SELF-INSTRUCTIONAL MODULE FOR ARTICLES ON TREATMENT

Citation: Gaziano JM, Sesso HD, Christen WG, et al. Multivitamins in the prevention of cancer in men: The Physicians' Health Study II Randomized Controlled Trial. JAMA 2012: 308(18):doi:10.1001/jama2012.14641.

I. APPRAISING DIRECTNESS

Does the study provide a direct enough answer to your clinical question in terms of type of patients (P), exposure/intervention (E) used and disease or outcome (O) of interest?

Yes. Among healthy adult males (P), how effective is multivitamin use (E) in preventing cancer (O)?

Possible follow-up questions:

a. What are the components of a focused clinical question?

P=the patient population of interest; E=the exposures being compared; O=the outcomes being measured.
Note: M (methodology) is not part of the question.

b. Comparing the clinical PEO with the Research PEO in a directness table.

Comparison Table (Ask group to generate):

	Clinical Question	Research Question
P	Healthy adult males	Relatively healthy adult males
E	Multivitamin vs Usual Care	Multivitamin vs Placebo
O	Prevention of cancer	Prevention of cancer

c. In what way can there be a mismatch in P?

The target population may not be exactly the same:
1) disease spectrum might be different (eg - mild vs severe disease),
2) the general population may be different (eg – adults vs. children), or
3) the degree of uncertainty in the diagnosis may be different (eg – angiographically confirmed coronary disease vs. a diagnosis based purely on clinical grounds).

Note: Avoid getting too specific with "P" at this point or there will never be a perfect match. The goal of questioning directness is just to decide if the paper might be useful for the reader.

d. In what ways can there be a mismatch in the E's assessed?

The drug(s) studied may not be exactly the same:
1) same class but different molecule (sometimes useful, eg – when the drug effect is a "class effect")
2) same molecule but different route (eg – IV study but drug is only available orally)
3) unfair comparisons (seldom useful clinically, eg - maximum does of one drug vs usual dose of another)

When can we be convinced there is class effect?

When previous studies show a consistent effect of different members of a class of drugs.

e. In what ways can there be a discrepancy in O?

The outcomes that researchers are interested in may be different:
1) Researchers often measure surrogate outcomes; clinicians and patients want clinical outcomes.
2) Researchers often monitor composite endpoints; clinicians and patients want to know the effect of treatment on individual components.

What are surrogate endpoints?

A laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint

Why do researchers use surrogate endpoints?

Shorter study, smaller sample size, treatment is new and they want to see if it might work.

When can we believe surrogate outcomes?

When previous studies show a strong and consistent relationship between the surrogate and the clinical outcome. (Almost never.)

What are composite endpoints?

A combination of clinical events wherein any of those events will count as part of that composite endpoint.

Why do researchers use composite endpoints?

To decrease sample size requirements.

When can we believe composite endpoints?

When components are of approximately equal importance; when the effect of the intervention(s) on each component is consistent.

II. APPRAISING VALIDITY

1. Were patients randomly assigned to treatment groups?
Yes. Please refer to title and 1st sentence of Design in page E2. "The PHS II was a randomized, double-blind, placebo-controlled…"

Possible follow-up questions:

a. Why do we need to randomize?
To make 2 groups equal.
Suggested discussion: go through methods of randomization such as use of computer generated sequence of allocation or table of random numbers.

b. What are the consequences of 2 groups being unequal?
Treatment can look better or worse.
Label: This is called BIAS.

c. How does randomization make sure groups are comparable as to baseline characteristics?
Through sheer numbers; suggested example - toss of a coin.

d. Can we make 2 groups equal without randomization?
Only for known risk factors.

e. Should we insist on randomized trials for all treatment decisions we make?
No.

f. What are the exceptions?
Suggested approach: think of specific conditions where you wouldn't do an RCT (eg - surgery or no surgery for a ruptured appendix), then extract some general rules regarding exceptions:
1) illness with uniformly fatal (or adverse) outcome,
2) no known options for treatment (eg-some cancers),
3) treatment of few subjects reverses uniform adverse outcome (ex. Lorenzo's Oil, Awakening).
Suggested concept to introduce: equipoise or clinical equipoise - general uncertainty over whether a treatment will be beneficial.

