Homework 2 Review:

1. Example 1:

- *Outcome: Prevent HIV seroconversion in women*
- *Intervention: E = vaginal ring with ARV therapy; C = vaginal ring with placebo*

2. Example 2: Select a trial from your own field and evaluate it similarly.

Questions:

1. *How long should the intervention period be? What information will you use to determine this?*
2. *How will you capture the outcome data? Definition of outcome variable, including frequency of collection.*
3. *In what study sample could this trial be conducted ethically? Identify essential eligibility criteria that affect the Intervention-Outcome relationship.*

Required Reading:

Robiner WN. Enhancing adherence in clinical research. Contemp Clin Trials. 2005 Feb; 26(1):59-77.

Solution, Example 1:

For seroconversion to occur in the trial of Example 1, participating women have to have unprotected sex with HIV-infected men on multiple occasions.

- A design strategy to promote this would be to enroll discordant couples: infected men, uninfected women.
- Design features that would minimize the number of seroconversion events include:
    - An ethical constraint, however, requires the trial to counsel participating women to use condoms during intercourse.
    - Promotion of adherence to the study protocol would also encourage use of condoms.

- As seen in today's lecture, sample-size calculations for time-to-event outcomes identify the number of EVENTS that must be observed to detect the specified treatment effect.

Q.1: Enrollment might take longer than expected and the intervention period might need to be extended, depending on the impact of condom use on the overall event rate.

Q.2: Periodic assessments of HIV status are needed during the intervention period. As demonstrated in the Knowler et al trial discussed today, semi-annual assessments are sometimes the compromise frequency used in RCT follow-up. The length of follow-up should be driven by: the frequency of sexual intercourse, the per-act risk of failure to use a condom, and the per-act risk of HIV seroconversion.

Q.3: Essential eligibility criteria include conducting the study in a region of the world where many men are HIV infected.

Lecture 3

Learning Objectives:

Upon completion of this lesson, you should be able to do the following:

- Express a study objective as a hypothesis using statistical notation
- Identify the summary statistics that are relevant to report
- Summarize completeness of outcome ascertainment
- Identify outcomes that are continuous, binary, event times, counts, ordered or unordered categories and repeated measurements.
- Define type-1 and type-2 errors in hypothesis testing framework
- Use software to estimate the number of events required for a logrank comparison of two hazard functions to have (1 - β)% power with given α level

Lecture 3

Response Variables

Example from Homework 1: Knowler WC et al., 2002, N Engl J Med, 346(6):393-403.

The primary outcome was incidence of diabetes, diagnosed (and confirmed by a second test) on the basis of threshold plasma glucose levels following *fasting* or *glucose challenge*:

Plasma glucose levels measured at study visits (Participants also were asked to monitor their blood glucose):

- ≥ 126 mg per deciliter (7.0 mmol per liter) … in the fasting state; semiannually or if symptoms arose, or
- ≥ 200 mg per deciliter (11.1 mmol per liter) … two hours after a 75-g oral glucose load; annually.

Note frequency of measurement reflects practicalities of studying research sample; gives up accuracy of timing of diagnosis to avoid excessive imposition on participants and data-collection costs.

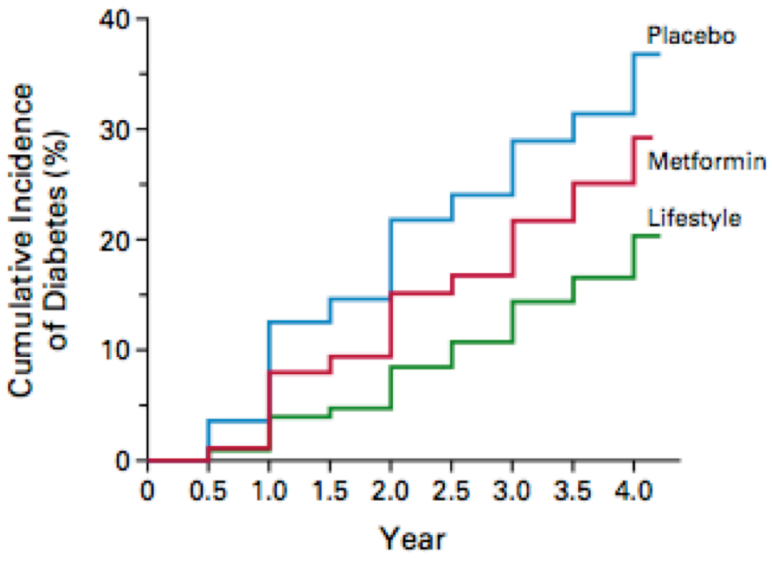


Figure 2. Cumulative Incidence of Diabetes According to Study Group.

