Segment 2: Randomized Studies

Section 04: Some Takeaways from Randomized Designs

Design vs. Analysis

- We have focused on features of the *design* of a randomized study
  - Discussed design decisions can impact key properties of causal estimates
  - Unbiasedness, efficiency
- Our discussion of *analysis* of the data from a particular study has been limited
  - Comparison of means (possibly within blocks)
  - Regression adjustment (more later)
  - No explicit standard error calculation
  - Some illustration of intuition with *randomization-based* inference
  - Key analysis consideration: **blocking/adjustment** for pre-treatment covariates thought to predict Y can reduce variance

Key Concept: (Conditional) Ignorability

$$Z \perp\!\!\!\perp Y^c, Y^t | \mathbf{X}$$

- Assignment to treatment does not depend on "the science" of the potential outcomes
  - (Within strata of X), treatment is not systematically assigned based on anticipated effect
- Unconfounded
- *Balanced* potential outcomes, on average, across randomizations
  - Does not guarantee balance for any *single* randomization
- Will be a key consideration in *observational* studies where the assignment mechanism is **unknown**
  - Ignorability may hold by *assumption*

Key Concepts: Unbiasedness/Efficiency

Key Goal: Create treatment/control groups that are as as similar as possible in Y^c, Y^t

- Need "balance" in potential outcomes to ensure *unbiasedness*
  - Balance on $\mathbf{X}$ can be a "proxy" for this
- Also need to think about how likely"unlucky" randomizations
  - Can try to minimize the chance of "unluckly" randomizations in the design
- Higher chance of groups similar on $Y^c, Y^t \rightarrow$ closer the estimate will be on average to the causal estimand $\rightarrow$ *efficiency*
  - The more similar the units, the smaller the variance of the estimate
  - The fewer "unlucky" randomizations, the smaller the variance of the estimate

Key Concept: Blocking

Conducting a completely randomized study within *blocks* of units defined by values of $\mathbf{X}$ is a key strategy for improving design

- Create subgroups (blocks, based on $\mathbf{X}$) that are expected to have similar (Y^c, Y^t)
- Block-specific homogeneity $\rightarrow$ reduced variance of effect estimates
- Need to account for the blocking in the analysis!
  - Analyze outcomes within block
  - Average block-specific estimates

Blocking → Adjustment

Blocking can be generalized to the concept of *adjusting* for pre-treatment covariates

- Consider a regression model fit to randomized study
- Leverages association between X and Y^c and/or Y^t to create homogeneity within strata defined by levels of $\mathbf{X}$.
- Mimics the idea of having blocked on the pre-treatment covariates
    - Like including a block for every possible value of $\mathbf{X}$, then positing some (parametric) relationship for how Y varies across blocks
- Has the effect of reducing variance of estimates across

Example: Hypothetical Dietary Experiment

Observed Data

Table: Observed Data from the Hypothetical Dietary Experiment, **Idealized Assignment**

Unit, i	Female, x_{1i}	Age, x_{2i}	Treatment Z_i	Potential Y_i^c	Potential Y_i^t
Audrey	1	40	0	140	135
Anna	1	40	1	140	135
Bob	0	50	0	150	140
Bill	0	50	1	150	140
Caitlin	1	60	0	160	155
Cara	1	60	1	160	155
Dave	0	70	0	170	160
Doug	0	70	1	170	160

Regression Model # 1:

Regression Model # 2:

Example: Hypothetical Dietary Experiment

Observed Data, Regression with/without covariate adjustment

Regression Model # 1: $E[Y_i] = \alpha^{(1)} + \tau^{(1)} Z_i$
Regression Model # 2: $E[Y_i] = \alpha^{(2)} + \tau^{(2)} Z_i + \beta_1^{(2)} X_{i1} + \beta_2^{(2)} X_{i2}$

- Balance is perfect
  - Point estimate $\hat{\tau}^{(1)} = \hat{\tau}^{(2)} = -7.5$
- Analyzing as completely randomized gives SE($\hat{\tau}^{(1)}$) $=$ 8.8
  - Accounting for “chance” of unlucky/poor balance with the completely randomized design
- Analyzing with covariate adjustment gives SE($\hat{\tau}^{(2)}$) $=$ 1.2
  - Does not consider the possibility of poor balance due to the “blocking” on age/sex

