# Leveraging coreference to identify arms in medical abstracts: An experimental study

Elisa Ferracane[1], Katrin Erk[1], Byron Wallace[2], Iain Marshall[3]

[1]University of Texas at Austin [2]Northeastern University [3]Kings College London

# Motivation

"Some experts estimate that *only 20 percent of medical practices are based on rigorous research evidence*. The rest are based on ... *a kind of folklore.*"

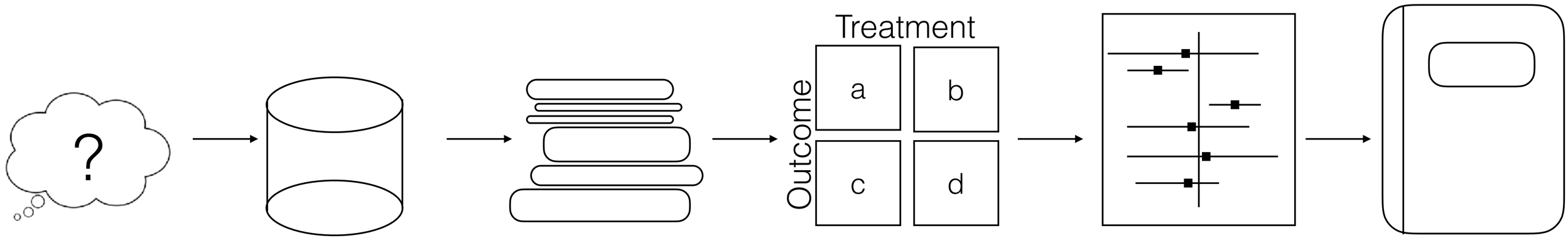
-New York Times, 2001

# Systematic Review

# Example

- **Question:** Treat childhood obesity?
- **Treatment:** Mandometer, scale that tracks the rate at which food leaves the plate.

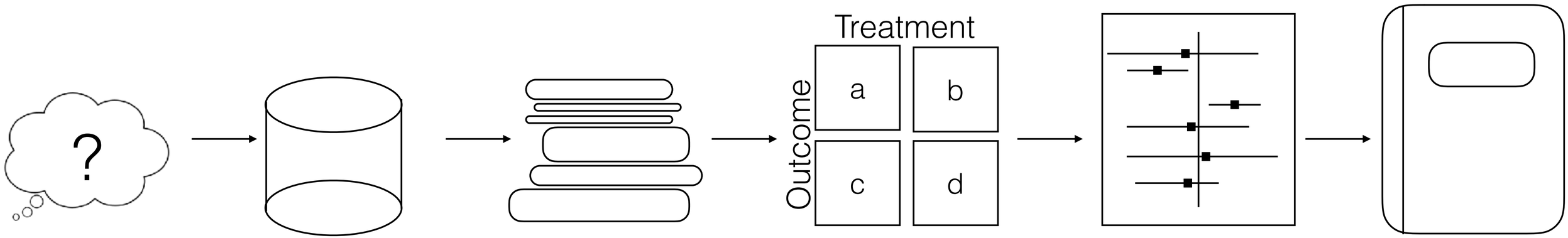
# Systematic Review




Topic Search Screen Extract Synthesize Report

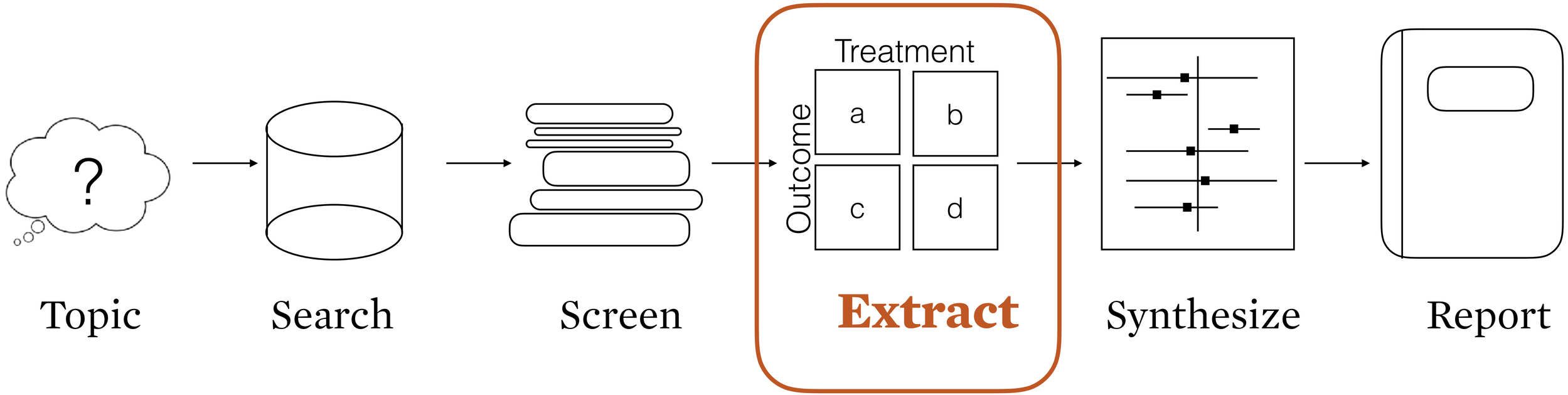

RCT
OLD
New

# Systematic Review




Topic Search Screen Extract Synthesize Report

# Systematic Review



# Arms

- Extract information from report of clinical trial:
- PICO:
  - **P**opulation
  - **I**ntervention
  - **C**ontrol/Comparator
  - **O**utcome