Arm-specific outcome estimates: *[N.B. could add 95% CIs]*

- The *crude* incidence was 11.0, 7.8, and 4.8 cases *per 100 person-years* for the placebo, metformin, and lifestyle-intervention arms, respectively.
- The estimated cumulative incidence of diabetes *at three years* was 28.9%, 21.7%, and 14.4% in the placebo, metformin, and lifestyle-intervention arms, respectively.

Between-Arm outcome estimates:

- **[Aim 1]** The incidence of diabetes was 58% lower (95% CI, 48% to 66%) in the lifestyle-intervention arm and 31% lower (95% CI, 17% to 43%) in the metformin arm than in the placebo arm.
- **[Aim 2]** The incidence of diabetes was 39% lower (95% CI, 24% to 51%) in the lifestyle-intervention arm than in the metformin arm.

Completeness of outcome ascertainment:

- 3234 study participants randomized: 1082 to placebo, 1073 to metformin, and 1079 to the intensive lifestyle intervention. → *Why not exactly equal?*
- *Average* follow-up: 2.8 years (range, 1.8 to 4.6). → *Drives maximum time-point of plot above and selection of "3 years" as time-point for outcome estimates.*
- At the close of the study, 99.6% of the participants were alive, of whom 92.5% had attended a scheduled visit within the previous *five* months. → *Spin: Doesn't specifically say what proportion of glucose measurements were analyzable.*

Your turn: Credible? Valid?

Analysis:

"The results of all three pairwise group comparisons were statistically significant by the group-sequential log-rank test." *[from Knowler et al]*

- Collectively: 3 p-values. → *How did they adjust for multiplicity?*

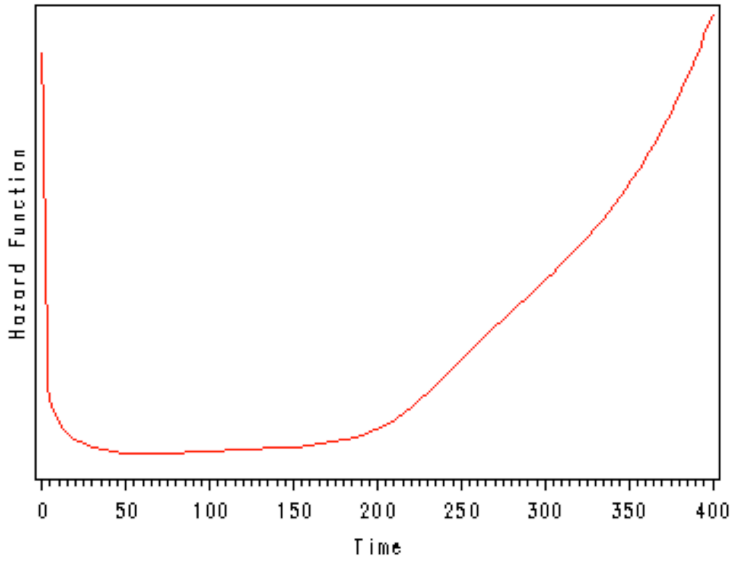
- Testing method: Logrank test
  - This is a longitudinal study design with multiple outcome assessments per participant over time.
  - The outcome is a "time-to-event" (or "survival") endpoint: Participants are assessed repeatedly until the event occurs or follow-up ends (the study ends, or the participant dies or is lost to follow-up).
  - Between-arm estimates: Expressed as % reductions → Parameter that captures treatment effect is a ratio, not a difference.

- "Group-sequential": → Tells us a Data Monitoring Committee evaluated efficacy data during the trial.