Adjusting for pre-treatment variables that predict $Y \rightarrow$ creates homogeneity in Y^t, Y^c for units with similar $\mathbf{X} \rightarrow$ taking advantage of a “naturally arising” randomized block design

Key Concept: Internal vs. External Validity

- **Internal Validity:** Ability to recover causal effects averaging over the sample in the study
  - A design/analysis with internal validity should be good at recovering “the truth” as defined in the population and the study conditions represented in the study sample
  - Results are valid with respect to the particular environment of the study
  - Randomized studies are a “gold standard” for internally valid causal inference

- **External Validity:** Ability to recover causal effects in environments or populations *other* than that represented by the study
  - Effects measured in a particular environment may or may not correspond to what would happen in other environments
  - Randomized studies may or may not hold external validity

Why Randomized Studies May Not Be Externally Valid

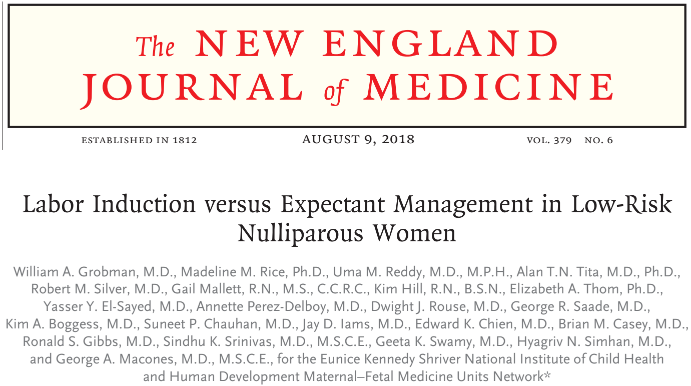
In general, the specifics of the experiment may not correspond to how the world works "in the wild"

- The units in the experiment do not represent the general population of interest
  - Only some units are eligible for participation in the experiment
  - Only some units voluntarily participate in the experiment
- The conditions of the experiment do not reflect how the treatment would be applied in general
  - People behave differently when enrolled in an experiment
  - Treatment in a controlled environment does not look the same as treatment in general settings

Examples of Threats to External Validity

- Email a customer base inviting them to participate in an online survey where they will be randomly assigned some questions about one of two products
  - Not everyone responds to the invitation to participate
  - Inference will be *internally valid* for responders
  - Effect estimate of randomization to product may not be *externally valid* for the entire customer base
- Early COVID vaccine trials enrolled patients who were > 18 years old and had no other serious illnesses
  - Results may not be *externally valid* to people ≤ 18 or with other serious illness
- Students are enrolled in a study where a teacher will administer a randomly assigned experimental reading training or a standard reading instruction
  - Results may not be *externally valid* to widespread deployment of the training when it is to be administered by *parents* instead of *teachers*

Example: Labor Induction



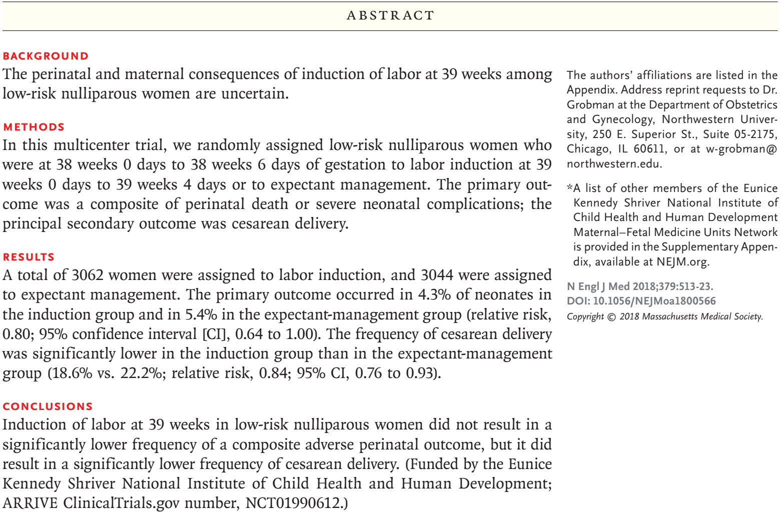
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 AUGUST 9, 2018 VOL. 379 NO. 6

