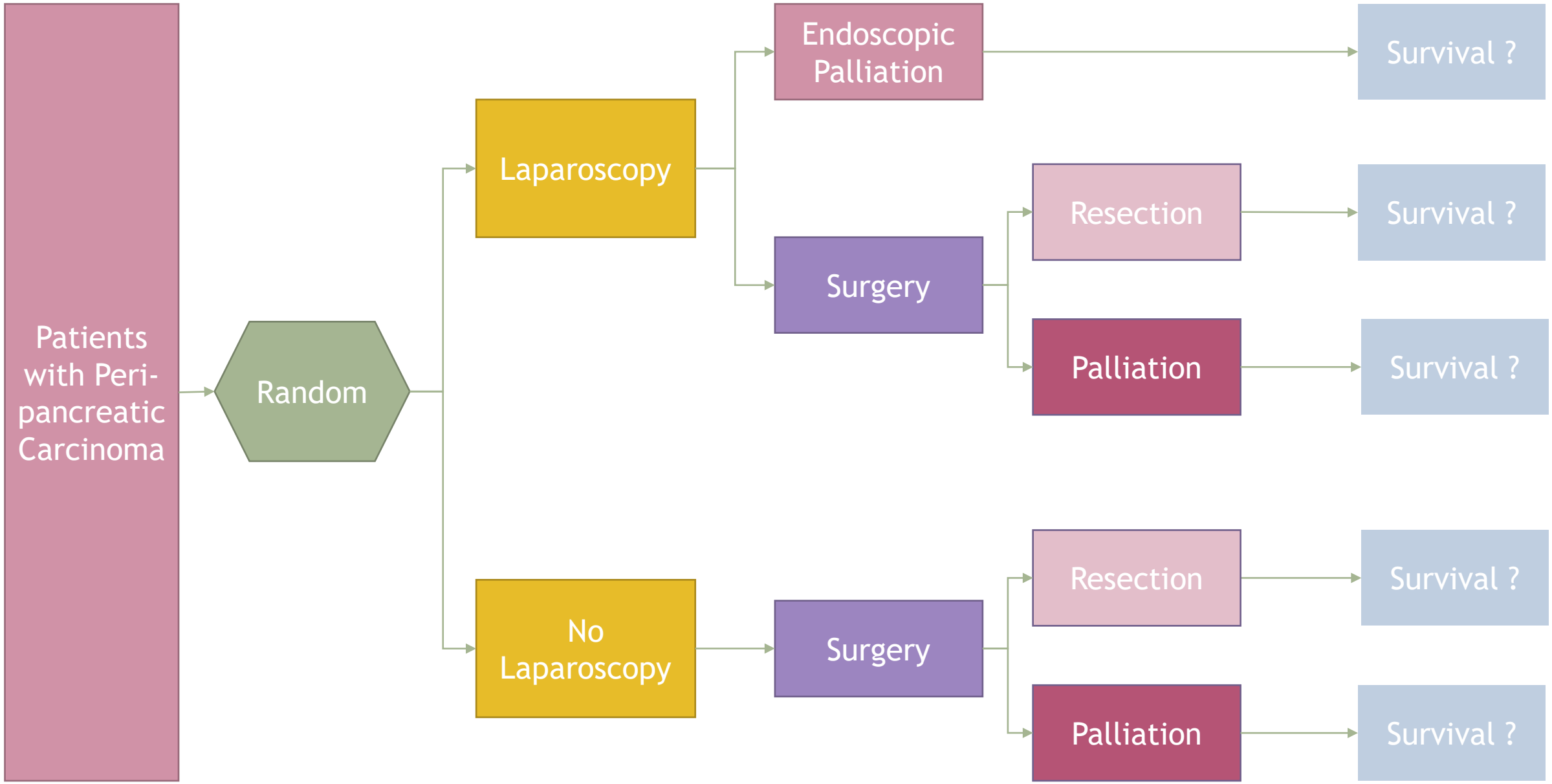
# Add-On: More efficient design

ANNALS OF SURGERY
Vol. 237, No. 1, 66–73


## Laparoscopic Staging and Subsequent Palliation in Patients With Peripancreatic Carcinoma

Els J. M. Nieveen van Dijkum, MD,* Mark G. Romijn, MD,§‖ Caroline B. Terwee, PhD,† Laurens Th. de Wit, MD,*
Jan H. P. van der Meulen, PhD,† Han S. Lameris, MD,§# Erik A. J. Rauws, MD,‡ Huug Obertop, MD,*
Casper H. J. van Eyck, MD,‖ Patrick M. M. Bossuyt, PhD,† and Dirk J. Gouma, MD*

*From the Departments of *Surgery, †Clinical Epidemiology and Biostatistics, ‡Gastroenterology and Hepatology, and §Radiology, Academic Medical Center, University of Amsterdam, The Netherlands; and the Departments of ‖General Surgery and #Radiology, Erasmus Medical Center, Erasmus University, Rotterdam, The Netherlands*


Patients with Peri-pancreatic Carcinoma
Random
Laparoscopy
No Laparoscopy
Endoscopic Palliation
Surgery
Surgery
Resection
Palliation
Resection
Palliation
Survival ?
Survival ?
Survival ?
Survival ?
Survival ?