Sample-size calculation:

"The study design provided 90% power to detect a 33% reduction from an incidence of 6.5 cases of diabetes per 100 person-years, with a 10% rate of loss to follow-up per year." *[from Knowler et al]*

- Knowler et al hypothesize the same treatment effect for both active-intervention arms so we only need to show one.
- The incidence of the outcome within a very short interval (*i.e., an instant*) is called the hazard of the outcome. (Knowler et al studied the incidence of diabetes within 6-month intervals – relatively crude!)
  - To denote a hazard that changes over time, we denote it by *h(t)* in the sample and $\lambda(t)$ in the population, with *t* ranging from *0* at randomization [not start of treatment] to *T* at the end of the longest follow-up (eg, 4.6 years).

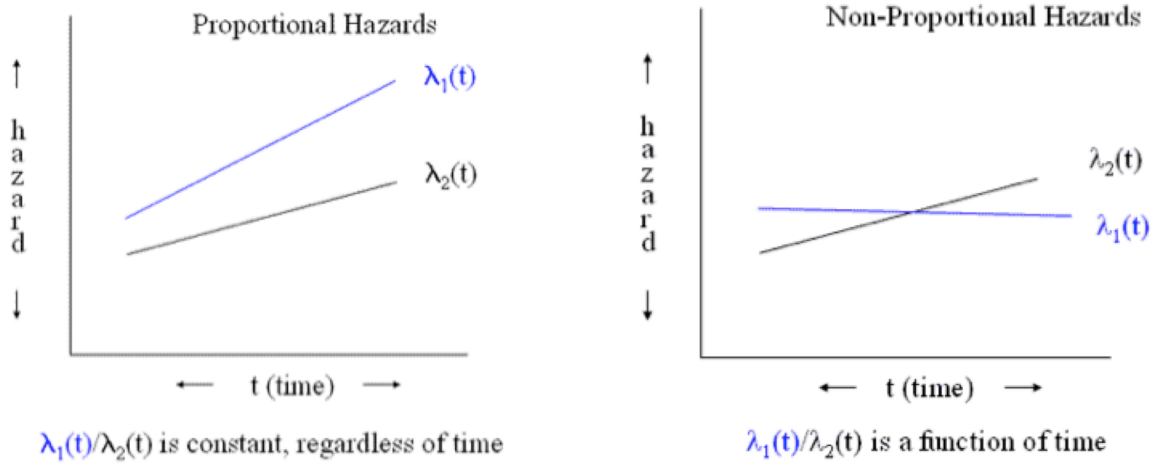


Example left: Heart transplant patients.

(More common to plot the cumulative hazard function because we cannot capture an instantaneous value.)

Proportional Hazards Assumption

- Typically, we assume the hazard ratio is constant throughout the study period, as depicted in left but not right plot ... *[In left-hand plot, which slope would represent the placebo arm of Knowler et al's study?]*



- The logrank test can be applied when one can assume the hazards are proportional.

- In "time-to-event" analyses, we typically estimate the treatment effect via the hazard ratio. N.B. If $\lambda_E(t) / \lambda_P(t) < 1$ then the incidence of diabetes is lower in the experimental arm. *[Note role of absolute value bars in 2-sided tests.]*

$$H_0: | \lambda_E(t) / \lambda_P(t) | = 1 \text{ versus } H_A: | \lambda_E(t) / \lambda_P(t) | \neq 1, \text{ where } t > 0$$

	Intervention Groups		
Group mean	*Placebo* ($N_C =$)	*Experimental* ($N_E =$)	Ratio
Estimated (Sample)	$h_P(t)$	$h_E(t)$	$h_E(t)/h_P(t)$
True (Population)	$\lambda_P(t)$	$\lambda_E(t)$	$\lambda_E(t) / \lambda_P(t)$

- If we assume proportional hazards (equivalently, a (time-)constant hazard ratio), then we don't express the HR as a function of time:

$$HR = h_E(t)/h_P(t).$$

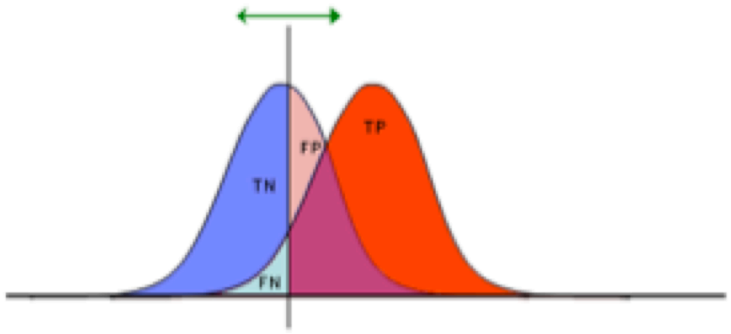
At the end of the trial, we want to be able to draw a firm conclusion: "reject H_0" or "do not reject H_0." However, our conclusion will be based on data, and errors can occur.

