Lesson 17: Missing Data in Human Subjects Research

Nicky Wakim

June 2, 2025

Road Map for today

Missing Data is unavoidable

- Various reasons for missing data in human subjects research
  - Respondents may skip any or all questions in a survey
  - Loss of follow-up in longitudinal studies
  - Missed appointments

- Overall, medical research downplays missing data
  - Rarely happens in a vacuum
  - Can be linked to other variables or our outcome

Biggest issues with current practices

- Missing data not mentioned in any results

- Some imputation of missing data was performed but not discussed
  - Imputation = filling in missing data

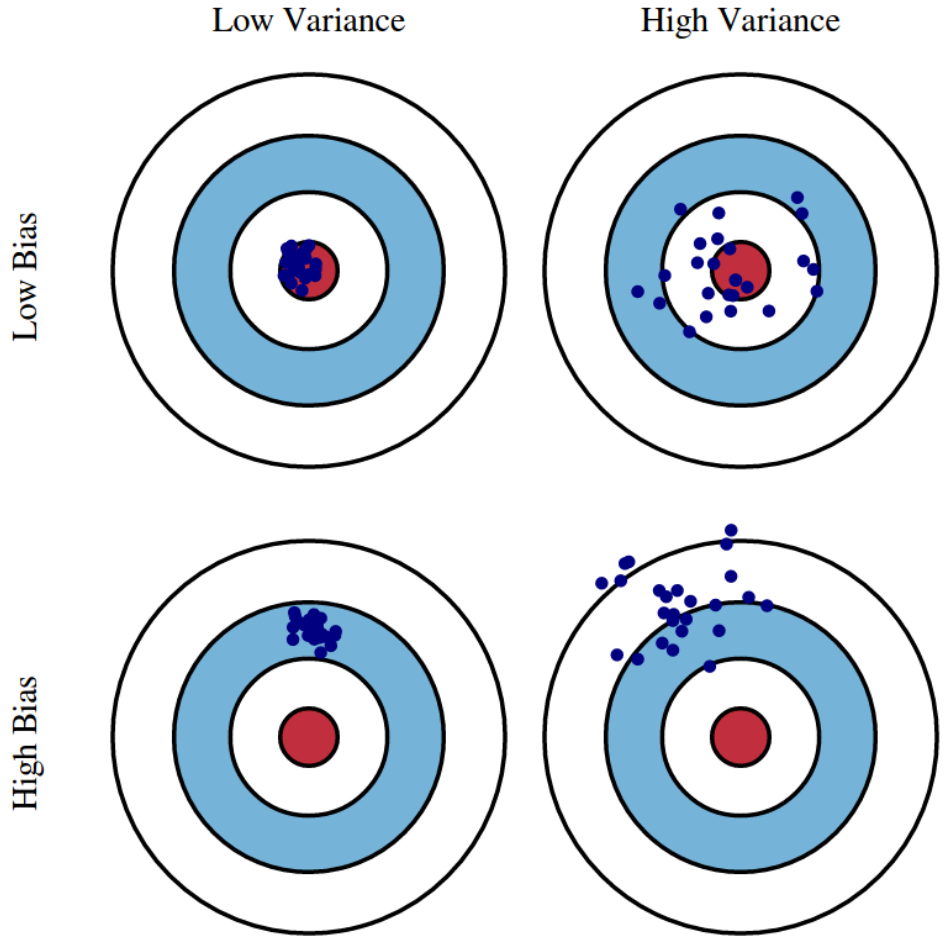
- Research has different tables with different sample sizes
  - No clear presentation of the sample size used in final analysis

Consequences of missing data

- Most researchers simply drop observations with any missing data

- Reduces the power of your analysis
  - Smaller sample leads to less power
  - Harder to prove differences