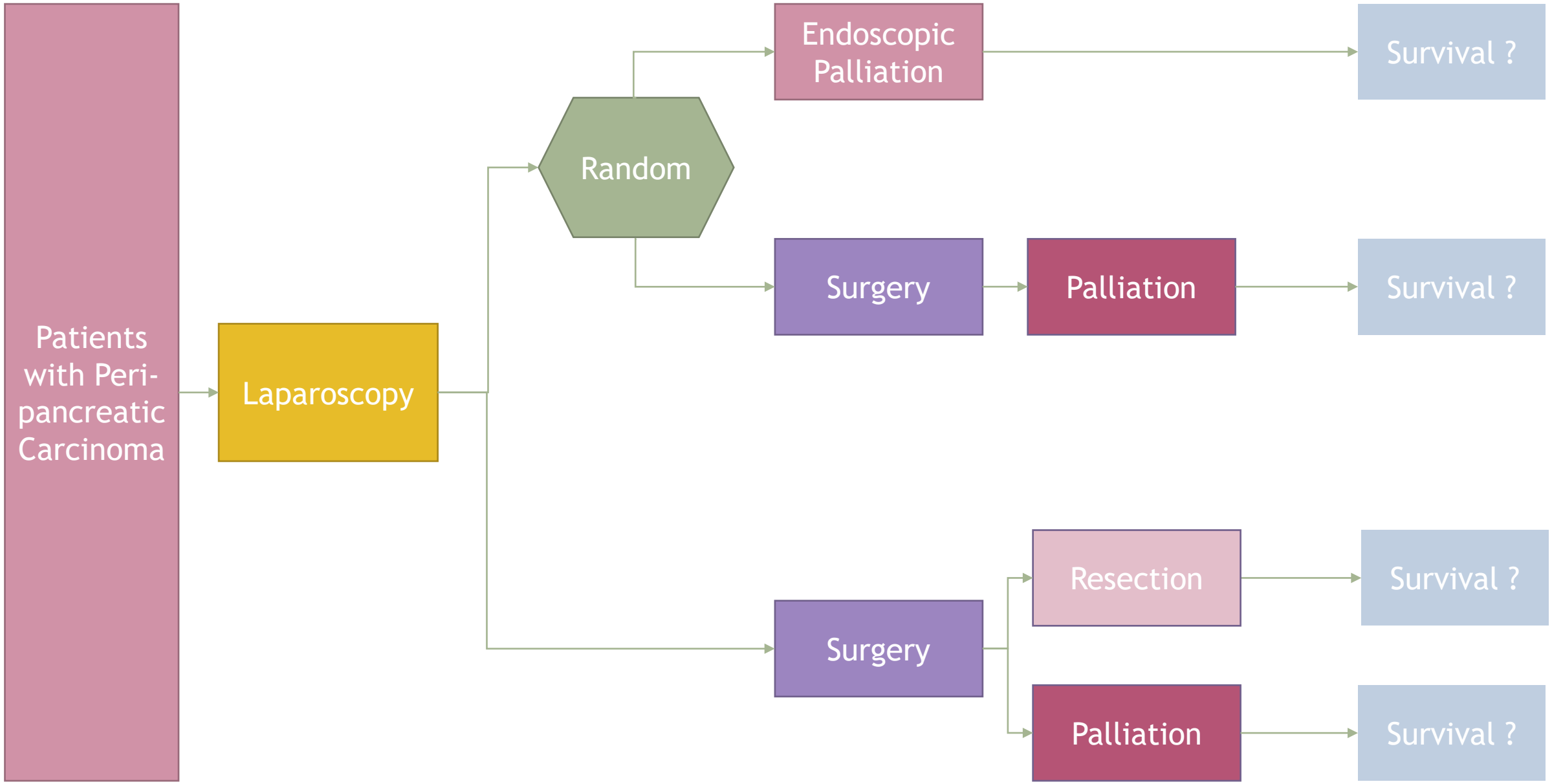
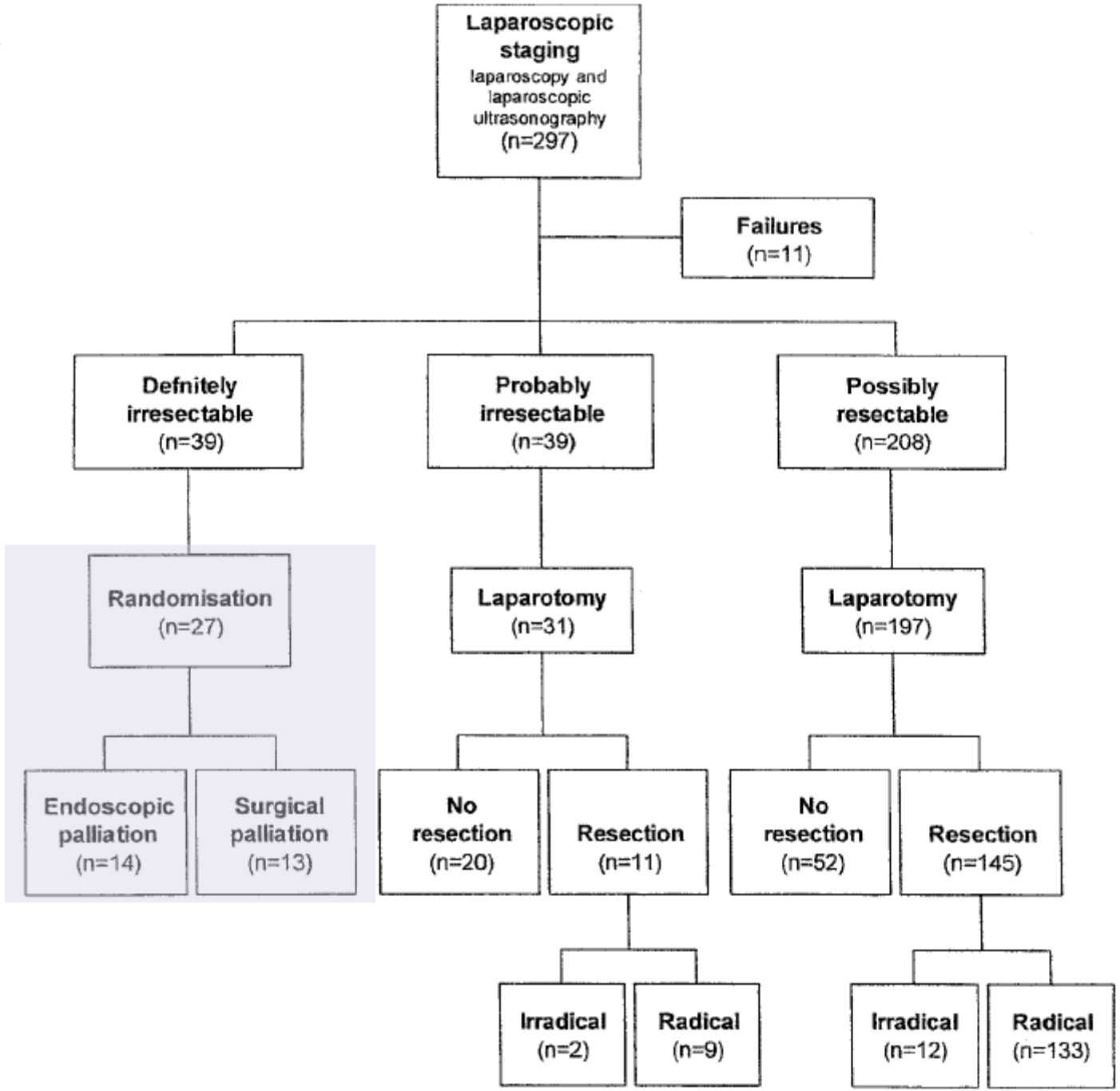