# Arms

- Extract information from report of clinical trial:
- PICO:
  - **P**opulation
  - **Intervention**
  - **Control/Comparator**
  - **O**utcome

# Arm

- Arm: **group** in a study

intervention arm(s) vs. control arm(s)

*mandometer group vs. standard lifestyle modification*

*To determine whether modifying eating behaviour with use of a feedback device facilitates weight loss in obese adolescents.*

# Experiment

- **Task**: label *token* in a *document* as part of an arm or not
- **Token**: word type (*Mandometer*)
- **Document**: abstract of clinical trial report
  - Why abstract and not full text?
    - less noise: more explicit and succinct
    - readily available
    - annotated*

*Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Huperff, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 372–377. IEEE.

# Example Abstract

OBJECTIVE: To determine whether modifying eating behaviour with use of a feedback device facilitates weight loss in obese adolescents.

DESIGN: Randomised controlled trial with 12 month intervention.

SETTING: Hospital based obesity clinic.

PARTICIPANTS: 106 newly referred obese young people aged 9-17.

INTERVENTIONS: A computerised device, Mandometer, providing real time feedback to participants during meals to slow down speed of eating and reduce total intake; standard lifestyle modification therapy.

MAIN OUTCOME MEASURES: Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Secondary outcomes were body fat SDS, metabolic status, quality of life evaluation, change in portion size, and eating speed.

RESULTS: Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care (baseline adjusted mean difference 0.24, 95% confidence interval 0.11 to 0.36). Similar results were obtained when analyses included only the 91 who attended per protocol (baseline adjusted mean difference 0.27, 0.14 to 0.41; P<0.001), with the difference maintained at 18 months (0.27, 0.11 to 0.43; P=0.001) (n=87). The mean meal size in the Mandometer group fell by 45 g (7 to 84 g). Mean body fat SDS adjusted for baseline levels was significantly lower at 12 months (0.24, 0.10 to 0.39; P=0.001). Those in the Mandometer group also had greater improvement in concentration of high density lipoprotein cholesterol (P=0.043).

CONCLUSIONS: Retraining eating behaviour with a feedback device is a useful adjunct to standard lifestyle modification in treating obesity among adolescents.

TRIAL REGISTRATION: ClinicalTrials.gov NCT00407420.

# Coreference

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

Randomised controlled trial with 12 month intervention. Change in body mass index (BMI) standard deviation score (SDS) over 12 months with assessment 18 months after the start of the intervention. Using the last available data on all participants (n=106), those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprot

- mandometer: 2 chains, 3 x in same chain
- intervention: 1 chain, 2 x in same chain

# Experiment

- Train binary classifier (Support Vector Machine in scikit-learn) to label a token as part of an arm or not

  — class weights inversely proportional to class frequencies

  — linear kernel

- pair token with abstract it appears in to create [abstract, token] pairs

- 5-fold cross-validation

# Features

- **Baseline model:**
  - **b-o-w:** frequency of token in medical abstract
  - **drugbank:** whether token exists in DrugBank database
  - **tf-idf:** term frequency-inverse document frequency of token
- **Coreference model:**
  - **max_counts:** maximum number of times token appears in single chain
  - **num_chains:** number of chains the token appears in the same abstract

# Corpus

- 263 abstracts from the British Medical Journal (BMJ) annotated with the arms (and other PICO elements)
- Structured abstracts consisting of short phrases and incomplete sentences

| | |
|---|---|
| Number of Documents | 263 |
| Number of Tokens | 63,488 |
| Unique [abstract, token] pairs | 35,650 |
| Average no. tokens per document | 241 |
| Positive labels | 5,757 (9%) |

# Evaluation

- Arms are **spans** of text: *Mandometer group, behavioural intervention*
- Penalize the model for not getting the more important words (following Summerscales 2013)