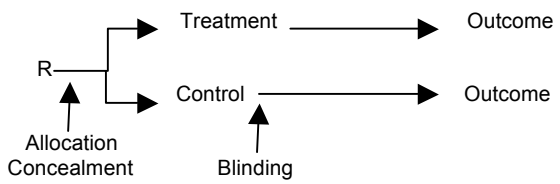
2. Was allocation concealed?
In order to determine the answer to this, the design paper must be retrieved (please refer to the "Design of Physicians' Health Study II" published in the Annals of Epidemiology in 2000).
Yes. Please see p. 130, under Randomization.

Possible follow-up questions:

a. Why is this important?
Randomization can be disturbed by good intentions (eg – the Australian study on open versus laparoscopic appendectomy).

b. How can it be done?
Randomization by 3rd party; computer randomization; opaque envelopes; telephone randomization.

c. How does allocation concealment differ from blinding?
Allocation concealment takes place before treatment is assigned. Blinding transpires after.
Allocation uses a 3rd party, opaque envelopes or a computer. Blinding uses a placebo.
Allocation starts the secrecy. Blinding continues it.



3. Were baseline characteristics similar at the start of the trial?
Yes. Please see p. E4, Table 1. Self-reported Baseline Characteristics According to Multivitamin Treatment Assignment in 14641 Men from PHS II. Under Results on p. E5, 2nd paragraph states that "all baseline characteristics had comparable distributions between the multivitamin and placebo groups."

Possible follow-up questions:

a. Why is this important?
This counter-checks if randomization was successful.

b. Why did the authors not give us a p-value when comparing baseline differences?
P values tell us the likelihood that the difference occurred by chance. What we are interested in is whether the differences can affect the results or not. This has to do with the magnitude of the difference rather than the probability that it occurred by chance.

Possible exercise if the answer cannot be extracted: Mean age reported in baseline characteristics of hypothetical studies on the effect of a statin on the risk of an MI. In which study are you more worried about the baseline characteristics? The correct answer of course is Study 2.

	Treatment	Control	P value
Study 1	71.8	69.9	0.04
Study 2	70.0	63.0	0.10

4. Were patients blinded to treatment assignment?
Yes. Please refer to abstract in the section of Design, Setting and Participants. "A large scale randomized, double-blind, placebo-controlled trial..." It was also noted in the 1st sentence under Design on p. E2.

Possible follow-up questions:
a. Why blind patients?
If patients know they're on placebo, are they likely to feel better or worse? (Answer is "worse").

b. Will this make the treatment look better or worse?
Answer is "better".

c. What are the strategies to blind patients?
Use of an identical placebo is the most common strategy, or an identical control treatment.

d. Can you blind patients to all kinds of treatments?
No. Some things are difficult to blind: dietary interventions, surgery, invasive procedures, physical therapy, educational interventions.

5. Were caregivers blinded to treatment assignment?
Yes. Please refer to abstract in the section of Design, Setting and Participants. "A large scale randomized, double-blind, placebo-controlled trial..." It was also noted in the 1st sentence under Design on p. E2.

Possible follow-up questions:
a. Why blind health caregivers? (i.e.-the doctors, nurses or therapists, etc)
If they know what their patient is on, they might try to change their management, ie, be more or less aggressive.

b. What are the strategies to blind healthcare caregivers?
Use of an identical placebo is the most common strategy, or an identical control treatment. Or, information can be simply withheld.

c. Can you blind caregivers to all kinds of treatments?
No. Some things are difficult to blind: dietary interventions, surgery, invasive procedures, physical therapy, educational interventions.

6. Were outcome assessors blinded to treatment assignment?
Yes. Please see p. E5, 1st column, 2nd paragraph. "All cancer and mortality end points were assessed and validated by medical record review by the PHS II Endpoints Committee composed of physicians blinded to treatment assignment."