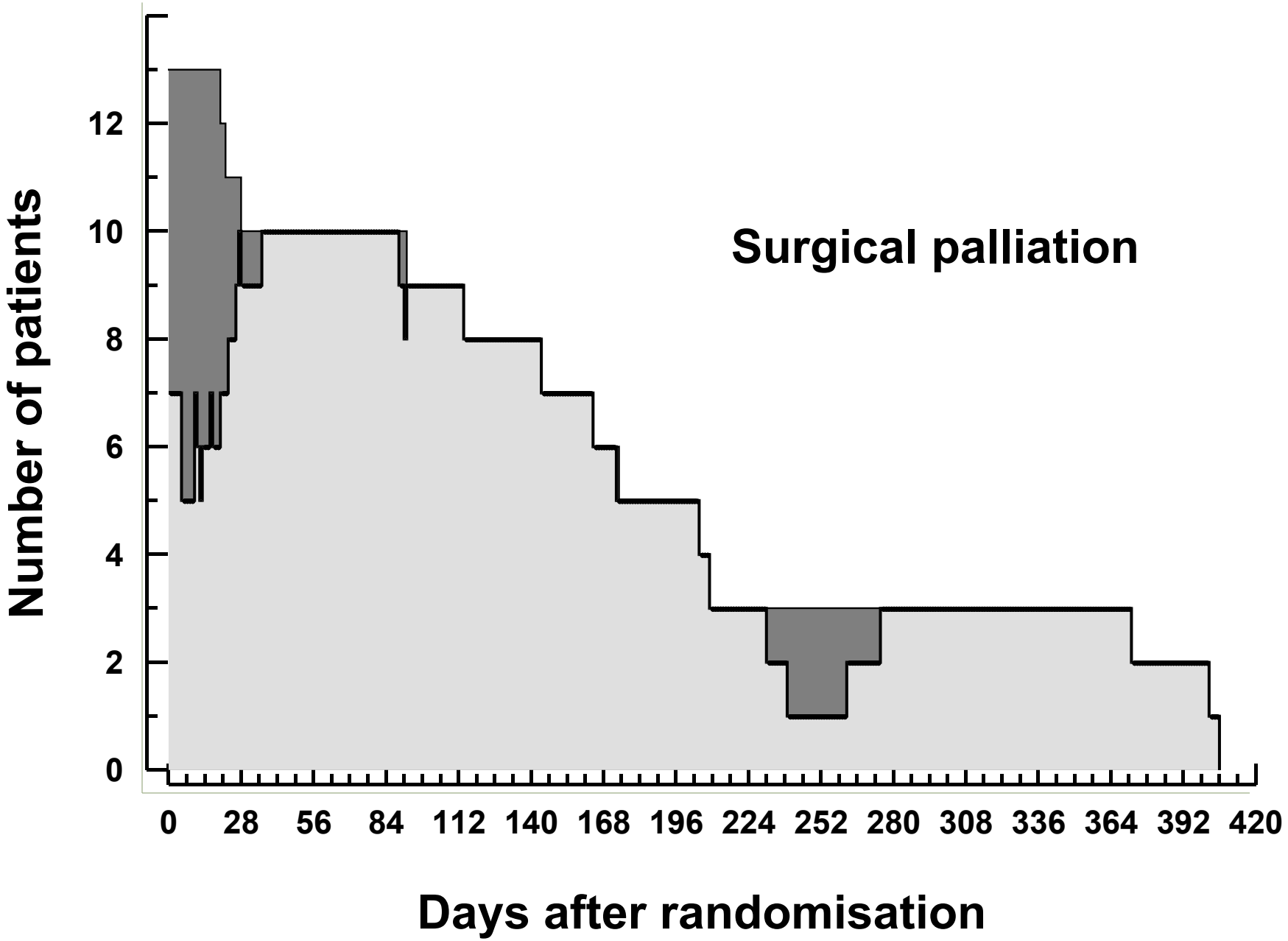
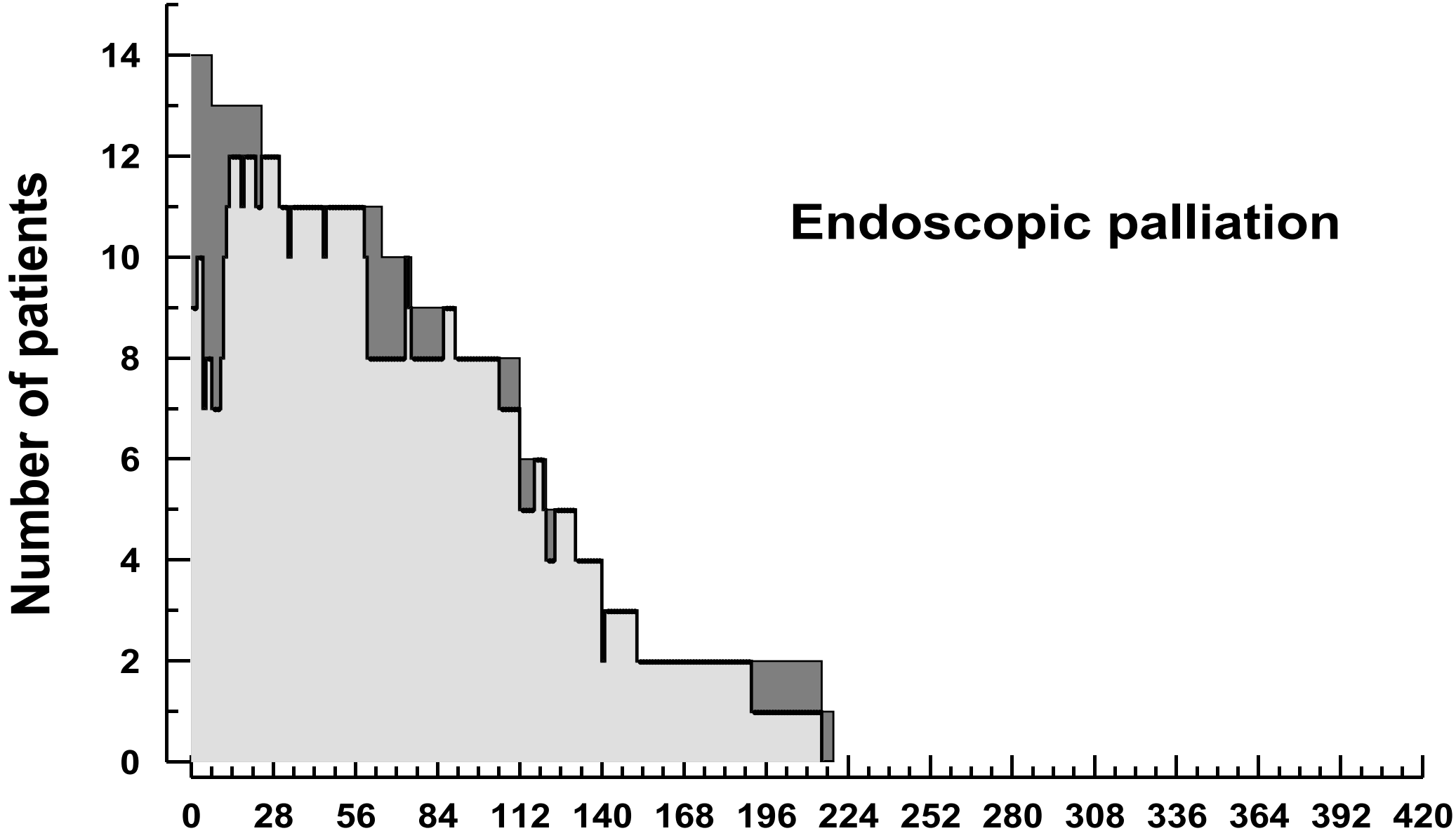

