



Στατιστική και κλινικές δοκιμές

Δημήτρης Καρλής
Τμήμα Στατιστικής
ΟΠΑ

Τι θα κάνουμε σήμερα

- Hot Topic εξαιτίας του COVID
- Αδύνατον να το συζητήσουμε σε 45 λεπτα!

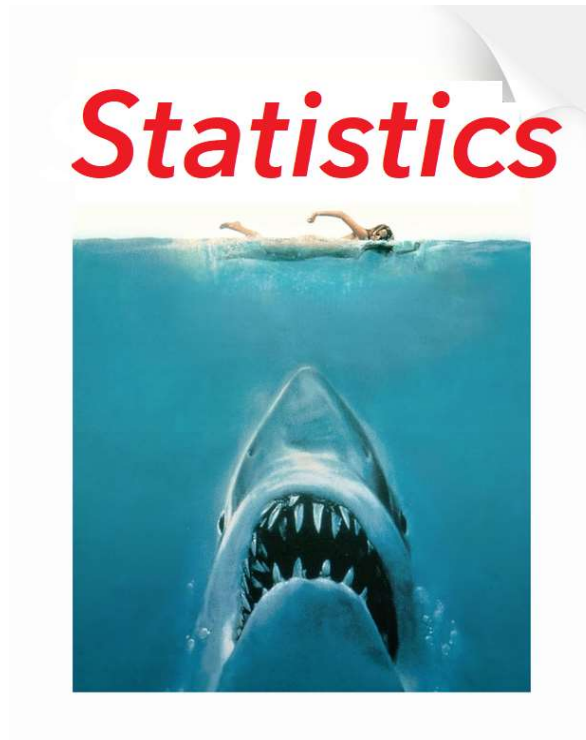
Εχω επιλέξει μερικά θέματα που
χρειάζονται σημαντικά στατιστική
Για να δούμε προβλήματα που
ζητάνε λύση και έχουν ενδιαφέρον



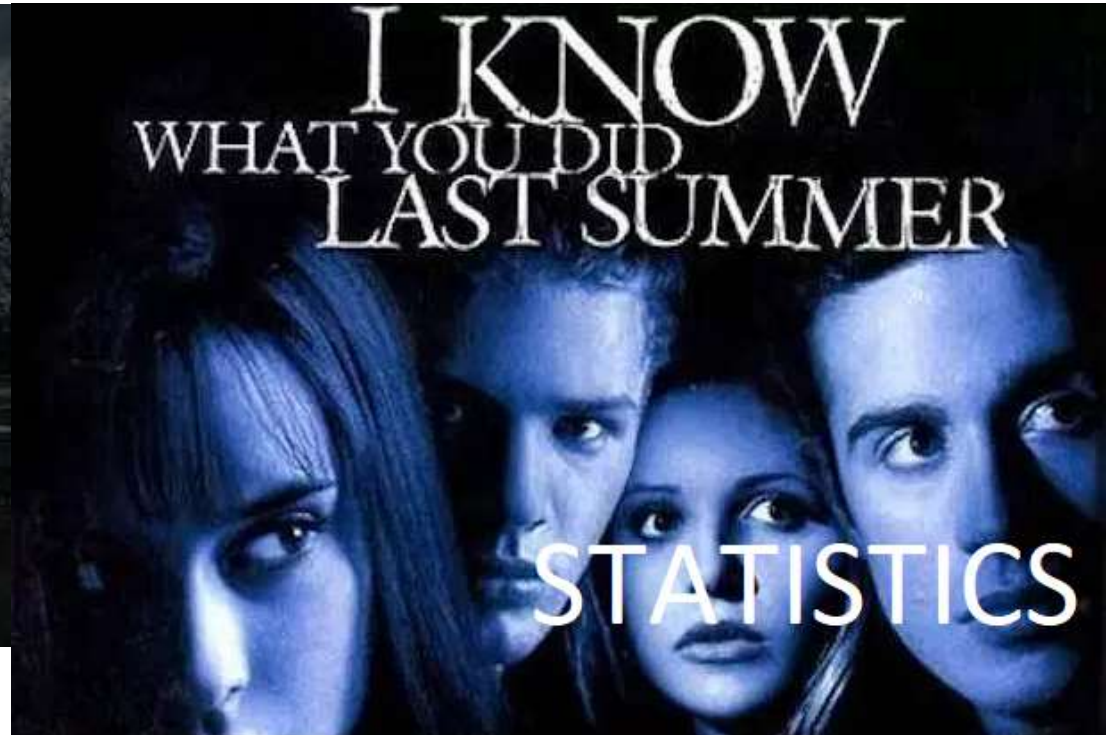
Εν συντομία

- Outcome selection
- Randomization
- Sample size determination
- Multiplicity Problem
- Interim analyses
- Futility
- Crossover
- Adaptive Designs
- Methodological Challenges (few of them)

Στατιστική: κάτι μου θυμίζει



Στατιστική: κάτι μου θυμίζει



Τι είναι μια κλινική δοκιμή

«Κλασσικός» ορισμός:

Μια προοπτική μελέτη παρέμβασης σε ομάδα ανθρώπων, η οποία συγκρίνει τα ευρήματα αναφορικά με την έκβαση μιας νόσου ή μιας κατάστασης, με μια άλλη ομάδα ελέγχου.

- Ερευνητική μελέτη σε ανθρώπους
- Εξετάζει την ορθότητα μιας ερευνητικής υπόθεσης
- Προτείνει τρόπους για πρόληψη ή θεραπεία
- Κατάλληλη για αιτιολογικές συσχετίσεις

Ένας εναλλακτικός ορισμός...

- Οποιαδήποτε μορφή σχεδιασμένου πειράματος στο οποίο λαμβάνουν μέρος άνθρωποι για να μπορέσει κανείς να εκτιμήσει ποιά είναι η καταλληλότερη θεραπεία για κάποια συγκεκριμένη πάθηση από μια σειρά από υπάρχουσες ή νέο-προτεινόμενες θεραπείες

Γιατί χρειάζονται οι κλινικές δοκιμές

Ένας ερευνητής έδωσε μαντζουράνα σε έναν ασθενή με έλκος και το έλκος ιατρεύτηκε.

Ο ερευνητής δημοσίευσε ένα άρθρο:

"Η μαντζουράνα θεραπεύει το έλκος στομάχου."

Ο επόμενος ασθενής που δοκίμασε μαντζουράνα ως θεραπεία απεβίωσε.

Ο ερευνητής δημοσίευσε το επόμενο άρθρο:

"Νέες μελέτες δείχνουν ότι η μαντζουράνα θεραπεύει το 50% των περιπτώσεων έλκους στομάχου"

Κλινικές δοκιμές στην πράξη

- Εξαιρετικά δύσκολο, πολύπλοκο, χρονοβόρο και κοστοβόρο επιχείρημα
- Απαιτεί συνεργασία πολλών ανθρώπων, εξαιρετική προετοιμασία και βαθιά γνώση
- Συνήθως περιλαμβάνει πολλά κέντρα, άρα απαιτεί συντονισμό
- Και μην ξεχνάμε ότι πολλά φάρμακα περιμένουν τη σειρά τους, αρα μια αποτυχία κοστίζει πολλαπλά

Τι μπορεί να πάει στραβά;

Πολύπλοκο και δύσκολο επιχείρημα

- Κακός σχεδιασμός (πχ κάποιες υποθέσεις δεν ήταν ρεαλιστικές)
- Κακή εκτίμηση πόσους ασθενείς μπορούμε να βρούμε
- Κακή οργάνωση και εκτέλεση της μελέτης
- Νέες πληροφορίες να «ακυρώσουν» τη μελέτη
- Και πολλά άλλα...

και προφανώς το φάρμακο δεν ήταν όπως το περιμέναμε (έλλειψη αποτελεσματικότητας, τοξικότητες κλπ)

Φάσεις Κλινικών Δοκιμών

- **Φάσης I:**

- Η φαρμακοκινητική και οι βιολογικές επιδράσεις του φαρμάκου εξετάζονται για πρώτη φορά σε ανθρώπους. Συνήθως λίγοι ασθενείς
- Στόχοι: Διερεύνηση Ασφάλειας
Καθορισμός Ασφαλούς Δόσης

- **Φάσης II ή I/II:**

- Μικρές μελέτες ασφάλειας και αποτελεσματικότητας
- Στόχος: Επιλογή λίγων δραστικών θεραπειών για περαιτέρω μελέτη σε δοκιμή Φάσης III

Είδη Κλινικών Δοκιμών

- **Φάσης III:**
 - Μεγάλου μεγέθους συγκριτικές μελέτες της πειραματικής θεραπείας έναντι της καθιερωμένης θεραπείας ή εφόσον δεν υπάρχει τέτοια θεραπεία, έναντι placebo (εικονικής θεραπείας).
 - Στόχοι: Διερεύνηση **αποτελεσματικότητας**
Έγκριση Φαρμάκου
- **Φάσης IV:**
 - Μελέτες παρατήρησης σε μεγάλο αριθμό ασθενών μετά την κυκλοφορία του φαρμάκου στην αγορά.
 - Στόχος: Αναγνώριση πιθανών ανεπιθύμητων ενεργειών

Σχεδιασμός

- Όσο προοδεύει η γνώση μας οι φάσεις μπορούν να συμπυκνθούν
- Πχ **Φάση I/II** δηλαδή να συγχωνευτούν δυο φάσεις
- Ομοίως Φάση **II/III**

Σκοπός τους να βελτιωθεί ο χρόνος και το κόστος.