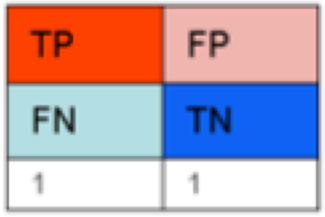
Statistical test	*Two possible true states*	
	H_A: $\lvert \lambda_E(t) / \lambda_P(t) \rvert \neq 1$	H_0: $\lvert \lambda_E(t) / \lambda_P(t) \rvert = 1$
HR ≠ 1: Reject H_0 … "positive trial"	*TP* = 1-β	*FP* = α
HR = 1: Do not reject H_0 … "negative trial"	*FN* = β	*TN* = 1-α

- When H_0 is true: *Specificity* = 1-α = Pr{Do no reject H_0, when H_0 is true}
- When H_A is true: *Sensitivity* = 1-β = power = Pr{Reject H_0, when H_0 is false}

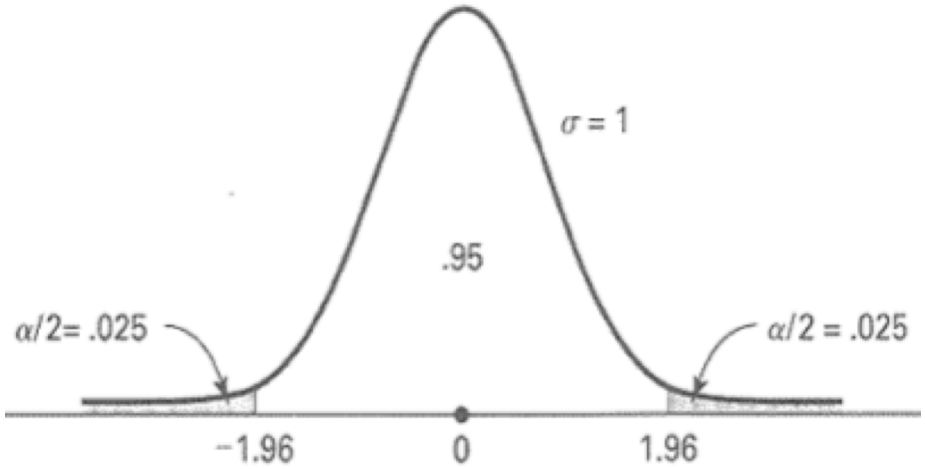
Sample-size calculations are approximate. Many make use of the normal distribution.

- Typical values of the type-1 error rate, α, are 0.05 and 0.01, and of the type-2 error rate, β, are 0.1 and 0.2.
  At β=0.10: $z_{1-\beta}$ = 1.28. At α=0.05: For two-sided test, $z_{1-\alpha/2}$ = 1.96 and for one-sided, $z_{1-\alpha}$ = 1.645.
  At β=0.20: $z_{1-\beta}$ = 0.84. At α=0.01: For two-sided test, $z_{1-\alpha/2}$ = 2.58 and for one-sided, $z_{1-\alpha}$ = 2.33.





The distribution of the data when H_0 is true (left) and when H_A is true (right).



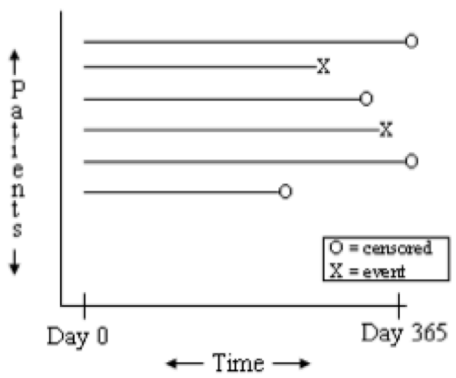
Standard normal probability density function (mean = 0, standard deviation = 1).