Patients with Peri-pancreatic Carcinoma
Laparoscopy
Random
Endoscopic Palliation
Survival ?
Surgery
Palliation
Survival ?
Surgery
Resection
Survival ?
Palliation
Survival ?

**Figure 1.** Flow chart of patients included for laparoscopic staging.




Surgical palliation
Number of patients
0
2
4
6
8
10
12
0
28
56
84
112
140
168
196
224
252
280
308
336
364
392
420
Days after randomisation


Endoscopic palliation
Number of patients
0
2
4
6
8
10
12
14
0
28
56
84
112
140
168
196
224
252
280
308
336
364
392
420

# 4.
# Avoidable Waste in Research

STARD reporting guidelines

researchwaste.net

Home | About | Events | Documents | News and Blog | Links

# Research

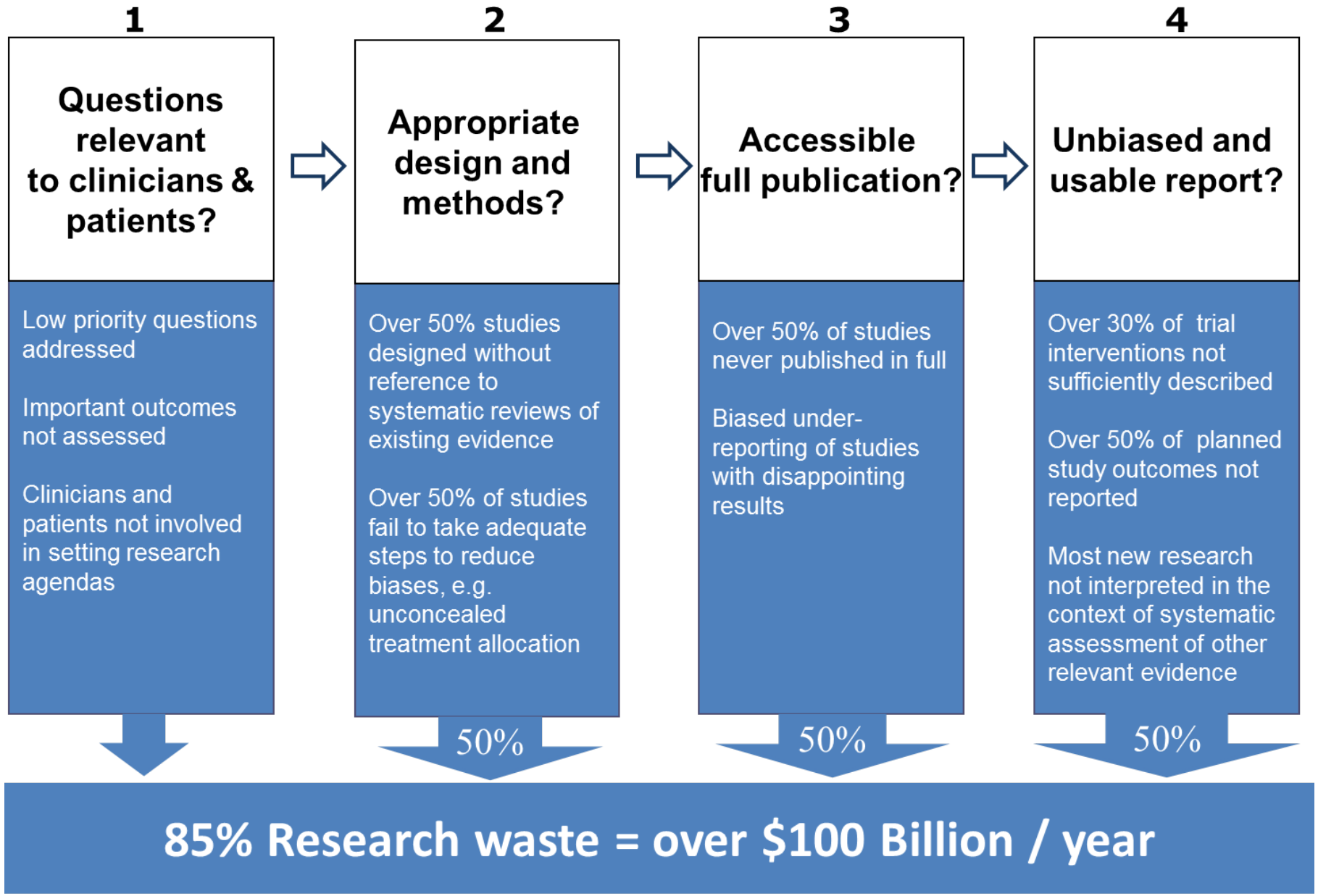
## Increasing value, reducing waste

It has been estimated that 85% of research is wasted, usually because it asks the wrong questions, is badly designed, not published or poorly reported. This diminishes the value of research and also represents a significant financial loss. However, many causes of this waste are simple problems that could easily be fixed, such as appropriate randomisation or blinding of a clinical trial. A first step towards increasing the value of research and increasing waste is to monitor the problems and develop solutions that aim to fix them.