Possible follow-up questions:
a. Who are the usual outcome assessors in a study?
They may be study personnel, caregivers, or even patients assessing themselves.

b. Why blind the outcome assessors in a study?
Knowledge of the treatment may affect assessment of the outcome.

c. What are the strategies to blind outcome assessors?
If the outcome assessors are patients or caregivers, then questions 4 and 5 already address this issue, that is, an identical placebo or control treatment may be used. On the other hand, if the outcome assessors are study personnel, then judgments may be used based on paper summaries, and information on treatment assignment can be withheld (eg – an independent adjudication committee)

d. Can you blind outcome assessors to all kinds of treatments?
Yes. This is particularly useful for lessening bias, especially when you can't blind patients and caregivers. That is why a PROBE design has evolved – a prospective, randomized, open, blinded endpoint study.

7. Were all patients analyzed in the groups to which they were originally randomized?
Yes. Please refer to section on Statistical Analysis, p. E5, 1st sentence. "All primary analyses classified study participants based on the intention-to-treat principle, in which all 14,641 randomized participants were classified to their randomized multivitamin treatment assignment…"

Possible follow-up questions:

a. What is the difference between a censored analysis and an ITT analysis?

Possible exercise if the answer cannot be extracted – Draw the following table, one column at a time:

		ANALYSIS A		ANALYSIS B
N	Treatments compared	Failure Rate (Compliers)	Failure Rate (Non-compliers)	% Failure (Total)
1,000	6-mo. Course	100/1000 = 10%	0/0	100/1000 = 10%
1,000	1-yr course	45/900 = 5%	90/100	135/1000 = 13.5%

1) Analysis A and B have different results – which one is better?
2) Which one is ideal world analysis, which one is real world analysis? ("A" is ideal, "B" is real world)
3) Which one addresses the question, "can the drug work?" Which one addresses the question "will the drug work?" ("A" addresses "can the drug work" question; "B" addresses "will the drug work".
4) Which one is ITT, which one is censored analysis? (Label: "A" is censored, "B" is ITT.)
5) Which analysis should MD's be interested in? (Usually "B")
6) Who would be interested in "A"? (Drug companies, researchers in general, because they try to bridge the gap between real and ideal)
7) Which one would patients be interested in? (Usually "A". They are not interested in asking "will the drug work even if I stop taking it?"
8) So which one should we report? (Preferably both analyses).

8. Was follow-up rate adequate?

No. Please refer to Figure 1 on page E3, 2nd to the last at the bottom of the figure. 48 were lost in the multivitamin group. 57 were lost in the placebo group. Since the benefit is quite small, the number of participants who dropped out affected the results assuming a worst case scenario on sensitivity analysis.

Possible follow-up questions:

a. How do study dropouts differ from non-compliant patients (usually referred to as withdrawn patients) addressed by criterion 7?

The non-compliers stopped taking the treatments, but they continued follow-up, so we know if they had an outcome. The dropouts never came back, so we have no idea about whether they had an outcome or not.

b. When should we worry about drop-outs?

When best and worst assumptions about their outcome may change the results.

Possible exercise if the answer cannot be extracted – Draw the following table. Start with row A and B then ask which dropout rate they would worry about (expected answer - study B). Then draw row C and D and ask the same question (expected answer – study D).

Study	Control drop-out rate	Treatment drop-out rate	Control death rate	Treatment death rate
A	1 %	1 %	20 %	10 %
B	1 %	1 %	0.2 %	0.1 %
C	10 %	10 %	50 %	10 %
D	10 %	10 %	10 %	5 %

From this example formulate general rules on when to worry:
1) When worst assumptions on what happened lead to opposite conclusions
2) When drop-out rates are greater than event rates

Note: If worst assumptions don't change the conclusions, then drop-outs don't matter. We say that the conclusions are "ROBUST". Otherwise, we call the conclusions "SOFT" and we accept the conclusions with a grain of doubt.

c. Perform a sensitivity analysis to evaluate the drop-out rate in the study (easier to use absolute numbers if the denominators for treatment and control approximate each other).

If treatment event rate (in ITT analysis) is 1290/7317 and control is 1379/7324, RR = 0.92 (0.86, 0.998)
In a sensitivity analysis we shall add the 48 who were lost-to-follow up in the treatment event rate (assuming these patients had cancer).
The treatment event rate would then be 1338/7317 and the control will remain as 1379/7324. The RR then will become 0.97 (0.91, 1.04).

d. Will the drop-outs affect the results in this study?