Αντίτιμο: πιο δύσκολη ανάλυση και σχεδιασμός

Ορολογία

- Έκβαση (outcome)
- Τυχαιοποίηση (randomization)
- Τυφλότητα (blindness)
- Ομάδα ελέγχου (control group)
- Placebo (placebo)
- Ενδιάμεση ανάλυση (interim analysis)
- Κριτήρια ένταξης (Eligibility criteria)
- Πρωτόκολλο (Protocol)

Έλεγχος Υποθέσεων

- **Μηδενική Υπόθεση (H_0):**

Συνήθως η H_0 είναι μία δήλωση «μη ύπαρξης διαφοράς» ή «μη επίδρασης» και γενικά θέλουμε να μην ισχύει

- Ο μέσος του πληθυσμού ισούται με κάποια υποθετική τιμή μ_0 ,
 $H_0 : \mu = \mu_0$.
- Η διάμεση επιβίωση ατόμων που πήραν την θεραπεία B είναι 3 χρόνια, $H_0 : \delta=3$
- Το ποσοστό ασθενών που πήραν τη νέα θεραπεία και είχαν κλινικά μετρήσιμη βελτίωση ήταν 20%, $H_0 : \pi=0.20$

Έλεγχος Υποθέσεων

- **Εναλλακτική Υπόθεση (H_A):**

Η εναλλακτική υπόθεση είναι η δήλωση που ελπίζουμε να αληθεύει αντί της μηδενικής

- $H_A : \mu \neq \mu_0$.
- Διάμεση επιβίωση < 3 χρόνια, $H_A : \delta < 3$.
- Ποσοστό ασθενών είναι 30% $H_A : \pi > 0.20$, $\pi = 0.30$

Παρουσίαση των Υποθέσεων

- λ_T = Κίνδυνος εμφάνισης νόσου (Risk of a disease event) για την ομάδα του νέου φαρμάκου
- λ_S = Κίνδυνος εμφάνισης νόσου για την ομάδα παρατήρησης

Λόγος Κινδύνων, Hazard Ratio

$$\underline{\vartheta = HR} = \lambda_T / \lambda_S$$

$HR > 1$ σημαίνει $\lambda_T > \lambda_S$, **το νέο φάρμακο χειρότερο**

$HR = 1$ σημαίνει $\lambda_T = \lambda_S$, **καμία διαφορά**

$HR < 1$ σημαίνει $\lambda_T < \lambda_S$, **το νέο φάρμακο καλύτερο**

Παρουσίαση των Υποθέσεων

- Δουλεύουμε με το λογάριθμο του HR (Δ) για να μπορούμε να κάνουμε χρήση της κανονικής κατανομής
- Εκτιμώμενος λόγος κινδύνων στη λογαριθμική κλίμακα [$\ln(HR)$] για εμφάνιση νόσου μεταξύ της ομάδας του νεο φαρμάκου και της ομάδας παρατήρησης:

$$\Delta = \ln(HR) = \ln(\lambda_T) - \ln(\lambda_S)$$

όπου, Δ είναι **η αληθινή σχετική ελάττωση του κινδύνου εμφάνισης νόσου** για θεραπεία ενός έτους με το νεο φάρμακο έναντι απλής παρατήρησης

Παρουσίαση των Υποθέσεων

Λόγος Κινδύνων, Hazard Ratio $\vartheta = HR = \lambda_T / \lambda_S$

- Μηδενική Υπόθεση (Null):

$$H_0: \quad HR=1 \quad \text{OR} \quad \lambda_T = \lambda_S$$

- Εναλλακτική Υπόθεση (Alternative): **το νέο φάρμακο καλύτερο**

$$H_A: \quad HR \neq 1, HR < 1 \quad \text{OR} \quad \lambda_T < \lambda_S$$

Παρουσίαση των Υποθέσεων

Μηδενική Υπόθεση: $H_0: HR=1$ ($\lambda_T = \lambda_S$)

Εναλλακτική Υπόθεση: $H_A: HR \neq 1$, $HR < 1$ ($\lambda_T < \lambda_S$)

Κλινικά Σημαντική Διαφορά : 23%

Εάν στην πραγματικότητα HR είναι μικρότερος κατά 23%,
θα θέλαμε να υπάρχει βάσει της μελέτης **ΥΨΗΛΗ**
ΠΙΘΑΝΟΤΗΤΑ να βρεθεί αυτή η διαφορά.

Αυτή η πιθανότητα είναι η

ΣΤΑΤΙΣΤΙΚΗ ΙΣΧΥΣ, STATISTICAL POWER = 1- β

Έλεγχος Υποθέσεων

Το τελικό συμπέρασμα είναι ΠΑΝΤΑ ένα εκ των ακόλουθων δύο:

- Απορρίπτουμε τη Μηδενική Υπόθεση
- Αποτύχαμε να απορρίψουμε τη Μηδενική Υπόθεση

**ΠΡΟΣΟΧΗ!!! Η ΠΙΘΑΝΟΤΗΤΑ ΛΑΘΟΥΣ ΘΕΛΟΥΜΕ ΝΑ ΕΙΝΑΙ
ΟΣΟ ΜΙΚΡΟΤΕΡΗ ΓΙΝΕΤΑΙ**

Τύποι Σφαλμάτων

□ Στην εφαρμογή της τεχνικής του ελέγχου υποθέσεων, υπάρχουν δύο τύποι σφαλμάτων που μπορεί να διαπραχθούν.

- Το πρώτο καλείται **σφάλμα τύπου I**, και είναι επίσης γνωστό ως **σφάλμα απόρριψης**, ή **α σφάλμα**.

Η πιθανότητα να διαπράξει κανείς ένα σφάλμα τύπου I καθορίζεται από το επίπεδο σημαντικότητας του ελέγχου.

- Ο δεύτερος τύπος σφάλματος που μπορεί να διαπραχθεί κατά τη διάρκεια ενός ελέγχου υπόθεσης είναι ένα **σφάλμα τύπου II** ή **σφάλμα μη απόρριψης** ή **β σφάλμα**.

Η πιθανότητα να διαπράξει κανείς ένα σφάλμα τύπου II υπολογίζεται σε μία συγκεκριμένη τιμή της εναλλακτικής υπόθεσης.

Ισχύς

- Η **ισχύς** είναι η πιθανότητα να απορρίψουμε την H_0 όταν η μηδενική υπόθεση είναι λανθασμένη, δηλαδή είναι η πιθανότητα να αποφύγουμε ένα σφάλμα τύπου II.

$$\text{Ισχύς} = 1 - \beta$$

- Όπως και η β , η ισχύς πρέπει να υπολογιστεί για μία συγκεκριμένη τιμή για τη παράμετρο του πληθυσμού στην εναλλακτική υπόθεση.

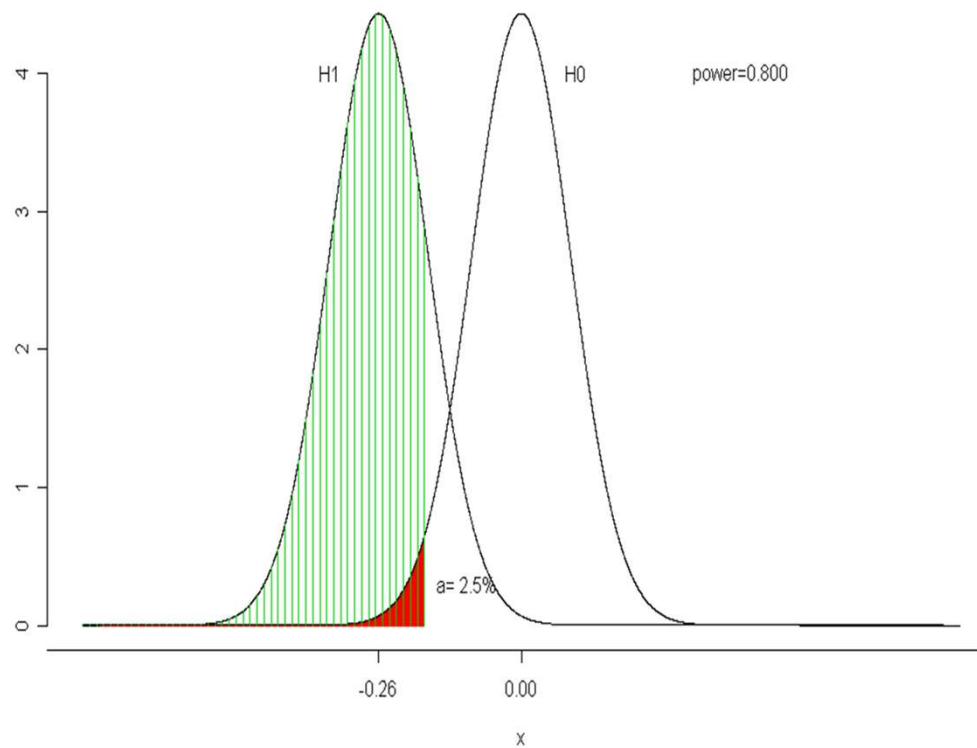
Ο πίνακας της αλήθειας

		Πραγματικότητα	
		H_0 αληθής	H_0 όχι αληθής
Απόφαση βάσει μελέτης	Δεν απορρίπτουμε H_0 $\Delta = 0$ $\ln(\text{HR}) = 0, \text{HR} = 1$ $\ln(\lambda_T) - \ln(\lambda_S) = 0$	 $(1-\alpha)$	Σφάλμα τύπου II (β)
	Απορρίπτουμε H_0 $\Delta < 0, \text{HR} < 1$	Σφάλμα τύπου I (α)	 $(\text{Ισχύς} = 1-\beta)$

Παρατήρηση

- Μεγάλες υπάρχουσες διαφορές μπορούν να ανιχνευτούν εύκολα κι επομένως στην πράξη χρειαζόμαστε μικρότερο μέγεθος δείγματος
- Μικρές διαφορές ανιχνεύονται πιο δύσκολα και άρα χρειαζόμαστε μεγαλύτερο δείγμα

HERA Trial (HR=0.77 vs. HR=1)