- Analogously to the hazard function, we usually work with cumulative probabilities.

Sample-size calculations address ***precision*** rather than validity or bias.

- We want a precise estimate (i.e., a narrow 95% CI) for the treatment effect
- We want adequate power to detect a *clinically relevant* treatment effect
  - Clinically relevant is NOT what has been found previously, although that is a guide.
  - Clinically relevant is a treatment effect that would change practice
- The best estimate of a standard deviation comes from prior research.

Types of Outcomes



Time-to-event outcomes are common in medical research because we often design studies with

- staggered enrollment (and randomization) dates
- a fixed end-of-study date (say, after 5 years)
- → together these result in unequal follow-up times across participants

During f'up of a randomized participant, if the study outcome (eg, diagnosis of diabetes)

- was not observed, his/her outcome value is right-censored – represented by "o"
- was observed, the outcome value is not right-censored – represented by "x".

- Reasons for right censoring include: death; loss to f'up due to competing events, relocation, loss of motivation to participate; end of trial.

- Left censoring and interval also occur. **Homework 3:** Define and cite examples.

To analyze (right) censored event-time data, requires

- **Unique to each participant:** randomization date, follow-up dates, binary status at intermediate follow-up times, censoring status at final follow-up time-point → *Keep all of the details! Don't overwrite earlier data with later data!*
- **Common to all:** start of follow-up time-point (eg, randomization); frequency of follow-up times; end of follow-up (eg, trial termination). → As depicted above, for analysis, modify the dataset to examine follow-up time from common starting point.

Other outcomes (and their distributions) commonly used in clinical trials include:

Continuous measurements	Normal (or normalizable)
Counts	Poisson
Binary endpoints	Binomial
Unordered categories	Multinomial
Ordered categories	Ordered Multinomial

Sample-size procedures depend on the type of outcome variable. Usually we convert the outcome distribution to "standard normal."

Recall that we are interested in summarizing the results of an RCT both within-arm and between arms (see p.3).

Examples of population summaries we want to estimate:

- Within-arm: μ_E and μ_C ; Between-arm: $\delta = \mu_E - \mu_C$
- Within-arm: π_E and π_C ; Between-arm: $\delta = \pi_E - \pi_C$ or $OR = [\pi_E(1-\pi_C)] / [\pi_C(1-\pi_E)]$ or $RR = \pi_E / \pi_C$
- Within-arm: $\lambda_E(t)$ and $\lambda_P(t)$; Between-arm: $HR = \lambda_E(t) / \lambda_P(t)$

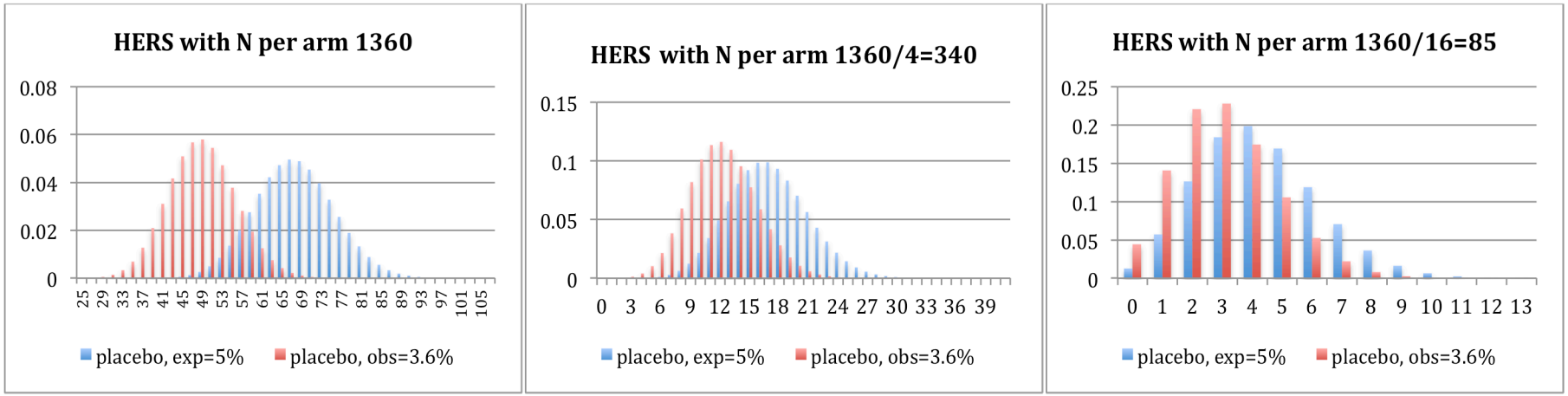
Example: In the *HERS Trial*, about 1360 patients were allocated to the placebo arm.