**Access articles**


researchwaste.net is a place to share and exchange documentation, information, and resources on how to increase the value of both basic and applied research and reduce or avoid wasting research. It is based on a series of articles that were published in the medical journal *The Lancet* in 2014.

# Waste at four stages of research



Clinical Chemistry 60:4
651–659 (2014)



# Publication and Reporting of Test Accuracy Studies Registered in ClinicalTrials.gov

Daniël A. Korevaar,[1*] Eleanor A. Ochodo,[1] Patrick M.M. Bossuyt,[1] and Lotty Hooft[2]

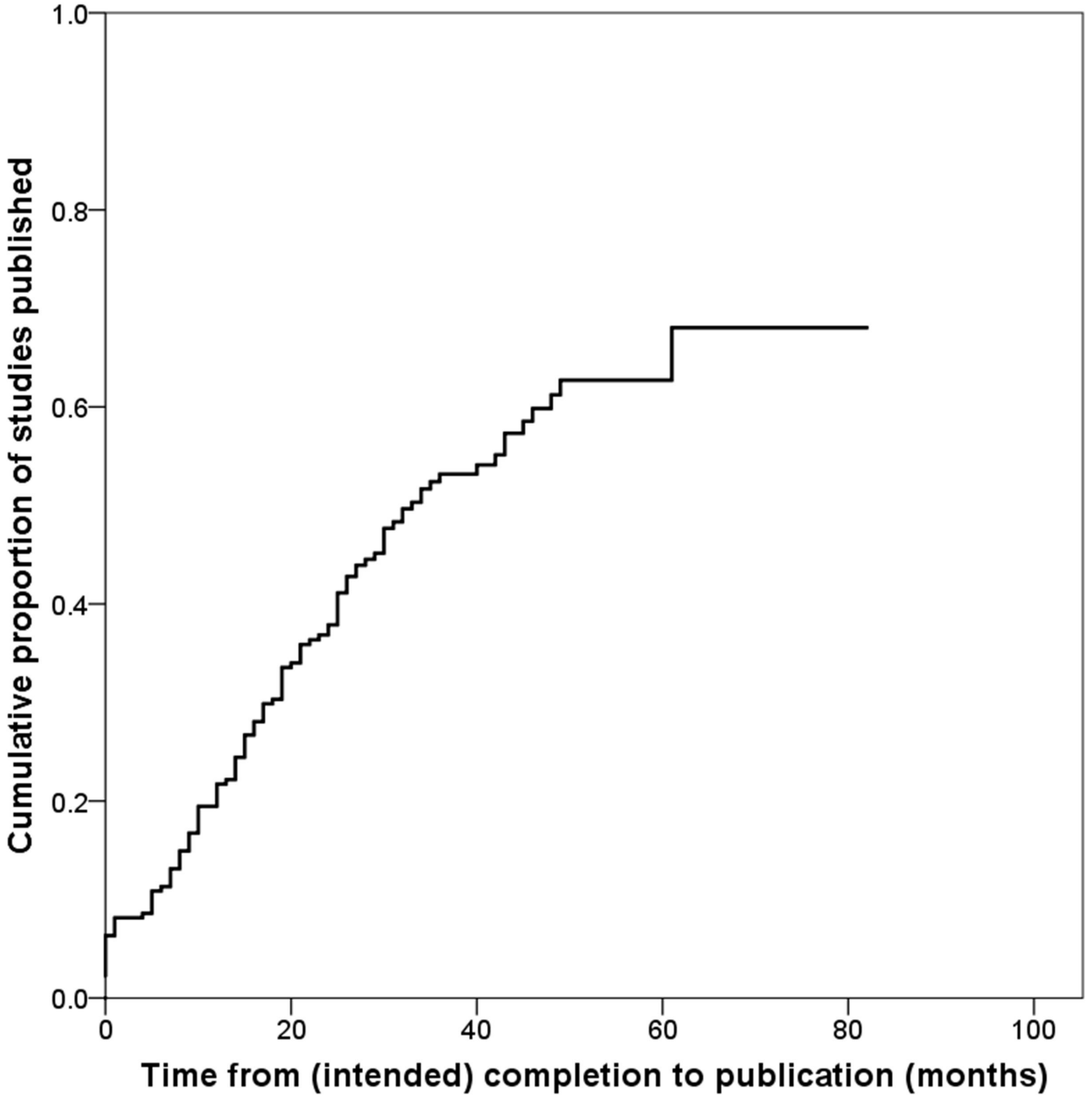
**BACKGROUND:** Failure to publish and selective reporting are recognized problems in the biomedical literature, but their extent in the field of diagnostic testing is unknown. We aimed to identify nonpublication and discrepancies between registered records and publications among registered test accuracy studies.

**METHODS:** We identified studies evaluating a test's accuracy against a reference standard that were registered in ClinicalTrials.gov between January 2006 and December 2010. We included studies if their completion date was set before October 2011, allowing at least 18 months until publication. We searched PubMed, EMBASE, and Web of Science and contacted investigators for publications.

should be further promoted among researchers and journal editors.



In recent years, failure to publish studies and selective reporting of research findings, each related to the strength and direction of outcomes *(1, 2)*, have been demonstrated several times in the biomedical literature *(3, 4)*. Studies with favorable results were shown to be more likely to be published than studies with negative or disappointing ones *(3, 5)*. This is regrettable for several reasons. The nonreporting of research results may lead to unnecessary duplication of research efforts, wasting time and money. Furthermore, the absence of information in the public domain can affect the evi-




ClinicalTrials.gov

418 evaluations of tests & markers registered

01-2006 – 12-2010

Excluding
94 registered after completion

N=324

(Daniel Korevaar et al. 2014)

# STARD (2003)

## Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative

Patrick M. Bossuyt,[1*] Johannes B. Reitsma,[1] David E. Bruns,[2,3] Constantine A. Gatsonis,[4] Paul P. Glasziou,[5] Les M. Irwig,[6] Jeroen G. Lijmer,[1] David Moher,[7] Drummond Rennie,[8,9] and Henrica C.W. de Vet,[10] for the STARD Group