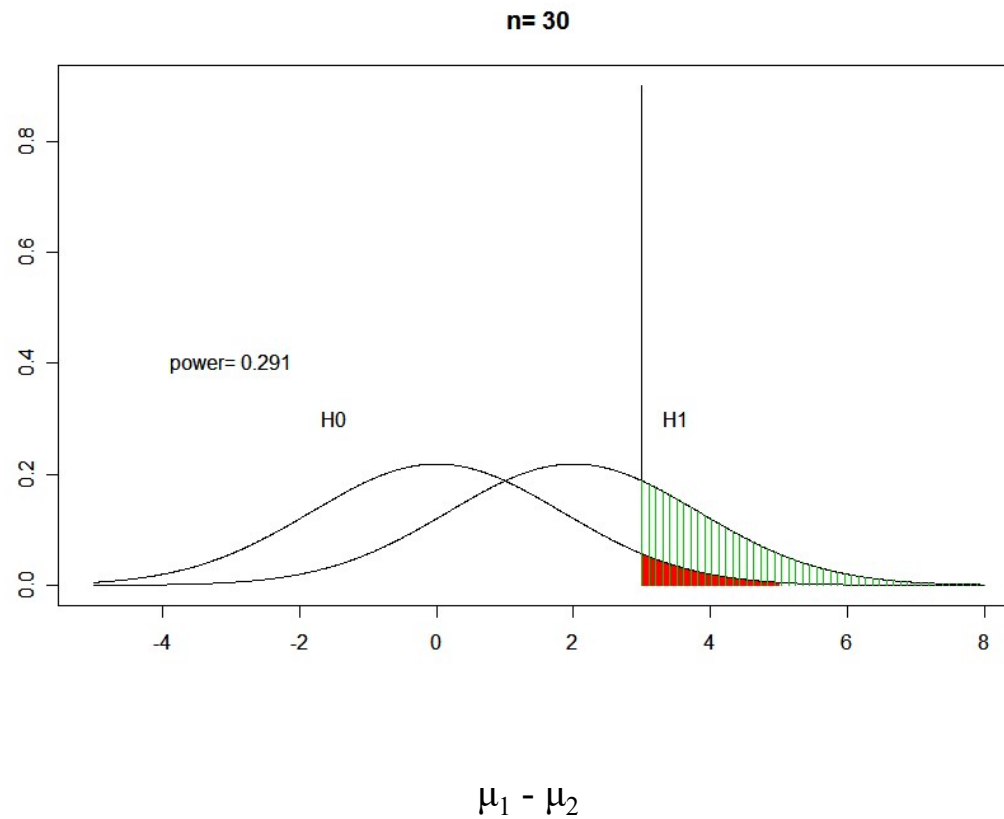


$$\Delta = \ln(\text{HR}) = \ln(0.77) = -\ln(1-0.23) = -0.26$$

Μέγεθος δείγματος=30, Ισχύς=29%

Sampling distn under H_0 : $\mu_1 - \mu_2 = 0$

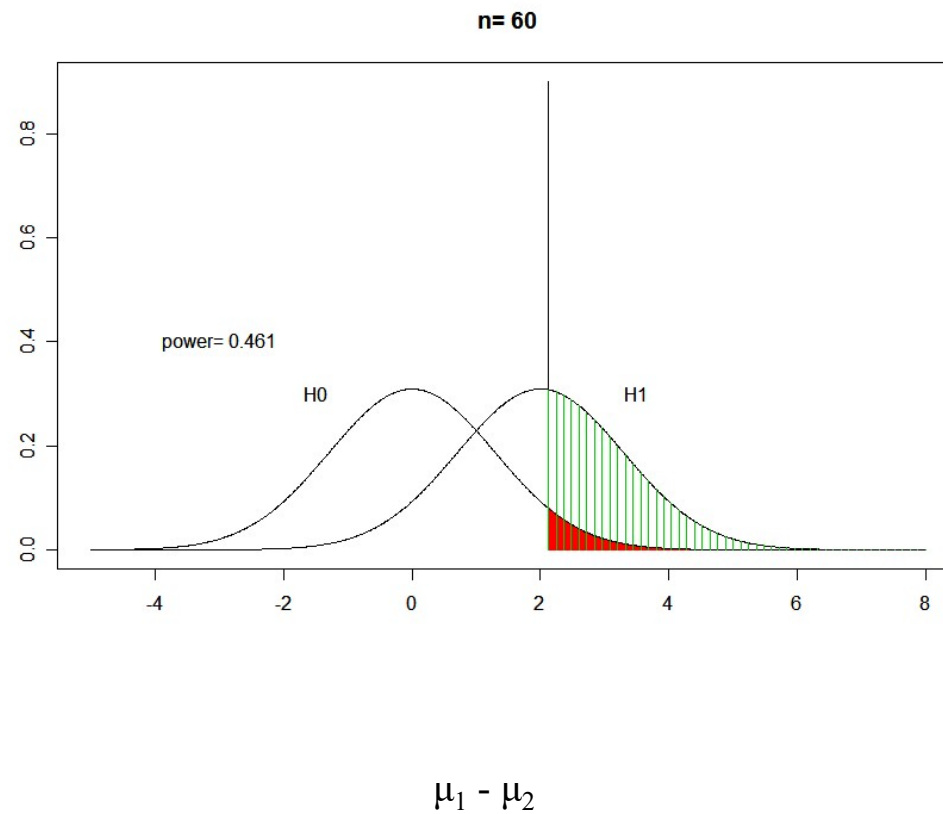
Sampling distn under H_1 : $\mu_1 - \mu_2 = 2$



Μέγεθος δείγματος=60, Ισχύς=46%

Sampling distn under H_0 : $\mu_1 - \mu_2 = 0$

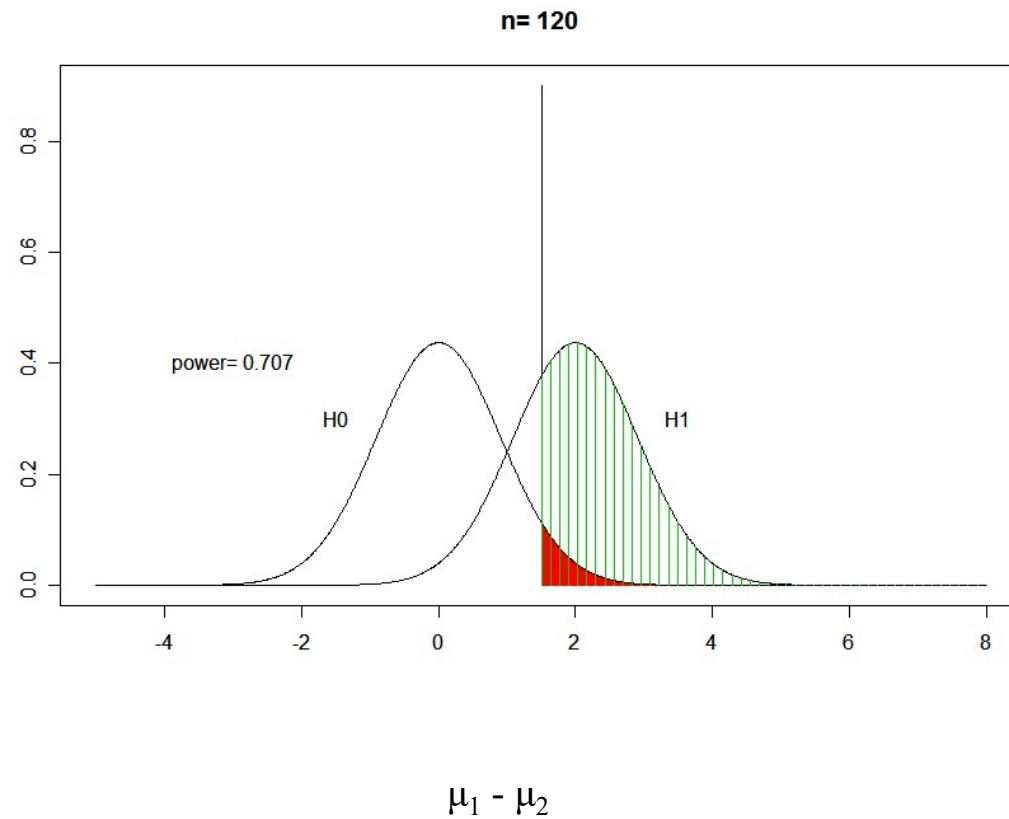
Sampling distn under H_1 : $\mu_1 - \mu_2 = 2$



Μέγεθος δείγματος=120, Ισχύς=71%

Sampling distn under $H_0: \mu_1 - \mu_2 = 0$

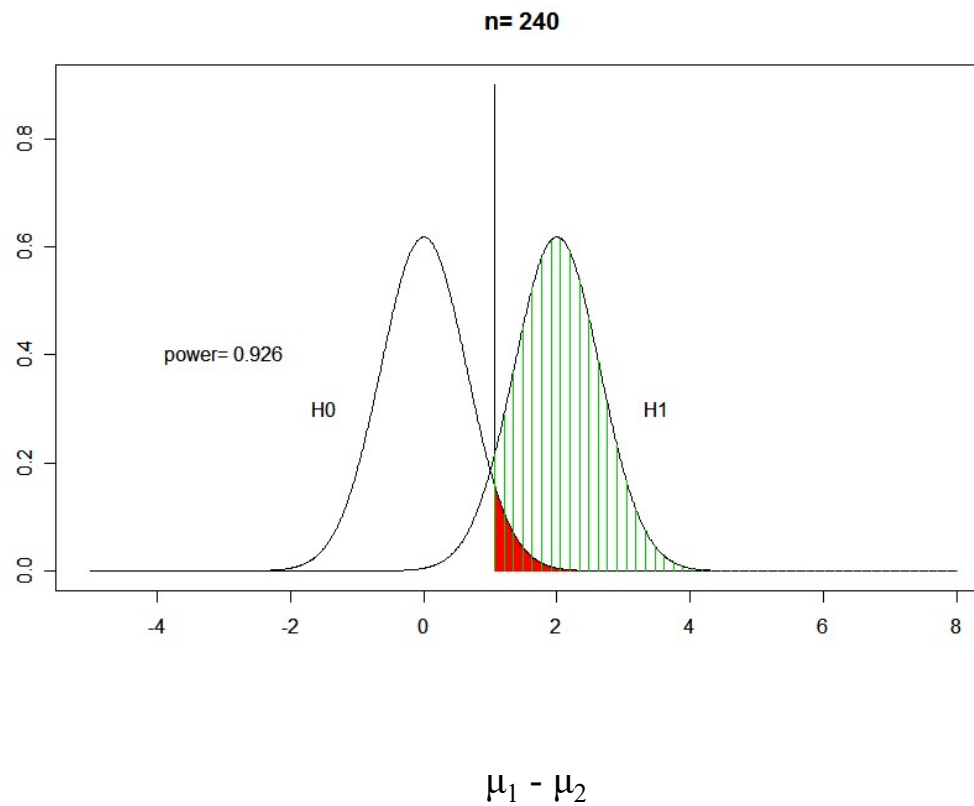
Sampling distn under $H_1: \mu_1 - \mu_2 = 2$