III. INTERPRETING RESULTS

1. How large was the effect of treatment?

Please refer to Table 2 on p. E6.

Possible follow-up questions:

a. What are the ways of expressing the effect of a treatment?

RR=Rt/Rc; ARR=Rc-Rt; RRR=Rc-Rt/Rc

Possible exercise if the answer cannot be extracted:

Scenario: If you weighed 80 KG after the Xmas holidays, and 60 KG after a summer diet, think of 3 ways of expressing your weight loss. (Draw this table to help learners think through the problem):

Control Weight	Treatment Weight			
80 KG	60 KG	a	b	c

Answers:
a – 25% (I lost 25% of my weight).
b – 75% (I am now 75% of what I used to weigh)
c – 20 KG (I lost 20 KG)

Add column headings, suggesting possible names for the figures they suggested:
a – the "relative weight reduction"
b – the "relative weight" (or weight that remains)
c – the "absolute weight reduction

Ask how they estimated these numbers; add a row of formulas (assigning the variables T and C respectively:

C	T	(C-T)/C	T/C	C-T

Add another row: if instead of weight, we talked of the risk of dying, in what ways would you express the change in risk?

Risk of Death Without Treatment=8%	Risk of Death With Treatment=6%	a	B	c

Answers:
a – 25% (I lost 25% of my risk).
b – 0.75 (My risk is now 0.75 of what I used to be.)
c – 2%(I lost 2% of my risk.)

Add another row, suggesting possible names (and eliciting formulas) for the figures they suggested:

Rc	Rt	RRR=(Rc-Rt)/Rc	RR=Rt/Rc	ARR=Rc-Rt

Note: RR is traditionally expressed in scientific notation, ARR and RRR in %. Why? To make things difficult for us.

Crude computation

Outcome: Total Cancer

Rc	Rt	RRR=(Rc-Rt)/Rc	RR=Rt/Rc	ARR=Rc-Rt
1379/7324 = 0.19	1290/7317 = 0.18	(0.19 – 0.18)/0.19 = 0.05	0.18/0.19 = 0.95	0.19 – 0.18 = 0.01
19%	18%	5%	0.95	NNT = 100

Outcome: Total Epithelial Cancer

Rc	Rt	RRR=(Rc-Rt)/Rc	RR=Rt/Rc	ARR=Rc-Rt
1244/7324 = 0.17	1158/7317 = 0.16	(0.17 – 0.16)/0.17 = 0.06	0.16/0.17 = 0.94	0.17 – 0.16 = 0.01
17%	16%	6%	0.94	NNT = 100

b. Which of the figures above is the best expression of the effect of a treatment?

RR is just the arithmetic inverse of RRR, so their importance is equal. The advantage of over ARR is that it is a constant ratio, which can be transferred between different settings, eg, high risk or low risk populations.

ARR varies from setting to setting, even from patient to patient. Its strength is that it reflects the effect of treatment in a particular situation.

Possible exercise if the answer cannot be extracted:

Scenario: If a treatment reduces the risk of death in patients with an AMI by 25%,

Control Mortality	Treatment Mortaity	RRR/RR	ARR
20%	15%	?/?	?

Answers: 25%, 0.75, 5%

Then add another row:

10%	7.5%	?/?	?

Answers: 25%, 0.75, 2.5%

Then add another row:

4%	3%	?/?	?

Answers: 25%, 0.75, 1%

Then add another row:

2%	1.5%	?/?	?

Answers: 25%, 0.75, 0.5%

Then add a last row:

1%	0.75%	?/?	?

Answers: 25%, 0.75, 0.25%

What does this table show?
1) RR/RRR are constant, but ARR is not
2) ARR becomes smaller as baseline risk becomes smaller (and vise versa)
3) You need the baseline risk in specific settings to estimate ARR
4) RR/RRR usually seem bigger, so readers can be misled.
5) We need to look at ARR

c. When do RRR, ARR and RR reflect harm?
RRR < 0 ARR < 0 RR > 1

d. When do they reflect benefit?
RRR > 0 ARR > 0 RR < 1

e. When do they reflect "no difference"?
RRR ~ 0 ARR ~ 0 RR ~ 1

f. How would results be reported if continuous variables were used?