**Table 1. STARD checklist for the reporting of studies of diagnostic accuracy.**

| Section and Topic | Item # | | On page # |
|---|---|---|---|
| TITLE/ABSTRACT/ KEYWORDS | 1 | Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity'). | |
| INTRODUCTION | 2 | State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups. | |
| METHODS | | Describe | |
| *Participants* | 3 | The study population: The inclusion and exclusion criteria, setting and locations where the data were collected. | |
| | 4 | Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard? | |
| | 5 | Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected. | |
| | 6 | Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)? | |
| *Test methods* | 7 | The reference standard and its rationale. | |
| | 8 | Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard. | |
| | 9 | Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard. | |
| | 10 | The number, training and expertise of the persons executing and reading the index tests and the reference standard. | |
| | 11 | Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers. | |
| *Statistical methods* | 12 | Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals). | |
| | 13 | Methods for calculating test reproducibility, if done. | |
| RESULTS | | Report | |
| *Participants* | 14 | When study was done, including beginning and ending dates of recruitment. | |
| | 15 | Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers). | |
| | 16 | The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended). | |
| *Test results* | 17 | Time interval from the index tests to the reference standard, and any treatment administered between. | |
| | 18 | Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition. | |
| | 19 | A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard. | |
| | 20 | Any adverse events from performing the index tests or the reference standard. | |
| *Estimates* | 21 | Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals). | |
| | 22 | How indeterminate results, missing responses and outliers of the index tests were handled. | |
| | 23 | Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done. | |
| | 24 | Estimates of test reproducibility, if done. | |
| DISCUSSION | 25 | Discuss the clinical applicability of the study findings. | |

thebmj

*BMJ* 2015;351:h5527 doi: 10.1136/bmj.h5527 (Published 28 October 2015) 

RESEARCH METHODS & REPORTING

# STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies



Incomplete reporting has been identified as a major source of avoidable waste in biomedical research. Essential information is often not provided in study reports, impeding the identification, critical appraisal, and replication of studies. To improve the quality of reporting of diagnostic accuracy studies, the Standards for Reporting Diagnostic Accuracy (STARD) statement was developed. Here we present STARD 2015, an updated list of 30 essential items that should be included in every report of a diagnostic accuracy study. This update incorporates recent evidence about sources of bias and variability in diagnostic accuracy and is intended to facilitate the use of STARD. As such, STARD 2015 may help to improve completeness and transparency in reporting of diagnostic accuracy studies.

Patrick M Bossuyt [1], Johannes B Reitsma [2], David E Bruns [3], Constantine A Gatsonis [4], Paul P Glasziou [5], Les Irwig [6], Jeroen G Lijmer [7], David Moher [8 9], Drummond Rennie [10 11], Henrica C W de Vet [12], Herbert Y Kressel [13 14], Nader Rifai [15 16], Robert M Golub [17 18], Douglas G Altman [19], Lotty Hooft [20], Daniël A Korevaar [1], Jérémie F Cohen [1 21], for the STARD Group



ORIGINAL RESEARCH ■ SPECIAL REPORT

# STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies[1]

Patrick M. Bossuyt, PhD
Johannes B. Reitsma, MD, PhD
David E. Bruns, MD
Constantine A. Gatsonis, PhD
Paul P. Glasziou, MRCGP, FRACGP, PhD, MBBS
Les Irwig, MBBS, PhD
Jeroen G. Lijmer,MD, PhD
David Moher, MD, PhD
Drummond Rennie, MD, MACP, FRCP
Henrica C.W. de Vet, PhD
Herbert Y. Kressel, MD
Nader Rifai, PhD, DABCC, FACB
Robert M. Golub, MD
Douglas G. Altman, DSc
Lotty Hooft, PhD
Daniël A. Korevaar, MD
Jérémie F. Cohen, MD, PhD
For the STARD Group

Incomplete reporting has been identified as a major source of avoidable waste in biomedical research. Essential information is often not provided in study reports, impeding the identification, critical appraisal, and replication of studies. To improve the quality of reporting of diagnostic accuracy studies, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was developed. Here we present STARD 2015, an updated list of 30 essential items that should be included in every report of a diagnostic accuracy study. This update incorporates recent evidence about sources of bias and variability in diagnostic accuracy and is intended to facilitate the use of STARD. As such, STARD 2015 may help to improve completeness and transparency in reporting of diagnostic accuracy studies.

Radiology

**Clinical Chemistry** 61:12
1446–1452 (2015)

Special Reports

# STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies

Patrick M. Bossuyt,[1*] Johannes B. Reitsma,[2] David E. Bruns,[3] Constantine A. Gatsonis,[4] Paul P. Glasziou,[5] Les Irwig,[6] Jeroen G. Lijmer,[7] David Moher,[8,9] Drummond Rennie,[10,11] Henrica C.W. de Vet,[12] Herbert Y. Kressel,[13,14] Nader Rifai,[15,16] Robert M. Golub,[17,18] Douglas G. Altman,[19] Lotty Hooft,[20] Daniël A. Korevaar,[1] and Jérémie F. Cohen,[21,22] for the STARD Group

# STARD group members

Penny Whiting
Marie Westwood
Nandini Dendukuri
David Simel
Augusto Azuara-Blanco
Rita Horvath
Ann van den Bruel
Anne Rutjes
Lucas Bachmann
Jeffrey Blume
Frank Buntinx
Blanca Lumbreras
Chris Hyde
Carl Heneghan
Ewout Steyerberg
Eleanor Ochodo
Gianni Virgili
Holly Janes
Joris de Groot
Jac Dinnes
Carl Moons

Mariska Leeflang
Matthew Thompson
Margaret Pepe
Nynke Smidt
Nancy Obuchowski
Petra Macaskill
Katie Morris
Reem Mustafa
Rosanna
Steffen Petersen
Sally Lord
Holger Schunemann
Susan Mallett
Todd Alonzo
Andrew Vickers
Nancy L. Wilczynski
Yemisi Takwoingi
Nitika Pai
Sarah Byron
Stephanie Chang
Stefan Lange

Hans van Maanen
William Summerskill
Herbert Kressel
Nader Rifai
Robert Golub
Philippe Ravaud
Isabelle Boutron
Richelle Cooper
John Ioannidis
Iveta Simera
Andreas Ziegler
Doug Altman
Jon Deeks
Kenneth Fleming
Gordon Guyatt
Myriam Hunink
Jos Kleijnen
Andre Knottnerus
Erik Magid
Barbara McNeil
Matthew McQueen

Andrew Onderdonk
Christopher Price
Sharon Straus
Stephen Walter
Wim Weber
Constantine Gatsonis
Les Irwig
David Moher
Riekie de Vet
David Bruns
Paul Glasziou
Jeroen Lijmer
Drummond Rennie
Hans Reitsma
Jorgen Hilden
Harry Büller
Frank Davidoff
John Overbeke
Daniël Korevaar
Lotty Hooft
Jérémie Cohen
Patrick Bossuyt