Μέγεθος δείγματος=240, Ισχύς=93%

Sampling distn under H_0 : $\mu_1 - \mu_2 = 0$

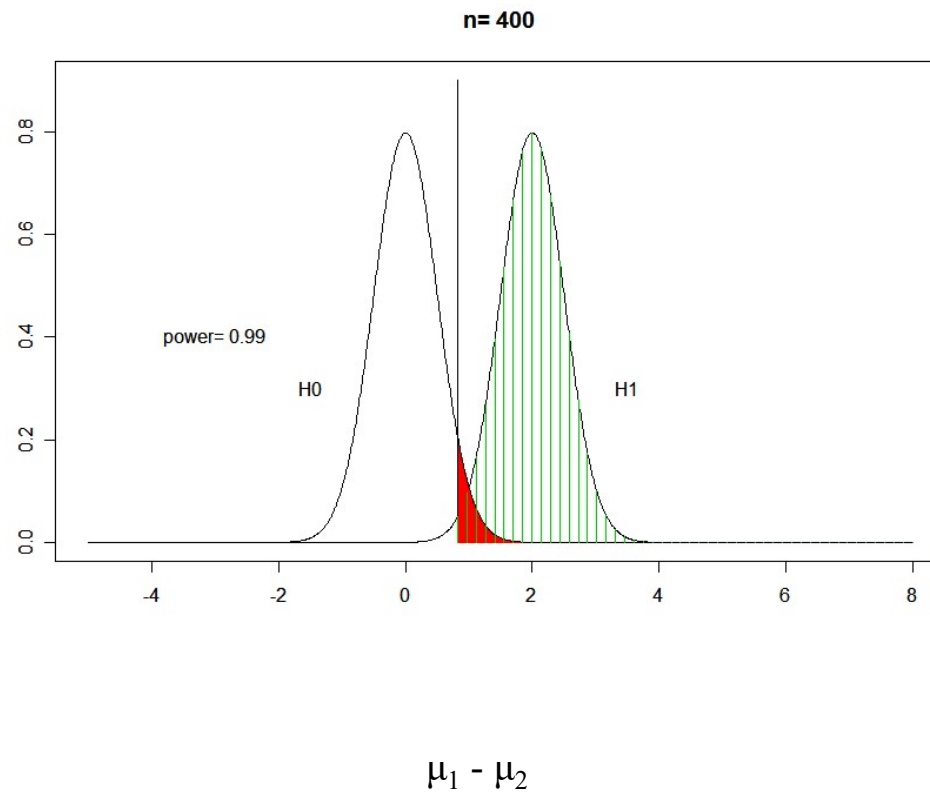
Sampling distn under H_1 : $\mu_1 - \mu_2 = 2$



Μέγεθος δείγματος=400, Ισχύς > 99%

Sampling distn under H_0 : $\mu_1 - \mu_2 = 0$

Sampling distn under H_1 : $\mu_1 - \mu_2 = 2$



Μέγεθος Δείγματος

- Ο πειραματισμός με περισσότερα άτομα από τον αναγκαίο αριθμό παραβιάζει τα ατομικά ανθρώπινα δικαιώματα των ασθενών
- Ο πειραματισμός με λιγότερα άτομα από τον αναγκαίο αριθμό παραβιάζει τα δικαιώματα των ασθενών στην κοινότητα

Μέγεθος δείγματος

- **Επαρκές δείγμα:** Μέγεθος Μελέτης ικανό για να αναγνωριστεί με μεγάλη πιθανότητα (Στατιστική Ισχύς) η ελαχίστη κλινικά σημαντική διαφορά ως στατιστικά σημαντική
 - Αύξηση αξιοπιστίας, εγκυρότητας
 - Μείωση δειγματοληπτικού, συστηματικού και τυχαίου σφάλματος
- **Αντιπροσωπευτικό δείγμα**

Μέγεθος Δείγματος

Επομένως για να σχεδιάσουμε μια μελέτη με σκοπό να δούμε αν υπάρχει κάποια κλινική διαφορά (έστω τ) ανάμεσα σε δύο θεραπείες το πόσο μεγάλη είναι αυτή διαφορά, και

- τις πιθανότητες σφαλμάτων που «επιτρέπουμε» να γίνουν (τύπου I και II)

$$n = \frac{2\sigma^2}{\tau^2} \left(z_{\alpha/2} + z_{\beta} \right)^2 ,$$

4 αλληλοεξαρτώμενες ποσότητες: α , β , τ , N

- α = ρυθμός ψευδώς θετικών (σφάλμα τύπου I)
- β = ρυθμός ψευδώς αρνητικών (σφάλμα τύπου II)
- τ = «η ελάχιστη κλινικά σημαντική διαφορά»
- N = το μέγεθος του δείγματος.

Δεδομένης της διαδικασίας ελέγχου, οι ποσότητες αυτές αλληλεξαρτώνται π.χ.

- τ αυξάνεται $\rightarrow$ πιο εύκολα το ανακαλύπτουμε $\rightarrow$ το $1-\beta$ αυξάνεται.

Κύριο ερώτημα – μεταβλητή απόκρισης

- Το κύριο ερώτημα θα πρέπει να είναι κάτι που μπορεί να παρατηρηθεί και να μετρηθεί με μια μεταβλητή, που λέγεται μεταβλητή απόκρισης (response variable).
- Παραδείγματα μεταβλητών απόκρισης είναι:
 - επίπτωση μιας ασθένειας, θνησιμότητα-νοσηρότητα, αριθμός υποτροπών,
 - συνολική επιβίωση, χρόνος ως την υποτροπή, χρόνος ως την παρουσίαση μιας ασθένειας,
 - διαστολική πίεση, CD4 counts, ποιότητα ζωής, βιολογικοί δείκτες (biomarkers) κ.α.

Γενικοί κανόνες για τη μεταβλητή απόκρισης

- Η μεταβλητή απόκρισης θα πρέπει να μπορεί να παρατηρηθεί και να καταγραφεί για όλους τους ασθενείς
- Η μεταβλητή θα πρέπει να αξιολογηθεί αμερόληπτα
Πχ. πώς καταγράφεται ο πόνος αμερόληπτα αν θεωρήσουμε ότι ήταν η μεταβλητή απόκρισης;

Είδη μεταβλητής απόκρισης

- **Συνεχείς μετρήσεις:** πίεση του αίματος.
- **Δίτιμα ή διχότομα:** υποτροπή εντός 30 ημερών.
- **Ονομαστικά:** Το αποτέλεσμα ανήκει σε μια κατηγορία από ένα σύνολο μη διατεταγμένων κατηγοριών (ιστολογία).
- **Διατεταγμένα:** Το αποτέλεσμα ανήκει σε μια κατηγορία από ένα σύνολο διατεταγμένων κατηγοριών (τοξικότητα).
- **Σύνολα:** Ο αριθμός των δερματικών στιγμάτων.
- **Επιβίωση ή χρόνος ως το γεγονός:** το καταληκτικό σημείο (*endpoint*) στις περισσότερες μελέτες θανατηφόρων ασθενειών.
- Ανάλογα με το είδος της μεταβλητής καθορίζεται και η στατιστική μεθοδολογία που θα χρησιμοποιηθεί για την ανάλυση.

Υπολογισμός μεγέθους δείγματος

- Σχετικά απλός για απλές περιπτώσεις
- Σύγχρονες κλινικές δοκιμές είναι εξαιρετικά πιο πολύπλοκες
- Απαιτητικό και ιδιαίτερα πιο δύσκολοι υπολογισμοί
- Πολλές φορές εφικτοί μόνο με προσομοίωση ή χρήση σεναρίων για να βρεθεί η ευαισθησία τους

Μέγεθος Δείγματος

- Είδαμε πριν ότι

$$n = \frac{2\sigma^2}{\tau^2} \left(z_{\alpha/2} + z_{\beta} \right)^2,$$

- Στην πράξη δεν γνωρίζουμε τη διακύμανση από πριν!
- Αν υποθέσουμε κάτι λανθασμένο είτε θα βρεθούμε με μεγαλύτερο δείγμα από αυτό που θα έπρεπε ή με μικρότερο
- Πως μπορούμε να διορθώσουμε?
- Sample size re-estimation approach

Μέγεθος Δείγματος

Σε πιο πολύπλοκους σχεδιασμούς δεν είναι ευκολο να βρούμε το μέγεθος με έναν απλό τύπο

Πχ πόσους ασθενείς χρειαζόμαστε για να βρούμε σε μια μελέτη ποια γονίδια συνδέονται με καλύτερη επίδραση ενός φαρμάκου/θεραπείας?