Labor Induction versus Expectant Management in Low-Risk Nulliparous Women

William A. Grobman, M.D., Madeline M. Rice, Ph.D., Uma M. Reddy, M.D., M.P.H., Alan T.N. Tita, M.D., Ph.D., Robert M. Silver, M.D., Gail Mallett, R.N., M.S., C.C.R.C., Kim Hill, R.N., B.S.N., Elizabeth A. Thom, Ph.D., Yasser Y. El-Sayed, M.D., Annette Perez-Delboy, M.D., Dwight J. Rouse, M.D., George R. Saade, M.D., Kim A. Boggess, M.D., Suneet P. Chauhan, M.D., Jay D. Iams, M.D., Edward K. Chien, M.D., Brian M. Casey, M.D., Ronald S. Gibbs, M.D., Sindhu K. Srinivas, M.D., M.S.C.E., Geeta K. Swamy, M.D., Hyagriv N. Simhan, M.D., and George A. Macones, M.D., M.S.C.E., for the Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network*

Example: Labor Induction



ABSTRACT

BACKGROUND

The perinatal and maternal consequences of induction of labor at 39 weeks among low-risk nulliparous women are uncertain.

METHODS

In this multicenter trial, we randomly assigned low-risk nulliparous women who were at 38 weeks 0 days to 38 weeks 6 days of gestation to labor induction at 39 weeks 0 days to 39 weeks 4 days or to expectant management. The primary outcome was a composite of perinatal death or severe neonatal complications; the principal secondary outcome was cesarean delivery.

RESULTS

A total of 3062 women were assigned to labor induction, and 3044 were assigned to expectant management. The primary outcome occurred in 4.3% of neonates in the induction group and in 5.4% in the expectant-management group (relative risk, 0.80; 95% confidence interval [CI], 0.64 to 1.00). The frequency of cesarean delivery was significantly lower in the induction group than in the expectant-management group (18.6% vs. 22.2%; relative risk, 0.84; 95% CI, 0.76 to 0.93).

CONCLUSIONS

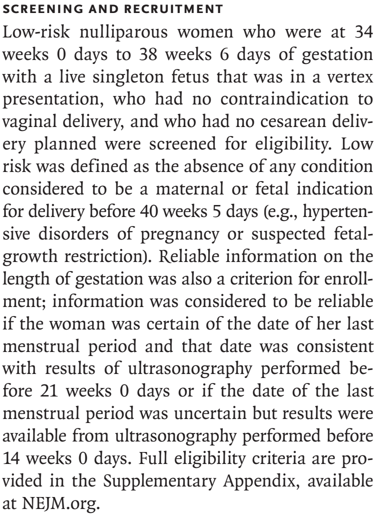
Induction of labor at 39 weeks in low-risk nulliparous women did not result in a significantly lower frequency of a composite adverse perinatal outcome, but it did result in a significantly lower frequency of cesarean delivery. (Funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development; ARRIVE ClinicalTrials.gov number, NCT01990612.)

The authors' affiliations are listed in the Appendix. Address reprint requests to Dr. Grobman at the Department of Obstetrics and Gynecology, Northwestern University, 250 E. Superior St., Suite 05-2175, Chicago, IL 60611, or at w-grobman@northwestern.edu.

*A list of other members of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Maternal–Fetal Medicine Units Network is provided in the Supplementary Appendix, available at NEJM.org.

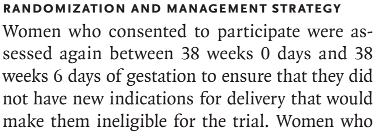
N Engl J Med 2018;379:513-23.
DOI: 10.1056/NEJMoa1800566

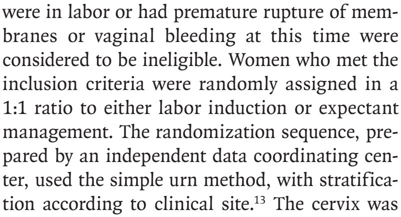
Example: Labor Induction



SCREENING AND RECRUITMENT

Low-risk nulliparous women who were at 34 weeks 0 days to 38 weeks 6 days of gestation with a live singleton fetus that was in a vertex presentation, who had no contraindication to vaginal delivery, and who had no cesarean delivery planned were screened for eligibility. Low risk was defined as the absence of any condition considered to be a maternal or fetal indication for delivery before 40 weeks 5 days (e.g., hypertensive disorders of pregnancy or suspected fetal-growth restriction). Reliable information on the length of gestation was also a criterion for enrollment; information was considered to be reliable if the woman was certain of the date of her last menstrual period and that date was consistent with results of ultrasonography performed before 21 weeks 0 days or if the date of the last menstrual period was uncertain but results were available from ultrasonography performed before 14 weeks 0 days. Full eligibility criteria are provided in the Supplementary Appendix, available at NEJM.org.