**Table 1**

**The STARD 2015 List**

| Section and Topic | No. | Item |
|---|---|---|
| TITLE OR ABSTRACT | | |
| | 1 | Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC) |
| ABSTRACT | | |
| | 2 | Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts) |
| INTRODUCTION | | |
| | 3 | Scientific and clinical background, including the intended use and clinical role of the index test |
| | 4 | Study objectives and hypotheses |
| METHODS | | |
| Study design | 5 | Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study) |
| Participants | 6 | Eligibility criteria |
| | 7 | On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry) |
| | 8 | Where and when potentially eligible participants were identified (setting, location and dates) |
| | 9 | Whether participants formed a consecutive, random or convenience series |
| Test methods | 10a | Index test, in sufficient detail to allow replication |
| | 10b | Reference standard, in sufficient detail to allow replication |
| | 11 | Rationale for choosing the reference standard (if alternatives exist) |
| | 12a | Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory |
| | 12b | Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory |
| | 13a | Whether clinical information and reference standard results were available to the performers/readers of the index test |
| | 13b | Whether clinical information and index test results were available to the assessors of the reference standard |
| Analysis | 14 | Methods for estimating or comparing measures of diagnostic accuracy |
| | 15 | How indeterminate index test or reference standard results were handled |
| | 16 | How missing data on the index test and reference standard were handled |
| | 17 | Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory |
| | 18 | Intended sample size and how it was determined |
| RESULTS | | |
| Participants | 19 | Flow of participants, using a diagram |
| | 20 | Baseline demographic and clinical characteristics of participants |
| | 21a | Distribution of severity of disease in those with the target condition |
| | 21b | Distribution of alternative diagnoses in those without the target condition |
| | 22 | Time interval and any clinical interventions between index test and reference standard |
| Test results | 23 | Cross tabulation of the index test results (or their distribution) by the results of the reference standard |
| | 24 | Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals) |
| | 25 | Any adverse events from performing the index test or the reference standard |
| DISCUSSION | | |
| | 26 | Study limitations, including sources of potential bias, statistical uncertainty, and generalisability |
| | 27 | Implications for practice, including the intended use and clinical role of the index test |
| OTHER INFORMATION | | |
| | 28 | Registration number and name of registry |
| | 29 | Where the full study protocol can be accessed |
| | 30 | Sources of funding and other support; role of funders |

| | | Report |
|---|---|---|
| *Participants* | | |
| | 6 | Eligibility criteria |
| | 7 | On what basis potentially eligible participants were identified<br>(such as symptoms, results from previous tests, inclusion in registry) |
| | 8 | Where and when potentially eligible participants were identified (setting, location and dates) |
| | 9 | Whether participants formed a consecutive, random or convenience series |

Radiology

# Overinterpretation and Misreporting of Diagnostic Accuracy Studies: Evidence of "Spin"[1]

Eleanor A. Ochodo, MBChB, MIH
Margriet C. de Haan, MD
Johannes B. Reitsma, MD, PhD
Lotty Hooft, PhD
Patrick M. Bossuyt, PhD
Mariska M. G. Leeflang, PhD

**Purpose:** To estimate the frequency of distorted presentation and overinterpretation of results in diagnostic accuracy studies.

**Materials and Methods:** MEDLINE was searched for diagnostic accuracy studies published between January and June 2010 in journals with an impact factor of 4 or higher. Articles included were primary studies of the accuracy of one or more tests in which the results were compared with a clinical reference standard. Two authors scored each article independently by using a pretested data-extraction form to identify actual overinterpretation and practices that facilitate overinterpretation, such as incomplete reporting of study methods or the use of inappropriate methods (potential overinterpretation). The frequency of overinterpretation was estimated in all studies and in a subgroup of imaging studies.

BMJ Open

# STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration

Jérémie F Cohen,[1,2] Daniël A Korevaar,[1] Douglas G Altman,[3] David E Bruns,[4] Constantine A Gatsonis,[5] Lotty Hooft,[6] Les Irwig,[7] Deborah Levine,[8,9] Johannes B Reitsma,[10] Henrica C W de Vet,[11] Patrick M M Bossuyt[1]

**To cite:** Cohen JF, Korevaar DA, Altman DG, *et al*. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;**6**:e012799. doi:10.1136/bmjopen-2016-012799

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-012799).

JFC and DAK contributed equally to this manuscript and share first authorship.

Received 26 May 2016
Revised 3 August 2016
Accepted 25 August 2016

For numbered affiliations see end of article.

**Correspondence to**
Professor Patrick M M Bossuyt; p.m.bossuyt@amc.uva.nl

## ABSTRACT

Diagnostic accuracy studies are, like other clinical studies, at risk of bias due to shortcomings in design and conduct, and the results of a diagnostic accuracy study may not apply to other patient groups and settings. Readers of study reports need to be informed about study design and conduct, in sufficient detail to judge the trustworthiness and applicability of the study findings. The STARD statement (Standards for Reporting of Diagnostic Accuracy Studies) was developed to improve the completeness and transparency of reports of diagnostic accuracy studies. STARD contains a list of essential items that can be used as a checklist, by authors, reviewers and other readers, to ensure that a report of a diagnostic accuracy study contains the necessary information. STARD was recently updated. All updated STARD materials, including the checklist, are available at http://www.equator-network.org/reporting-guidelines/stard. Here, we present the STARD 2015 explanation and elaboration document. Through commented examples of appropriate reporting, we clarify the rationale for each of the 30 items on the STARD 2015 checklist, and describe what is expected from authors in developing sufficiently informative study reports.

## INTRODUCTION

Diagnostic accuracy studies are at risk of bias, not unlike other clinical studies. Major sources of bias originate in methodological deficiencies, in participant recruitment, data collection, executing or interpreting the test or in data analysis.[1 2] As a result, the estimates of sensitivity and specificity of the test that is compared against the reference standard can be flawed, deviating systematically from what would be obtained in ideal circumstances (see key terminology in table 1). Biased results can lead to improper recommendations about testing, negatively affecting patient outcomes or healthcare policy.