Πολύ συχνά (ολοένα πιο συχνά) καταφεύγουμε σε μεθόδους Monte Carlo για να υπολογίσουμε κάτι τέτοιο με όσο πιο ρεαλιστικές συνθήκες μπορούμε

Τυχαιοποίηση

- Στη δοκιμή υπάρχει η νέα θεραπεία και συγκρίνεται με μια άλλη (πχ. την καλύτερη γνωστή θεραπεία) ή και με εικονική θεραπεία (placebo)
- Το ποια θεραπεία θα πάρει ο κάθε ασθενής προκύπτει τυχαία (πχ με χρήση κάποιου προγράμματος που παράγει τυχαίους αριθμούς)
- **Τυφλή Θεραπεία:**
 - Ο ασθενής δε γνωρίζει ποια θεραπεία πήρε.
 - **Διπλά Τυφλή:** Ούτε ο θεράπων γιατρός γνωρίζει

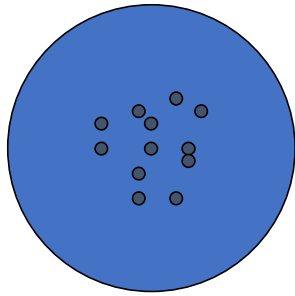
Τυχαιοποίηση

- Με την τυχαιοποίηση δημιουργούνται συγκρίσιμες ομάδες ασθενών. Η τυχαιοποίηση τείνει να εξισορροπήσει τους γνωστούς ή άγνωστους συγχυτικούς παράγοντες στις δύο ομάδες
- Αναιρείται η μεροληψία, ασυνείδητη ή συνειδητή του ερευνητή όταν επιλέγει τις ομάδες (selection bias)
- Εγγυάται ότι οι στατιστικοί έλεγχοι θα έχουν έγκυρα επίπεδα σημαντικότητας

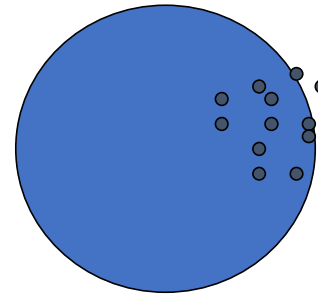
Bias - Variability

Μεροληψία - Μεταβλητότητα

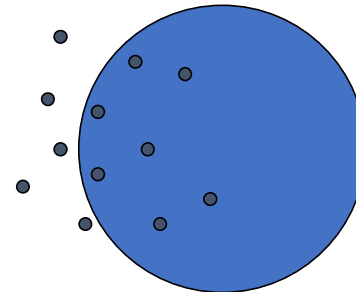
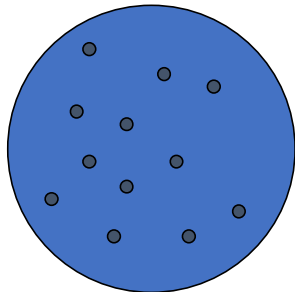
No bias,
small
variability



Bias,
small
variability



No bias,
larger
variability



Bias,
larger
variability

Απλή τυχαιοποίηση

- Ρίψη νομίσματος
 - Κεφάλι $\rightarrow$ Ομάδα Α; Γράμματα $\rightarrow$ Ομάδα Β
- Χρήση πίνακα τυχαίων αριθμών
- Χρήση τυχαίων αριθμών από Η/Υ
- **Στρωματοποίηση:** μερικές φορές η τυχαιοποίηση λαμβάνει υπόψη κάποια χαρακτηριστικά (όχι όμως πολλά) των ασθενών

Τυχαιοποίηση-Στρωματοποίηση

- Με την **τυχαιοποίηση** ελέγχεται:
 - Η μεταβλητότητα του θεραπευτικού αποτελέσματος (variability)
 - Το συστηματικό σφάλμα επιλογής (selection bias)
- Η **στρωματοποίηση** εξισορροπεί γνωστούς παράγοντες κινδύνου μεταξύ των θεραπειών υπό διερεύνηση
 - Η τυχαιοποίηση διεξάγεται εντός κάθε στρώματος

Τυχαιοποίηση

Στην πράξη χρειάζεται να κάνουμε πιο προσεκτικά την τυχαιοποίηση με στρώματα και να σιγουρευτούμε ότι δεν θα υπάρξει σοβαρό Imbalance ανάμεσα στις θεραπείες

Επίσης δεν θέλουμε να διακινδυνεύσουμε ότι κάποιος μπορεί να «μαντέψει» ποιο θα είναι η επιλογή για τον επόμενο ασθενή

Για αυτο το σκοπό χρησιμοποιούμε αρκετά πιο πολύπλοκα σχήματα όπως πχ Block Randomization

Παράδειγμα

Block of 4 – 2 θεραπείες

- AABB, ABBA, ABAB, BABA, BBAA, BAAB
- Διαλέγω τυχαία 20 ασθενείς AABB ABBA AABB ABAB BBAA
- Το μέγεθος των blocks μπορεί να διαφέρει

Adaptive Randomization

- Θέλω να τυχαιοποιώ στους κλάδους που έχουν δείξει καλύτερα αποτελέσματα
- Δηλαδή με βάση τα αποτελέσματα που έχω αλλάζω τις πιθανότητες να τυχαιοποιήσω κάποιον ασθενή σε αυτό τον κλάδο
- Χρήση του θεωρήματος του Bayes ώστε αν επαναπροσδιορίζω τις πιθανότητες σε κάθε κλάδο

Adaptive Designs

Κατά τη διάρκεια της δοκιμής και **με προδιαγεγραμμένο τρόπο** (πότε και πως) διάφορα χαρακτηριστικά της μελέτης μπορούν να αλλάξουν με βάση την πληροφορία που έχει συσσωρευτεί

Μπορούμε να κάνουμε αλλαγές σε

- Μέγεθος δείγματος
- Κλάδους της μελέτης (μπορεί κάποιοι να σταματήσουν)
- Τρόπος τυχαιοποίησης
- Έκβαση
- Ελεγχοσυνάρτηση
- κ.α.



Adaptive Pros

- **Planned flexibility** – change study parameters as trial progresses according to pre-specified algorithms
- **Program efficiency** – reduce overall development time and increase probability of success
- **Optimal sample size** – adjust sample size as study progresses
- **Patient outcomes** – increased exposure to effective doses, decreased exposure to unsafe or ineffective doses
- **Quicker decisions** – stop early for safety, futility, or efficacy



Adaptive Cons

- **Regulatory issues** – need regulatory buy-in for adaptive designs
- **Upfront planning** – modeling and simulation to evaluate design properties
- **Complicated design** – increased start-up cost time
- **Logistical issues** -- more difficult to implement
- **Statistical challenges** – Type 1 error control, confidence interval coverage and treatment effect estimation
- **Interpretation** – adaptation may introduce unexpected bias or variation

Παράδειγμα – two stage design

Simon's Design

Dichotomous outcome (usually in Phase II trial) Suppose that we want to test

$$H_0 : p = 0.25 \text{ versus } H_1 : p = 0.40$$

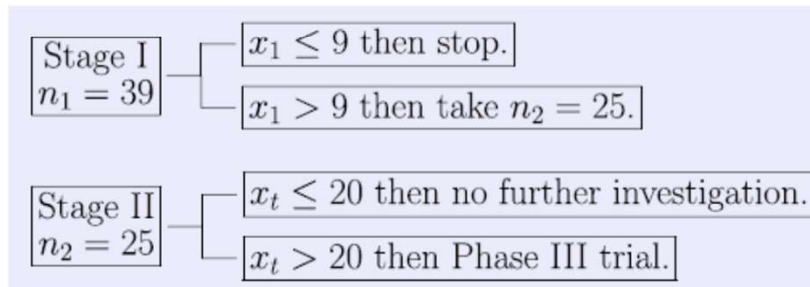
with $\alpha = 0.10$ and $\beta = 0.10$ (power = 0.90).

A conventional single stage design:

$N = 64$ and reject H_0 if $R > 20$, R being the number of responses.

Simon's two-stage design:

- In stage I, $n_1 = 39$ is accrued.
- If 9 or fewer responses are observed during stage I, then the trial is stopped for futility.
- Otherwise additional $n_2 = 25$ is accrued. If 20 or fewer responses are observed by the end of stage II, then no further investigation (i.e, Phase III trial) is warranted.



Crossover

Σε μια κλινική δοκιμή αφενός ενδιαφερόμαστε να

- δουμε αν το νέο φάρμακο είναι καλύτερο
- εκτιμήσουμε το πόσο καλό είναι (treatment effect)

Σε μερικές κλινικές δοκιμές είναι εφικτό τα άτομα να αλλάξουν
θεραπεία κατά τη διάρκεια της μελέτης, είτε λόγω σχεδιασμού είτε για
ηθικούς λόγους

HERA trial

- Γυναίκες με καρκίνο του μαστού (5081 ασθενείς)
- Αρχικά ο σχεδιασμός προέβλεπε 3 arms
 - 1-year trastuzumab
 - 2-years trastuzumab
 - Observation only

Μετα την προβλεπόμενη ενδιάμεση ανάλυση, το trastuzumab είχε μεγάλη επίδραση και αποφασίστηκε για ηθικούς λόγους να επιτραπεί στις υπόλοιπες ασθενείς να αλλάξουν κλάδο!

Στη δεύτερη ανάλυση η επίδραση είχε πια χαθεί (ITT ανάλυση)

Τελικά στο τέλος της μελέτης η επίδραση βρέθηκε αν και μικρή.