- At the start of the trial, the incidence of CHD *placebo* events was expected to be $\pi_P = 5.0\%$ per year (blue).
- At the end of the trial, the incidence of CHD *placebo* events was observed to be $p_P = 3.6\%$ per year (red).

Does it matter if the study design was "off" by 5.0% – 3.6% = 1.4% in its estimate of the reference rate?

Yes! The hypothesized difference was $\delta = \pi_E - \pi_P$ = 4.0% – 5.0% = -1.0% !!!

Probability distributions of CHD events, for 2 event rates at 3 sample sizes. {X,Y}-axes give {number of events; prob. of number of events}.



Other points: Plots show that

- … binomial distributions can be approximated by (symmetrical) normal distributions less well as N decreases → When N is small, need to use "exact" method.
- … Standard deviation depends on the rate, π: It's largest when $\pi = 0.50$ (... or 50%).

Example: Try to replicate sample-size calculation for event-time outcome of *Knowler et al*:

- If $h_P(t) = h_P = 0.065$, then the proportion diagnosed within 3 years is $1 - e^{-(0.065)*3} = 0.177$. (Median time to failure is $-\ln(0.5)/h_P = 10.7$ years.) → constant hazard in placebo arm
- Treatment effect: hypothesized 33% reduction → constant Hazard Ratio (HR)

- Given assumptions above, $h_E(t) = h_E = 0.0433$, then the proportion diagnosed within 3 years is $1 - e^{-(0.0433)*3} = 0.122$. (Median time to failure is $-\ln(0.5)/h_E = 16.0$ years.)

- $HR = \frac{0.0433}{0.065} = 0.67$ → $lnHR = -0.406$. $z_{1-\alpha/2} = 1.96$ and $z_{1-\beta} = 1.28$.

- Let $\gamma = n_E / n_P$ be the ratio of the sample sizes of two groups (experimental, placebo). Then if $\gamma = 1$:

$$x = \frac{(\gamma+1)^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\gamma(\ln HR)^2} = \frac{(1+1)^2 (1.96 + 1.28)^2}{(-0.406)^2} = 255 \text{ events (ie, cases of diabetes)}$$

and

$n_P = x/(\gamma\, h_E(t) + h_P(t)) = 255/(0.177 + 0.122) = 853$ and $n_E = \gamma\, n_P = 853$ participants.

Note: These per-arm sample sizes are smaller than they enrolled. In addition, the 3-year failure rates are lower than those reported: 28.9%, 21.7%, and 14.4% in the placebo, metformin, and lifestyle-intervention arms, respectively. These are associated with similar E:P hazard ratios to the value above (0.67) but higher hazards – which may explain why the trial was able to stop early:

Arm	*3-year Incidence*	*Median t-t-failure*	*(constant) Hazard*	*Hazard Ratio*	*% Reduction*	*N per arm*
Placebo	28.9%	6.1 yrs	0.114	*Ref*	*Ref*	*Ref*
Metformin	21.7%	8.5 yrs	0.082	0.717	24.9%	753
Lifestyle	14.4%	13.4 yrs	0.052	0.456	50.2%	160

Homework 3: Replicate this calculation using STATA *stpower* command (or SAS *proc power* or SPSS analog).