Endpoint: eg – blood pressure			
Mean decrease from baseline (treatment)	Mean decrease from baseline (control)	Mean difference	p-value
Δ_t	Δ_c	$\Delta_c - \Delta_t$	

g. Estimate the RRR, RR, ARR in this study for a beneficial outcome, a harmful outcome (if available).

h. Estimate the mean difference for continuous outcome if available.

2. How precise was the estimate of the treatment effect?
Please refer to Table 2 on p. E6.

Possible follow-up questions:
a. What is the difference between a point estimate and an interval estimate?
Suggested exercise:
1) Ask for an estimate of the average height or weight (or some other parameter) of people in the room.
2) Ask them regarding the probability that the estimate is correct – (very low).
3) Now ask for an interval estimate, then, ask them re the probability that this is correct.
Points to bring out:
1) interval estimates are humbler because they accept a range of possibilities;
2) interval estimates are more likely to be correct;
3) more useful because aside from suggesting statistical significance, they convey a message re magnitude of effect, ie, the best & worst scenario;
4) all point estimates have surrounding interval estimates.

b. What do these sample results mean?
RR = 0.8 (95% CI: 0.2, 0.9) Answer: definite benefit
RR = 0.8 (95% CI: 0.5, 1.2) Answer: possible benefit or harm
RR = 1.5 (95% CI: 1.2, 1.7) Answer: definite harm

Possibly introduce the concepts of non-inferiority and equivalence for advanced groups.

c. What does a p-value mean when reported with point estimate of a treatment effect? eg RR=0.7, p=0.01
p is the probability that the observed differences are coincidental.

d. How does this form of reporting relate with interval estimates of treatment effects?
p<0.05 implies the 95% CI of the RR does not contain the value 1.0.

e. What are the advantages and disadvantages of reporting treatment effects as interval estimates instead of point estimates with corresponding p-values?
95% CIs can be understood more intuitively than p-values (studies have shown this in the past)

IV. ASSESSING APPLICABILITY: *Are the results applicable to the patients you see?*

1. Are there biologic issues that may affect applicability of treatment?
There are no significant biological concerns regarding applicability of treatment.

Possible follow-up questions:
a. What biologic issues should be considered in assessing applicability? Are any of these issues of concern in this study?
Biologic issues: sex, comorbidity, race, age, and pathology.

2. Are there socio-economic issues affecting applicability of treatment?
Multivitamins (especially branded ones) are expensive and its cost may limit a person's opportunity to buy something more important.

Possible follow-up questions:

a. What socio-economic issues should be considered in assessing applicability? Are any of these issues of concern in this study?
Socio-economic issues: provider compliance and patient compliance

b. What is the difference between directness and applicability?

Directness	**Applicability**
Should I read this article?	Can I use the results in my patients?
Did they study the disease of interest?	Are there factors that can modify the treatment effect (RR)?
Broad criteria (P, E, O)	Specific criteria (SCRAP and socio-economic issues)
Like deciding what clothes you will buy.	Like trying on the clothes before buying them.
Etc	Etc

Note: these issues affect applicability if they could potentially lower or increase the RR or RRR, ie, if they can enhance or reduce the effectiveness of treatment. Changes in baseline risk for the outcome of interest should not affect applicability (but may affect individualization).

V. INDIVIDUALIZING THE RESULTS

1. Estimate the individualized NNT for your patient(s).

For total cancer

Step 1. Estimate baseline risk of your patient (Rc)	5%
Step 2. Estimate RR with treatment	0.92
Step 3. Estimate post-treatment risk (Rt = Rc x RR)	Rt = 0.05 x 0.92 = 0.046
Step 4. Estimate ARR (Rc-Rt)	0.05 – 0.046 = 0.004
Step 5. Estimate Individualized NNT	1/0.004 = 250

Possible follow-up questions:

a. How do we estimate baseline risk?
A risk calculator is available for many diseases, or a risk scoring system, cohort studies on untreated patients, and the control group of the study (last choice because it is the least generalizable)

2. Would you offer the treatment to your patients?
You can present option, but need to be able to present all pros & cons since benefit is small.