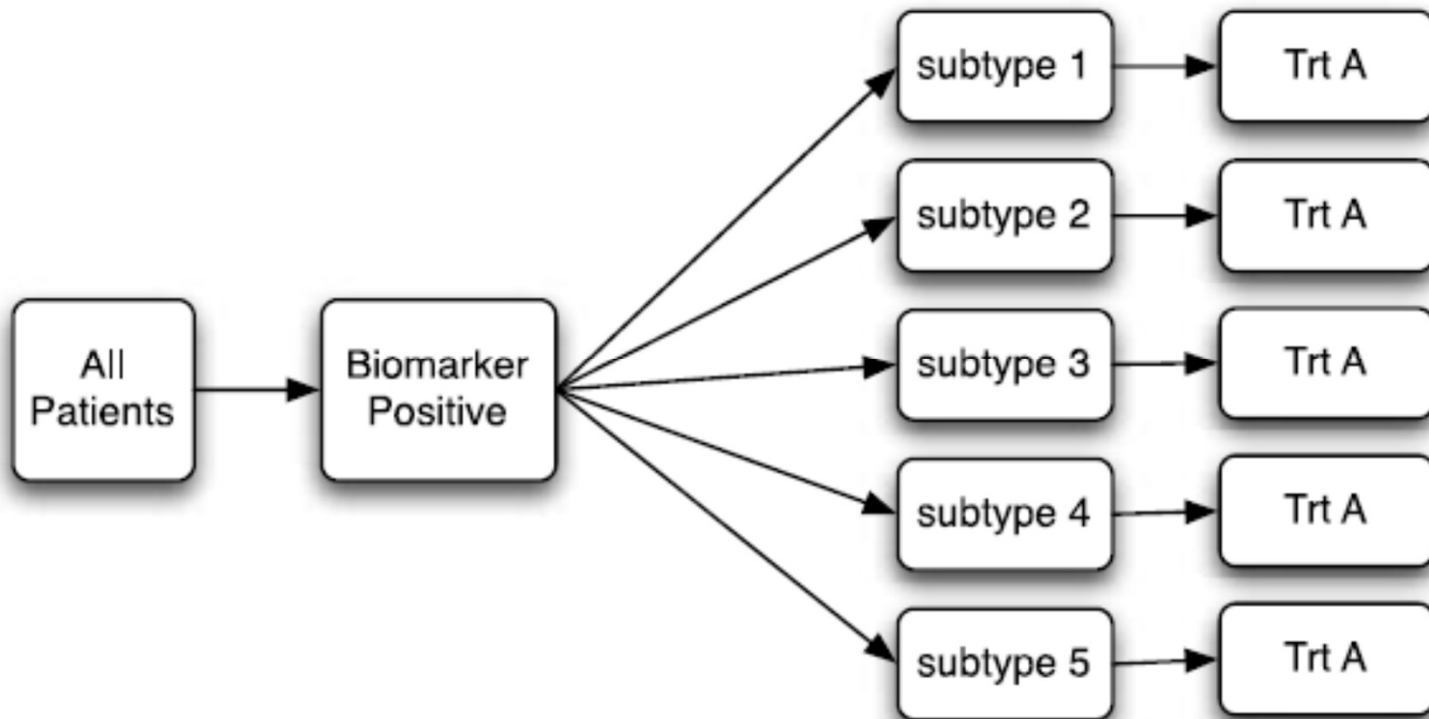
HERA TRIAL

- Το φάρμακο εγκρίθηκε και ήδη χρησιμοποιείτε
- Δεν είναι ακόμα ξεκάθαρο ποια είναι η επίδραση του, δηλαδή πόσο βελτιώνει την επιβίωση των ασθενών
- Υπάρχουν αρκετές μέθοδοι για να αναλυθούν τέτοια δεδομένα και να εκτιμηθεί η επίδραση αλλά ακόμα δεν υπάρχει κάποια μέθοδος που να απαντάει με σιγουριά το ερώτημα!

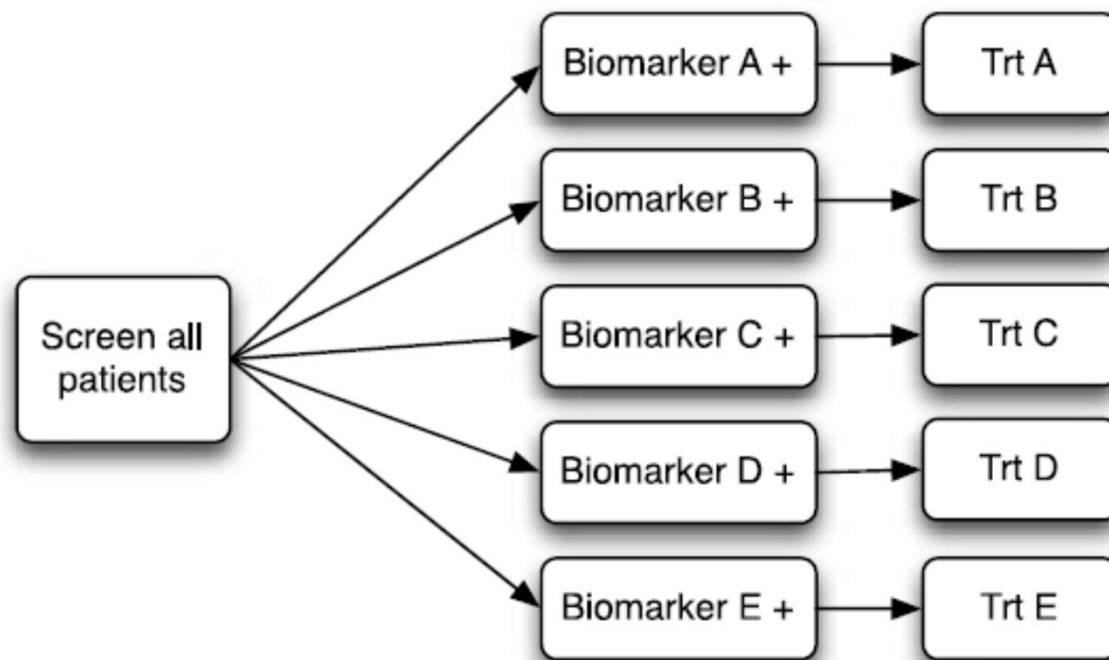
Σύγχρονοι σύνθετοι σχεδιασμοί

- **Basket trial:**
one targeted therapy is tested on more than one disease or subtype of disease
- **Umbrella trial:**
more than one targeted therapy is tested for one or more subtypes of a single disease

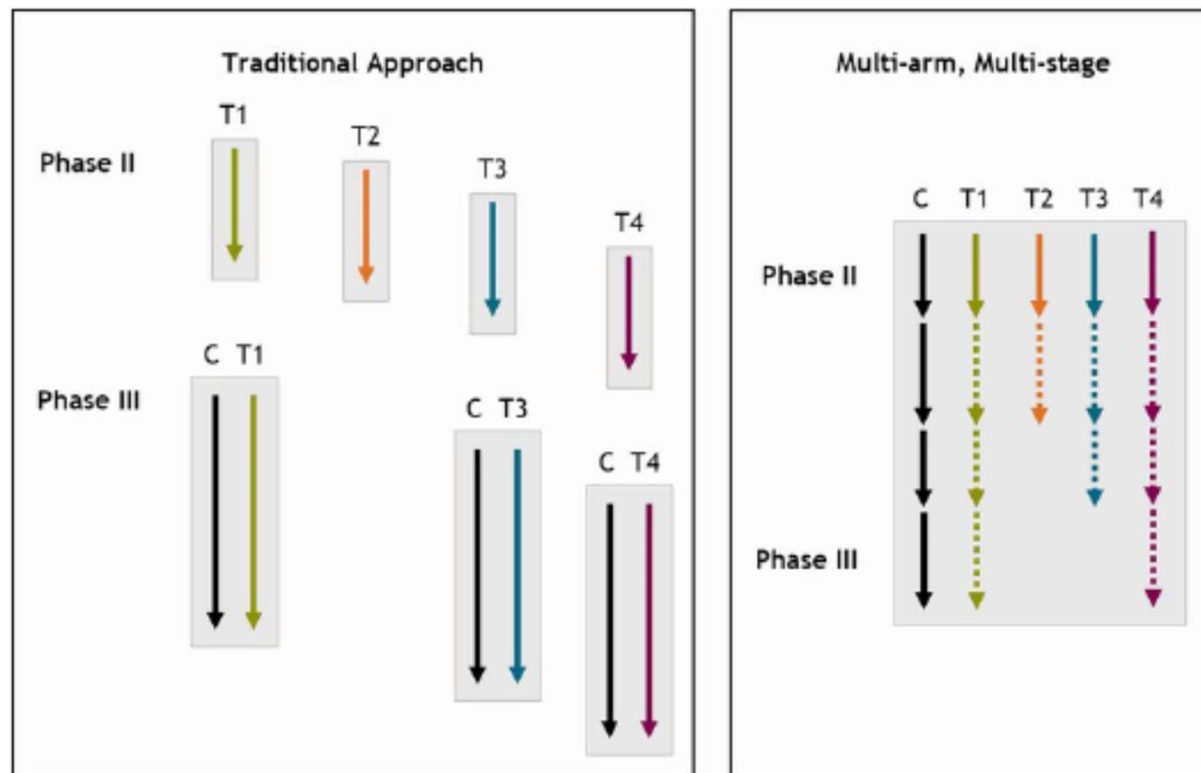
Basket σχεδιασμός



Umbrella σχεδιασμός



Multi Arm – Multi Stage



Μεθοδολογικά προβλήματα

Φανταστείτε τον παρακάτω σχεδιασμό.

- Πολλές επιμέρους ουσίες (η διαφορετικές δόσεις της ίδιας) μελετώνται συγχρόνως.
- Μετά από το πρώτο στάδιο η καλύτερη (ή οι καλύτερες) συνεχίζουν στο επόμενο στάδιο.
- Πως θα ελέγξουμε αν το φάρμακο που προκρίθηκε είναι τελικά κατάλληλο;
- Βλέπετε κάποια προβλήματα;

Παράδειγμα

- Έστω k θεραπείες. Για κάθε μια κάνουμε έναν έλεγχο αν είναι καλύτερη από την placebo της μορφής: $H_0 : \theta_j = 0$ έναντι της εναλλακτικής $H_A : \theta_j < 0$. Έστω η ελεγκοσυνάρτηση για τον έλεγχο αυτό είναι η T_j .
- Αν προκρινουμε μόνο τη καλύτερη από τις k θεραπείες πρακτικά έχουμε διαλέξει την

$$T = \max (T_1, T_2, \dots T_k)$$

- άρα ο έλεγχος που κάνουμε είναι πιο σύνθετος και με άλλη ελεγκοσυνάρτηση και βέβαια άλλη κατανομή κάτω από τη μηδενική!
- Οι επιμέρους έλεγχοι μάλιστα δεν είναι ανεξάρτητοι.