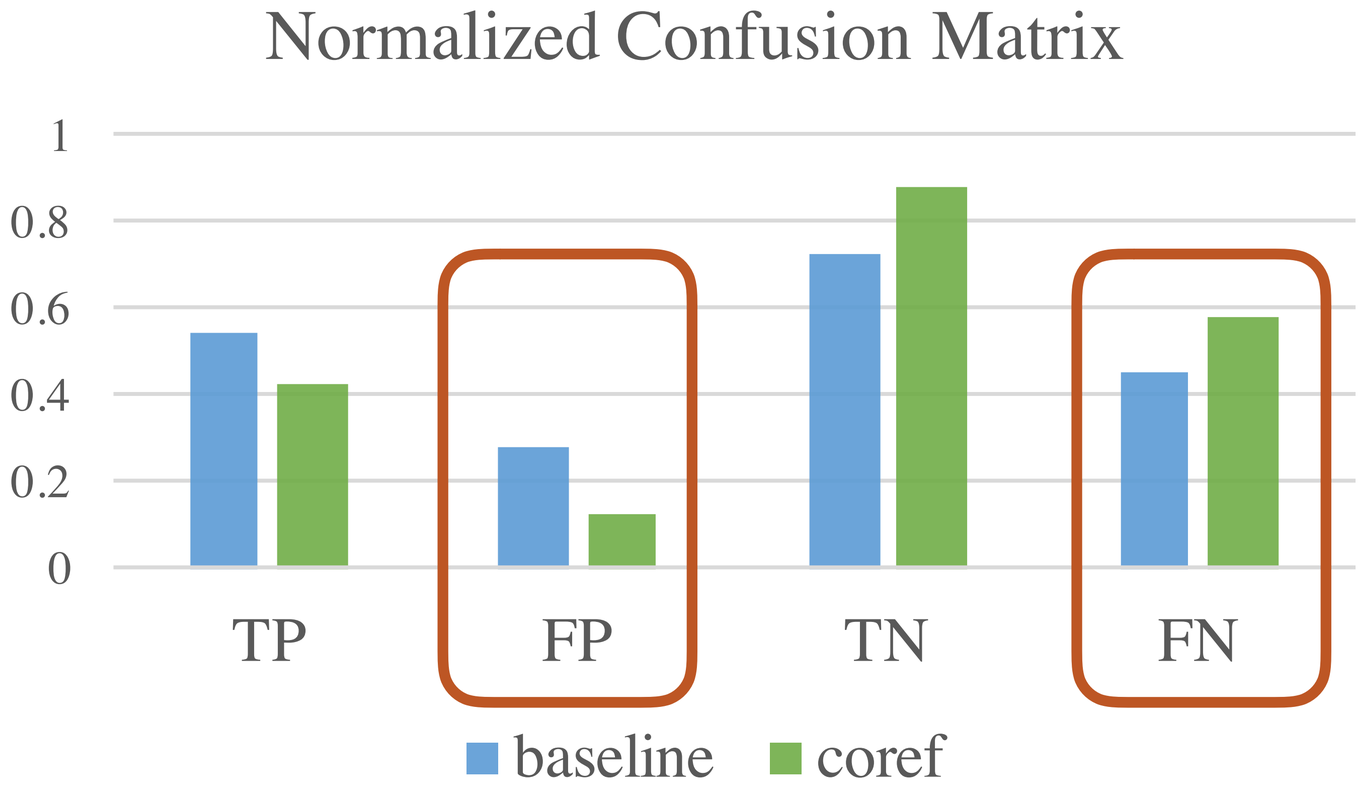
Gold: Mandometer group

Classifier: group > False Positive

Classifier: Mandometer > True Positive

# Results

| Model | P | R | F1 |
|---|---|---|---|
| Baseline | 12.9 | **88.6** | 22.5 |
| Coref | **19.7** | 82.7 | **31.8** |



# Error Analysis

- False Negatives:
- Poor recall for control arms:
  - — same words used across all abstracts >> low tf-idf
    - *control, sham, placebo, standard*
  - — usually mentioned once, if at all >> low coref scores
  - — but sometimes performs **better** than baseline:
    - some abstracts refer <u>often</u> to control arm: not salient across abstracts, but **salient in discourse**

# Error Analysis

- False Positives:

- *Annotation error:* **both** models got it “right”, but not annotated

... those in the Mandometer group had significantly lower mean BMI SDS at 12 months compared with standard care. The mean meal size in the Mandometer group fell by 45 g. Those in theMandometer group also had greater improvement in concentration of high density lipoprotein cholesterol.

- use for **bootstrapping**: annotator labels just 1 instance in abstract

# Error Analysis

- False Positives:

**Outcome** marked in black:

To determine the effects of a behavioural intervention for prevention of HIV and sexually transmitted diseases that identified , trained , and engaged leaders of Roma ( Gypsy ) men's social networks to counsel their own network members.

# Error Analysis

- False Positives:
  - labels other **salient PICO terms**!

**Outcome** marked in black:

To determine the effects of a behavioural intervention for prevention of HIV and **sexually transmitted** diseases that identified , trained , and engaged leaders of Roma ( Gypsy ) men's social networks to counsel their own network members.

# Conclusion

- Coreference **improves** arm identification over simple baseline
- Coreference identifies **discourse-salient** terms

  — help **identify other salient PICO terms** like outcomes!

# Future Work

- Will coreference still help with a strong baseline?

  — implement Summerscales 2013 as baseline

- Use FULL articles!

- What other discourse features can be used?

  — discourse connectives (*and, however*)

  — discourse template

# Q&A

## Leveraging coreference to identify arms in medical abstracts: An experimental study

elisa@ferracane.com

# Backup Slides

# Factorial design

Printed booklets : half had only general information from CancerBACUP about each patient's cancer and half had personalised information from the patient 's medical record plus selected general information; half were composed of information chosen interactively by the patient and half were produced automatically with a larger volume of material; and half had additional advice on anxiety management and half did not .

How many arms?