Primary outcome is closely related to changes in mean plasma glucose levels over time *[Knowler et al]:*

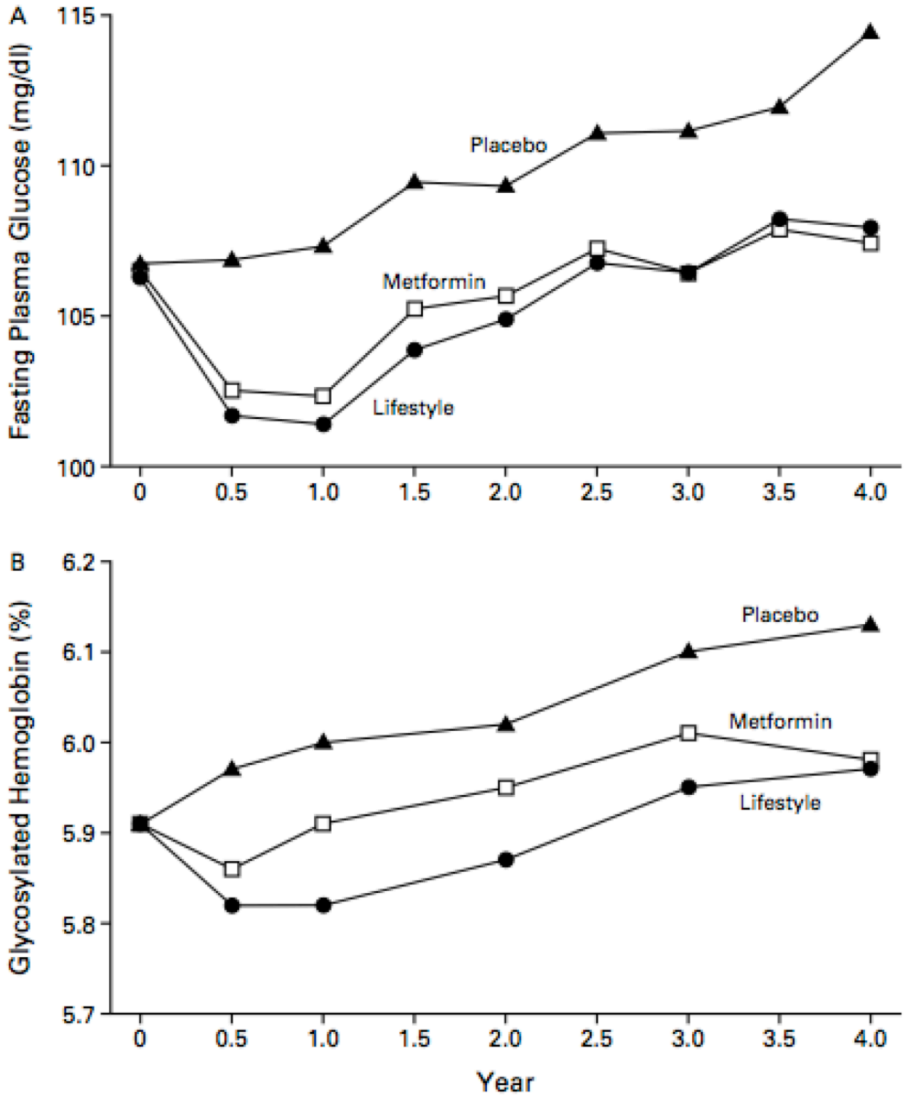


Figure 3. Fasting Plasma Glucose Concentrations (Panel A) and Glycosylated Hemoglobin Values (Panel B) According to Study Group.

The analysis included all participants, whether or not diabetes had been diagnosed. Changes in fasting glucose values over time in the three groups differed significantly ($P<0.001$). Glycosylated hemoglobin values in the three groups differed significantly from 0.5 to 3 years ($P<0.001$). To convert the values for glucose to millimoles per liter, multiply by 0.05551.

- By design, no participants had normal post-load glucose values at baseline.

- Metformin and Lifestyle had similar effects on fasting glucose levels. *[How could plasma glucose level been used as the primary outcome for sample-size calculation?]*

One should be able to see the relationships between mean glucose levels and incidence of diabetes; however, this plot does not exclude participants following diagnosis of diabetes.

- Metformin had less effect than Lifestyle on glycosylated hemoglobin levels

N.B.: Often the primary outcome (eg, clinical status; person-level) changes in reflection of changes in an intermediate outcome (eg, a drug target; cellular level).

One is secondarily interested in the effect of therapy on that outcome.