Μεθοδολογικά προβλήματα

- Πολλαπλοί έλεγχοι
- Ο καλύτερος προκρίνεται άρα αυτό που βλέπουμε ίσως είναι καλύτερο από ότι θα περιμέναμε
- Πως διορθώνουμε για αυτά?
- Χρειαζόμαστε πιο πολύπλοκους ελέγχους υποθέσεων για αυτό με τη χρήση πιο σύνθετων μηδενικών υποθέσεων και ελεγκτοσυναρτήσεων

Μεθοδολογικά προβλήματα

- Έστω μια μελέτη που μας ενδιαφέρουν δυο co-primary εκβάσεις
- Πχ η επιβίωση σε όλον τον πληθυσμό και η επιβίωση στις γυναίκες μόνο
- Οι δυο έλεγχοι είναι συσχετισμένοι (γιατί?)
- Πως θα διορθώσουμε κάτι τέτοιο?

Futility - Ματαιότητα

- Για ηθικούς λόγους πρέπει να σταματήσουμε μια έρευνα η οποία διαφαίνεται ότι δεν μπορεί να εκπληρώσει τους στόχους της
- Πχ έστω ότι θέλουμε να δούμε το ποσοστό απόκρισης σε ένα φάρμακο. Μας ενδιαφέρει να δείξουμε ότι το ποσοστό είναι $p > 0.6$ αλλά στην ενδιάμεση ανάλυση με 40 από τους 100 ασθενείς βρήκαμε 5 επιτυχίες άρα ποσοστό $5/40 = 12.5\%$.
- Πόσο πιθανό είναι να δούμε πάνω από 60 στο σύνολο? Άρα θέλουμε 55 τουλάχιστον επιτυχίες στους 60 επόμενους ασθενείς. Πόσο πιθανό είναι κάτι τέτοιο? Αξίζει να εκθέσουμε τους ασθενείς σε κίνδυνο για κάτι που είναι απίθανό (όχι αδυνατο) να συμβεί?

Ενδιάμεσες αναλύσεις – Interim analyses

Σκοπός

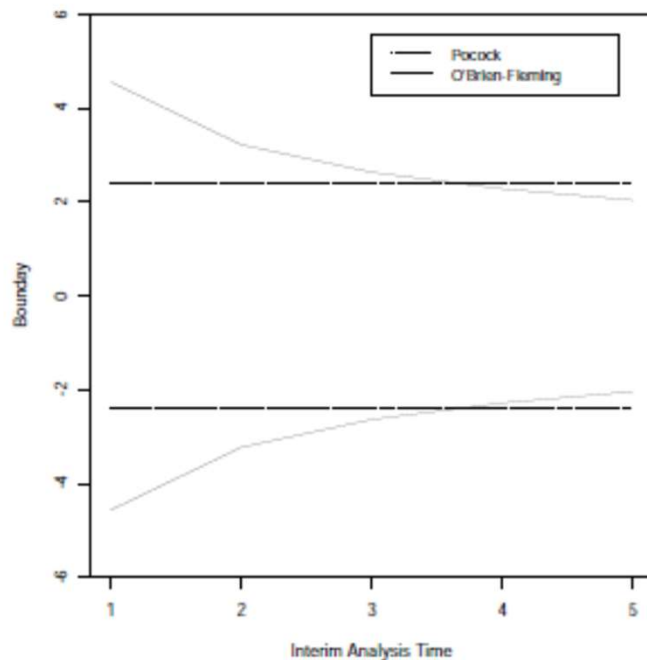
- Να δούμε τα δεδομένα ώστε αφενός μεν να προσαρμόσουμε αν χρειαστεί το σχεδιασμό
- Να σταματήσουμε έγκαιρα αν τα αποτελέσματα είναι πολύ καλά
- Να σταματήσουμε για ματαιότητα αν δεν είναι
- Να απορρίψουμε κλάδους που μοιάζει να μην έχουν επιτυχία ή να επιλέξουμε τον καλύτερο

Κάθε φορά που βλέπουμε τα δεδομένα σπαταλάμε επίπεδο σημαντικότητας (γιατί?)

Ενδιάμεσες αναλύσεις – Interim analyses

- Πως θα μοιράσουμε το α ανάμεσα στις διαφορετικές αναλύσεις?
- Μας συμφέρει να δώσουμε λίγο α στην αρχή καθώς έχουμε πολύ λίγα δεδομένα ακόμα
- Πότε θα ήταν το καλύτερο χρονικό σημείο για να κάνουμε την ανάλυση (αν είναι πολύ νωρίς ίσως δεν έχουμε αρκετή πληροφορία, αν αργήσουμε ίσως δεν κερδίσουμε κάτι)
- Πόσες φορές/αναλύσεις?

Παράδειγμα



Δυο διαφορετικές προσεγγίσεις.

- Στη μια απορρίπτουμε με πολύ πιο ακραία τιμή της ελεγχσυνάρτησης και όσο περνάει η ώρα αυτό μικραίνει (O'Brien-Fleming approach)
- Στο δεύτερο έχουμε σταθερά όρια (Pocock approach) πιο δύσκολα όμως από αυτά που θα είχαμε αν κάναμε έναν μόνο έλεγχο
- Στην πράξη υπάρχουν και άλλες προσεγγίσεις

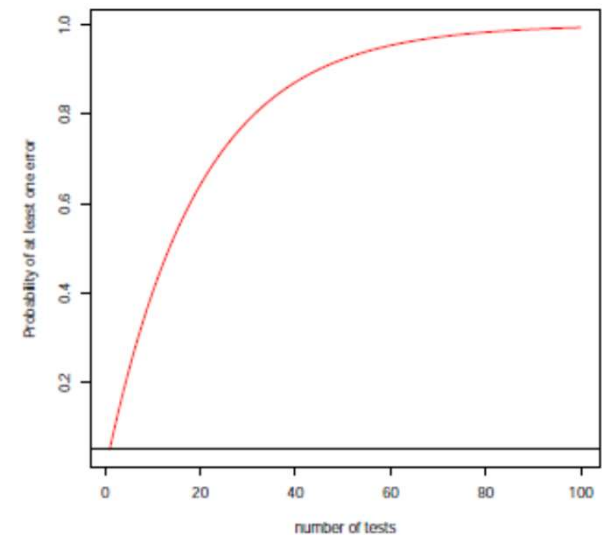
Multiplicity

Το πρόβλημα

Συνήθως κάνουμε έναν έλεγχο υπόθεσης σε $\alpha=5\%$

Αν κάνουμε πολλούς ελέγχους τότε το σφάλμα Τύπου Ι δεν είναι πια 5% αλλά πολύ μεγαλύτερο

$$P(\text{at least one Type I error among } m \text{ tests}) = 1 - (1 - \alpha)^m$$



Γιατί αυτό μας ενδιαφέρει

Subgroup analyses

- Συνήθως σε μια δοκιμή εκτός από το βασικό αποτέλεσμα κοιτάμε και πολλά άλλα σε υποομάδες
- Αν κοιτάξουμε πολλές υποομάδες τότε θα βρούμε κάποια στατιστικά σημαντικά απλά από λάθος
- Πρέπει να περιορίσουμε αυτά τα ψευδώς θετικά αποτελέσματα
- Χρειάζεται κατάλληλη μεθοδολογία

Multiplicity Story

- ISIS2 study: μελέτη για την επίδραση της ασπιρίνης στη θρομβόλυση για την θεραπεία εμφραγμάτων
- Τα αποτελέσματα έδειξαν σημαντικό κέρδος για τους ασθενείς.
- Όμως ανέδειξαν ότι αν το ζώδιο του ασθενούς ήταν Ζυγός ή Δίδυμος τότε η θεραπεία δεν ήταν χρήσιμη!
- Το ίδιο αποτέλεσμα εμφανίστηκε για τους Δίδυμους σε παρόμοια μελέτη μερικά χρόνια μετά.
- Αυτό είναι αποτέλεσμα του προβλήματος της Πολλαπλότητας (multiplicity)

Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

Gordon C S Smith, Jill P Pell

Abstract

Objectives To determine whether parachutes are effective in preventing major trauma related to gravitational challenge.

Design Systematic review of randomised controlled trials.

Data sources: Medline, Web of Science, Embase, and the Cochrane Library databases; appropriate internet sites and citation lists.

Study selection: Studies showing the effects of using a parachute during free fall.

Main outcome measure Death or major trauma, defined as an injury severity score > 15 .

Results We were unable to identify any randomised controlled trials of parachute intervention.

Conclusions As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomised controlled trials. Advocates of evidence based medicine have criticised the adoption of interventions evaluated by using only observational data. We think that everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute.

accepted intervention was a fabric device, secured by strings to a harness worn by the participant and released (either automatically or manually) during free fall with the purpose of limiting the rate of descent. We excluded studies that had no control group.

Definition of outcomes

The major outcomes studied were death or major trauma, defined as an injury severity score greater than 15.⁶

Meta-analysis

Our statistical approach was to assess outcomes in parachute and control groups by odds ratios and quantified the precision of estimates by 95% confidence intervals. We chose the Mantel-Haenszel test to assess heterogeneity, and sensitivity and subgroup analyses and fixed effects weighted regression techniques to explore causes of heterogeneity. We selected a funnel plot to assess publication bias visually and Egger's and Begg's tests to test it quantitatively. Stata software, version 7.0, was the tool for all statistical analyses.

Results

Our search strategy did not find any randomised controlled trials of the parachute.