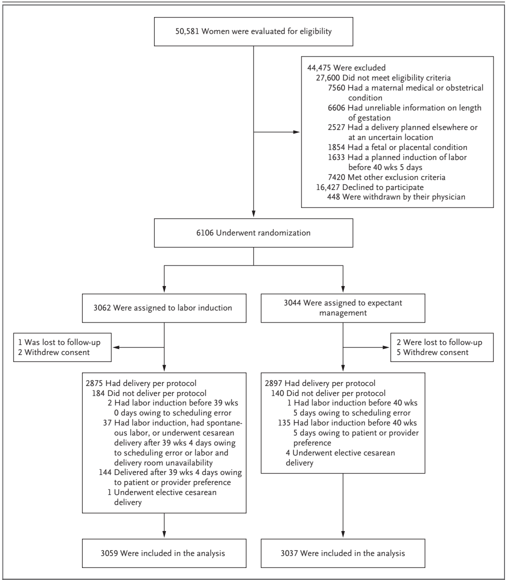




RANDOMIZATION AND MANAGEMENT STRATEGY

Women who consented to participate were assessed again between 38 weeks 0 days and 38 weeks 6 days of gestation to ensure that they did not have new indications for delivery that would make them ineligible for the trial. Women who were in labor or had premature rupture of membranes or vaginal bleeding at this time were considered to be ineligible. Women who met the inclusion criteria were randomly assigned in a 1:1 ratio to either labor induction or expectant management. The randomization sequence, prepared by an independent data coordinating center, used the simple urn method, with stratification according to clinical site.[13] The cervix was

Example: Labor Induction



50,581 Women were evaluated for eligibility

44,475 Were excluded
- 27,600 Did not meet eligibility criteria
  - 7560 Had a maternal medical or obstetrical condition
  - 6606 Had unreliable information on length of gestation
  - 2527 Had a delivery planned elsewhere or at an uncertain location
  - 1854 Had a fetal or placental condition
  - 1633 Had a planned induction of labor before 40 wks 5 days
  - 7420 Met other exclusion criteria
- 16,427 Declined to participate
  - 448 Were withdrawn by their physician

6106 Underwent randomization

3062 Were assigned to labor induction

1 Was lost to follow-up
2 Withdrew consent

2875 Had delivery per protocol
- 184 Did not deliver per protocol
  - 2 Had labor induction before 39 wks 0 days owing to scheduling error
  - 37 Had labor induction, had spontaneous labor, or underwent cesarean delivery after 39 wks 4 days owing to scheduling error or labor and delivery room unavailability
  - 144 Delivered after 39 wks 4 days owing to patient or provider preference
  - 1 Underwent elective cesarean delivery

3059 Were included in the analysis

3044 Were assigned to expectant management

2 Were lost to follow-up
5 Withdrew consent

2897 Had delivery per protocol
- 140 Did not deliver per protocol
  - 1 Had labor induction before 40 wks 5 days owing to scheduling error
  - 135 Had labor induction before 40 wks 5 days owing to patient or provider preference
  - 4 Underwent elective cesarean delivery

3037 Were included in the analysis

Figure 1. **Eligibility, Randomization, Delivery, and Assessment.**

Per-protocol delivery in the induction group was defined as electively induced labor from 39 weeks 0 days to 39 weeks 4 days or spontaneous labor or medically indicated delivery on or before 39 weeks 4 days (this also included delivery

Balancing Internal vs. External Validity

Key Question: To whom do the results of my experiment *generalize*?

- Does the sample included in my experiment represent the population I care about?
  - If not exactly, do I believe the results from my experiment are relevant to the broader population of interest?
- E.g., In the ARRIVE pregnancy trial, goal was to make inference in low-risk nulliparous women in 38^{th} week of gestation
  - *Many* other criteria for inclusion
  - Implications for general practice?
- Frequently a balance between:
  - Practical considerations
  - Need for good information on all units
  - Desire to isolate the effect of interest as much as possible
  - Generalizability beyond the confined study sample