Diagnostic accuracy is not a fixed property of a test. A test's accuracy in identifying patients with the target condition typically varies between settings, patient groups and depending on prior testing.[2] These sources of variation in diagnostic accuracy are relevant for those who want to apply the findings of a diagnostic accuracy study to answer a specific question about adopting the test in his or her environment. Risk of bias and concerns about the applicability are the two key components of QUADAS-2, a quality assessment tool for diagnostic accuracy studies.[3]

Readers can only judge the risk of bias and applicability of a diagnostic accuracy study if they find the necessary information to do so in the study report. The published study report has to contain all the essential information to judge the trustworthiness and relevance of the study findings, in addition to a complete and informative disclose about the study results.

Unfortunately, several surveys have shown that diagnostic accuracy study reports often fail to transparently describe core elements.[4–6] Essential information about included patients, study design and the actual results is frequently missing, and recommendations about the test under evaluation are often generous and too optimistic.

To facilitate more complete and transparent reporting of diagnostic accuracy studies, the STARD statement was developed: Standards for Reporting of Diagnostic Accuracy Studies.[7] Inspired by the Consolidated Standards for the Reporting of Trials or CONSORT statement for reporting randomised controlled trials,[8 9] STARD contains a checklist of items that should be reported in any diagnostic accuracy study.

The STARD statement was initially released in 2003 and updated in 2015.[10] The objectives of this update were to include recent evidence about sources of bias and variability and other issues in complete reporting, and



Downloaded from http://bmjopen.bmj.com/ on November 27, 2016 - Published by group.bmj.com



Table 1 Key STARD terminology

| Term | Explanation |
|---|---|
| Medical test | Any method for collecting additional information about the current or future health status of a patient |
| Index test | The test under evaluation |
| Target condition | The disease or condition that the index test is expected to detect |
| Clinical reference standard | The best available method for establishing the presence or absence of the target condition. A gold standard would be an error-free reference standard |
| Sensitivity | Proportion of those with the target condition who test positive with the index test |
| Specificity | Proportion of those without the target condition who test negative with the index test |
| Intended use of the test | Whether the index test is used for diagnosis, screening, staging, monitoring, surveillance, prediction, prognosis or other reasons |
| Role of the test | The position of the index test relative to other tests for the same condition (eg, triage, replacement, add-on, new test) |
| Indeterminate results | Results that are neither positive or negative |

make the STARD list easier to use. The updated STARD 2015 list now has 30 essential items (table 2).

Below, we present an explanation and elaboration of STARD 2015. This is an extensive revision and update of a similar document that was prepared for the STARD 2003 version.[11] Through commented examples of appropriate reporting, we clarify the rationale for each item and describe what is expected from authors.

We are confident that these descriptions can further assist scientists in writing fully informative study reports, and help peer reviewers, editors and other readers in verifying that submitted and published manuscripts of diagnostic accuracy studies are sufficiently detailed.

## STARD 2015 ITEMS: EXPLANATION AND ELABORATION

### Title or abstract

*Item 1. Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values or AUC)*

**Example**. 'Main outcome measures: Sensitivity and specificity of CT colonography in detecting individuals with advanced neoplasia (i.e., advanced adenoma or colorectal cancer) 6 mm or larger'.[12]

**Explanation**. When searching for relevant biomedical studies on a certain topic, electronic databases such as MEDLINE and Embase are indispensable. To facilitate retrieval of their article, authors can explicitly identify it as a report of a diagnostic accuracy study. This can be performed by using terms in the title and/or abstract that refer to measures of diagnostic accuracy, such as 'sensitivity', 'specificity', 'positive predictive value', 'negative predictive value', 'area under the ROC curve (AUC)' or 'likelihood ratio'.

In 1991, MEDLINE introduced a specific keyword (MeSH heading) for indexing diagnostic studies: 'Sensitivity and Specificity.' Unfortunately, the sensitivity of using this particular MeSH heading to identify diagnostic accuracy studies can be as low as 51%.[13] As of May 2015, Embase's thesaurus (Emtree) has 38 check tags for study types; 'diagnostic test accuracy study' is one of them, but was only introduced in 2011.

*In the example*, the authors mentioned the terms 'sensitivity' and 'specificity' in the abstract. The article will now be retrieved when using one of these terms in a search strategy, and will be easily identifiable as one describing a diagnostic accuracy study.

### Abstract

*Item 2. Structured summary of study design, methods, results and conclusions (for specific guidance, see STARD for Abstracts)*

**Example**. See STARD for Abstracts (*manuscript in preparation; checklist will be available at* http://www.equator-network.org/reporting-guidelines/stard/).

**Explanation**. Readers use abstracts to decide whether they should retrieve the full study report and invest time in reading it. In cases where access to the full study report cannot be obtained or where time is limited, it is conceivable that clinical decisions are based on the information provided in abstracts only.

In two recent literature surveys, abstracts of diagnostic accuracy studies published in high-impact journals or presented at an international scientific conference were found insufficiently informative, because key information about the research question, study methods, study results and the implications of findings were frequently missing.[14 15]

Informative abstracts help readers to quickly appraise critical elements of study validity (risk of bias) and applicability of study findings to their clinical setting (generalisability). Structured abstracts, with separate headings for objectives, methods, results and interpretation, allow readers to find essential information more easily.[16]

Building on STARD 2015, the newly developed STARD for Abstracts provides a list of essential items that should be included in journal and conference abstracts of diagnostic accuracy studies *(list finalised; manuscript under development)*.

### Introduction

*Item 3. Scientific and clinical background, including the intended use and clinical role of the index test*

# STARD for Abstracts

| Section | Item |
|---|---|
| | Identify abstract as a report of a diagnostic accuracy study<br>(using at least one measure of accuracy, such as sensitivity, specificity, predictive values, or area under the ROC curve) |
| | Describe: |
| **BACKGROUND** | |
| | Study objectives |
| **METHODS** | |
| | Data collection: whether this is a prospective or retrospective study |
| | Eligibility criteria for participants and the settings where the data were collected |
| | Whether participants formed a consecutive, random or convenience series |
| | Description of the index test and reference standard |
| **RESULTS** | |
| | Number of participants with and without the target condition included in the analysis |
| | Estimates of accuracy with measures of statistical uncertainty |
| **DISCUSSION** | |
| | General interpretation of the results |
| | Implications for practice, including the intended use of the index test |

# Learning objectives

After this session, students should be able to explain

- some of the difficulties in imaging RCT
- more efficient designs for randomized trials in imaging
- how STARD 2015 can reduce waste in imaging research

# Outcome Studies

Patrick MM Bossuyt

Amsterdam UMC
University Medical Centers