Department of
Obstetrics and
Gynaecology,
Cambridge
University,
Cambridge
CB2 2QQ

Gordon C S Smith
professor

Department of
Public Health,
Greater Glasgow
NHS Board,
Glasgow G3 8YU
Jill P Pell
consultant

Correspondence to:
G C S Smith
gcscs2@cam.ac.uk

BMJ 2003;327:1459-61



OPEN ACCESS



Check for updates

Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial

Robert W Yeh,¹ Linda R Valsdottir,¹ Michael W Yeh,² Changyu Shen,¹ Daniel B Kramer,¹ Jordan B Strom,¹ Eric A Secemsky,¹ Joanne L Healy,¹ Robert M Domeier,³ Dhruv S Kazi,¹ Brahmajee K Nallamothu⁴ On behalf of the PARACHUTE Investigators

ABSTRACT OBJECTIVE

To determine if using a parachute prevents death or major traumatic injury when jumping from an aircraft.

DESIGN

Randomized controlled trial.

SETTING

Private or commercial aircraft between September 2017 and August 2018.

PARTICIPANTS

92 aircraft passengers aged 18 and over were screened for participation. 23 agreed to be enrolled and were randomized.

INTERVENTION

Jumping from an aircraft (airplane or helicopter) with a parachute versus an empty backpack (unblinded).

MAIN OUTCOME MEASURES

Composite of death or major traumatic injury (defined by an Injury Severity Score over 15) upon impact with the ground measured immediately after landing.

RESULTS

Parachute use did not significantly reduce death or major injury (0% for parachute v 0% for control; $P>0.9$). This finding was consistent across multiple subgroups. Compared with individuals screened but not enrolled, participants included in the study were on aircraft at significantly lower altitude (mean of 0.6 m for participants v mean of 9146 m for non-participants; $P<0.001$) and lower velocity (mean of 0 km/h v mean of 800 km/h; $P<0.001$).

CONCLUSIONS

Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention. However, the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps. When beliefs



OPEN ACCESS



Check for updates

Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial

Robert W Yeh,¹ Linda R Valsdottir,¹ Michael W Yeh,² Changyu Shen,¹ Daniel B Kramer,¹ Jordan B Strom,¹ Eric A Secemsky,¹ Joanne L Healy,¹ Robert M Domeier,³ Dhruv S Kazi,¹ Brahmajee K Nallamothu⁴ On behalf of the PARACHUTE Investigators

WHAT IS ALREADY KNOWN ON THIS TOPIC

Parachutes are routinely used to prevent death or major traumatic injury among individuals jumping from aircraft, but their efficacy is based primarily on biological plausibility and expert opinion

No randomized controlled trials of parachute use have yet been attempted, presumably owing to a lack of equipoise

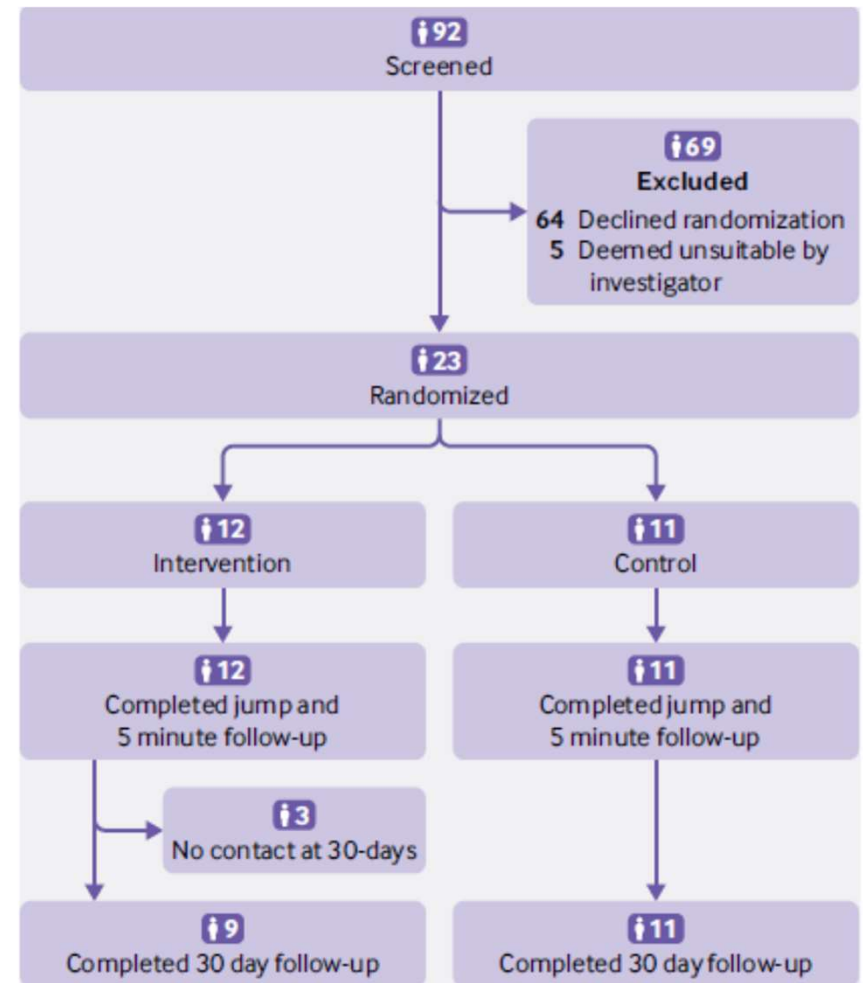
WHAT THIS STUDY ADDS

This randomized trial of parachute use found no reduction in death or major injury compared with individuals jumping from aircraft with an empty backpack

Lack of enrolment of individuals at high risk could have influenced the results of the trial

Statistical analysis

The primary efficacy analysis tested the hypothesis that parachute use is superior to the control in preventing death and major traumatic injury. Based on an assumption of an average jump altitude of 4000 meters (typical of skydiving) and the anticipated effect of impact with the Earth at terminal velocity on human tissue, we projected that 99% of the control arm would experience the primary outcome at ground impact with a relative risk reduction of 95% in the intervention arm. A sample size of 14 (7 in each arm) would yield 99% power to detect this difference at a two sided α of 0.05. In anticipation of potential withdrawal after enrolment owing to last minute anxieties, a total sample size of 20 participants was targeted. Analysis was performed on an intention-to-treat basis. We performed secondary subgroup analyses stratified by aircraft type (airplane v helicopter) and previous parachute use through formal tests of statistical interaction.



Results

Table 3 Event rates for primary and secondary endpoints. Values are numbers (percentages) unless stated otherwise				
Endpoint	Parachute	Control	Mean difference (95% CI)	P value
On impact				
Death or major traumatic injury	0 (0)	0 (0)	0	>0.9
Mean (SD) Injury Severity Score	0 (0)	0 (0)	0	>0.9
30 days after impact				
Death or major traumatic injury	0 (0)	0 (0)	0	>0.9
Mean (SD) Injury Severity Score	0 (0)	0 (0)	0	>0.9
Health status				
Mean (SD) Short Form Health Survey score	43.9 (1.8)	44.0 (2.4)	0.1 (–2.0 to 2.2)	0.9
Mean (SD) physical health subscore	19.6 (0.7)	19.7 (0.5)	0.04 (–0.5 to 0.6)	0.9
Mean (SD) mental health subscore	24.3 (1.3)	24.3 (2.1)	0.08 (–1.6 to 1.8)	0.9

Table 1 | Baseline characteristics of participants randomized to parachute versus control. Values are numbers (percentages) unless stated otherwise

Characteristics	Parachute	Control
Total	11 (100)	12 (100)
Demographics		
Median (SD) age (years)	38.1 (8.7)	38.6 (11.0)
Women	4 (36)	6 (50)
Men	7 (64)	6 (50)
Ethnic group:		
American Indian or Alaska Native	0 (0)	0 (0)
East Asian or South Asian	4 (36)	4 (33)
Black or African American	0 (0)	0 (0)
More than one race	0 (0)	0 (0)
White	7 (64)	8 (67)
Mean (SD) height (cm)	171.8 (9.1)	171.7 (8.4)
Mean (SD) weight (kg)	75.9 (24.4)	74.6 (13.0)
Medical history		
Broken bones	4 (36)	5 (42)
Acrophobia	3 (27)	6 (50)
Parachute use	3 (27)	0 (0)
Family history of parachute use	2 (18)	0 (0)
Frequent flier (average >4 flights per month)	0 (0)	4 (33)
Flight		
International v domestic:		
International	0 (0)	0 (0)
Domestic	11 (100)	12 (100)
Aircraft type:		
Jetliner	0 (0)	0 (0)
Biplane	5 (46)	6 (50)
Helicopter	6 (55)	6 (50)
Mean (SD) velocity (km/h)	0 (0)	0 (0)
Mean (SD) altitude (m)	0.6 (0.1)	0.6 (0.1)

The parachute study

Οι συγγραφείς παραδέχτηκαν ότι αν και προσπάθησαν κανείς δεν δέχτηκε να τυχαιοποιηθεί ενώ το αεροπλάνο πετούσε (γιατι άραγε;)

Πέρα από το αστείο της υπόθεσης η δοκιμή αυτή δείχνει αρκετά χρήσιμα συμπεράσματα σχετικά με τις μελέτες και την ανάγκη για σωστό σχεδιασμό (και μεταξύ μας ο σχεδιασμός είχε τυχαιοποίηση, επαρκή ισχύ και μέγεθος δείγματος).

Τι πήγε στραβά?

Αντί επιλόγου

- Ολοένα και πιο πολύπλοκες και απαιτητικές κλινικές δοκιμές
- Πολλά ενδιαφέροντα νέα θεωρητικά προβλήματα
- Είναι οι τυχαιοποιημένες μελέτες η μόνη απάντηση?
- Την εποχή των **big data** υπάρχει σφοδρό ενδιαφέρον για τη χρήση και μη τυχαιοποιημένων μελετών
- Η πανδημία οδήγησε σε νέες ανάγκες/κατευθύνσεις (αλλά και νέα προβλήματα)