Encourage debate. There are no right or wrong answers here. In the end, we may agree on the numbers but disagree in action plans because we value the numbers differently. Use this to illustrate that EBM is not just about evidence.

APPRAISING AN ARTICLE ON THERAPY

DIRECTNESS	Why is it Important	What to Look For
Does the study provide a direct enough answer to your clinical question in terms of patients (P), exposure (E) and outcome (O)? In this situation the exposure is the treatment that is being evaluated, and outcome is the condition being prevented.	Many times, the P, E and O are not exactly the same as those studied by the authors of a paper. If this is the case, you need to decide if you can use the study results at all. The decision requires some expertise on the disease under question.	Seek the opinion of an expert (this might be you), or your colleagues.
VALIDITY	**Why is it Important**	**What to Look For**
Were patients randomly assigned to treatment groups?	Randomization makes the treatment and control groups equal with regards all known or unknown prognostic factors.	Look for the word "randomize", "randomly allocated" in methods.
Was allocation concealed?	Concealment of the allocation helps protect the process of randomization.	Look for strategies such as use of opaque envelopes, randomization by third party, or randomization by computer.
Were baseline characteristics similar at the start of the trial?	This criterion assesses comparability of the groups being compared. The magnitude of any difference is more important than the reported p-value.	Look for a comparison in tables or in the text.
Were patients blinded to treatment assignment?	Blinding patients is necessary especially when subjective complaints are used as outcome measures. Patients are more likely to feel bad if they know they are not on active therapy.	Look for blinding strategies such as use of a placebo. This may not always be feasible or necessary.
Were caregivers blinded to treatment assignment?	Blinding caregivers is necessary because knowledge of what patients are receiving may affect how well they look after these patients .	Look for blinding strategies such as use of a placebo. This may not always be feasible or necessary.
Were outcome assessors blinded to treatment assignment?	Blinding outcome assessors separately may be necessary, especially when patients and caregivers cannot be blinded.	Look for strategies to withhold information regarding patient assignment.
Were all patients analyzed in the groups to which they were originally randomized?	Non-compliant patients should be evaluated as if they received a treatment because non-compliance is part of the effect of treatment.	Look for the term “intention-to-treat” under the planned analysis.
Was follow-up rate adequate?	If there are too many drop-outs with unknown outcome, the validity of a study is threatened.	Drop-outs should be stated explicitly in a paper. If not, compare number recruited with number of patients analyzed at end of study.

RESULTS	**Why is it Important**	**What to Look For**
How large was the effect of treatment?	This tells us how effective, ineffective or harmful the treatment is.	For dichotomous outcomes – look for hazards ratios, relative risk, relative risk reduction, or absolute risk reduction (also known as risk difference). For continuous outcomes – look for the mean difference.
How precise was the estimate of the treatment effect?	Because studies just give us estimates of the effect of a treatment, we need to know the range of possibilities rather than just a single number.	Look for 95% confidence intervals around the estimates of treatment effect mentioned above.
APPLICABILITY	**Why is it Important**	**What to Look For**
Are there biologic issues that may affect applicability of estimates of treatment effectiveness? (Consider the influence of sex, co-morbidity, race, age and pathology)	Sometimes, effectiveness of an intervention may depend on sex, presence of co-morbidities, race, age, or pathology of the disease in question.	Prior knowledge and experience with the disease will be useful.
Are there socio-economic issues that may affect applicability of estimates of treatment effectiveness?	Social, cultural and economic context may potentially affect how well a treatment works	Prior knowledge and experience with the disease will be useful.
INDIVIDUALIZING RESULTS	**Why is it Important**	**What to Look For**
What is the likely effect of the treatment on your individual patient?	Studies report average effects but the effect on your patient may not be average,	Using the patient's baseline risk for the outcome (based on clinical presentation) and the risk reduction (based on the study), one can estimate the individualized absolute risk reduction.