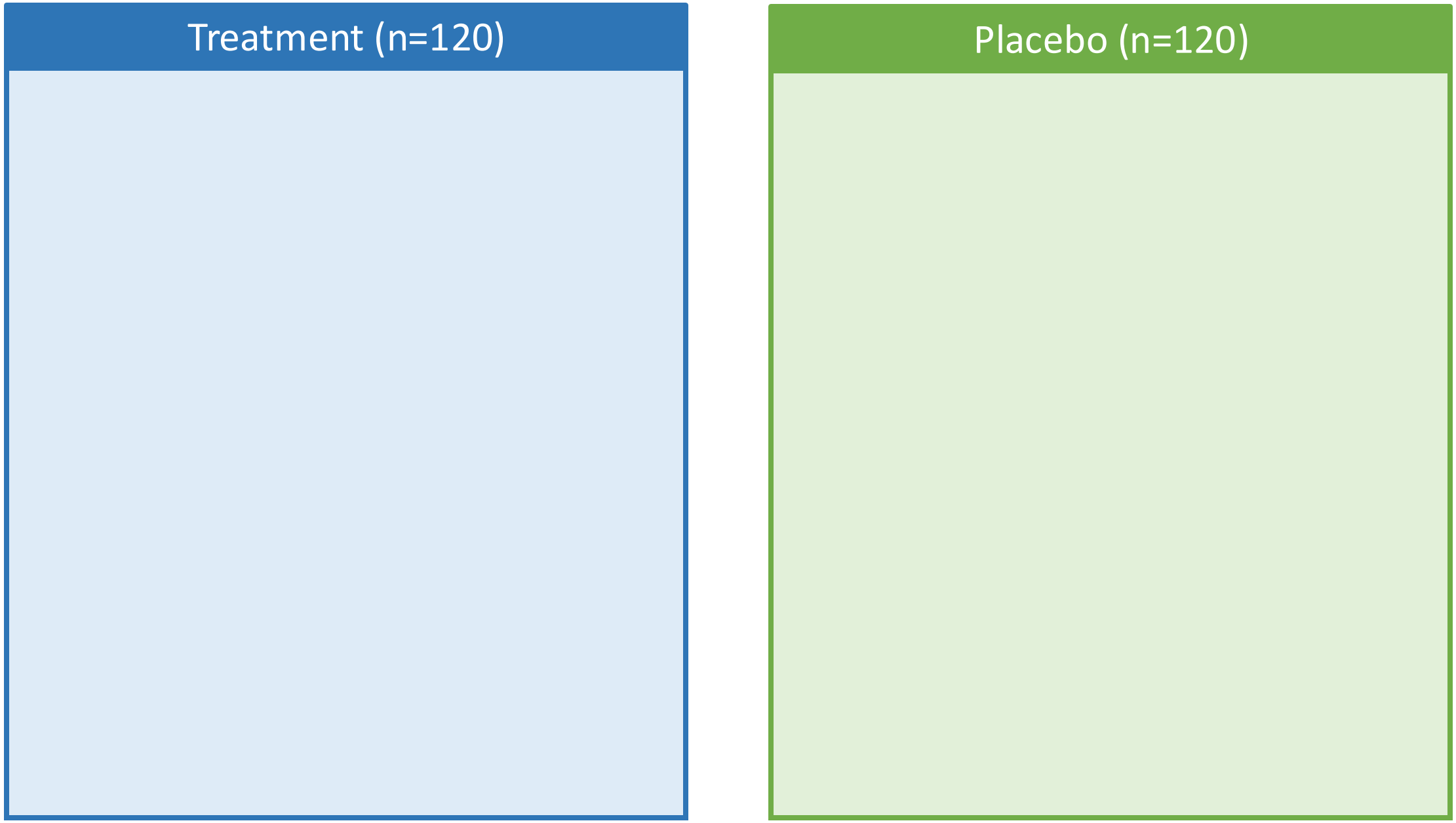
- Can lead to biased results
  - Can overestimate or underestimate the true results



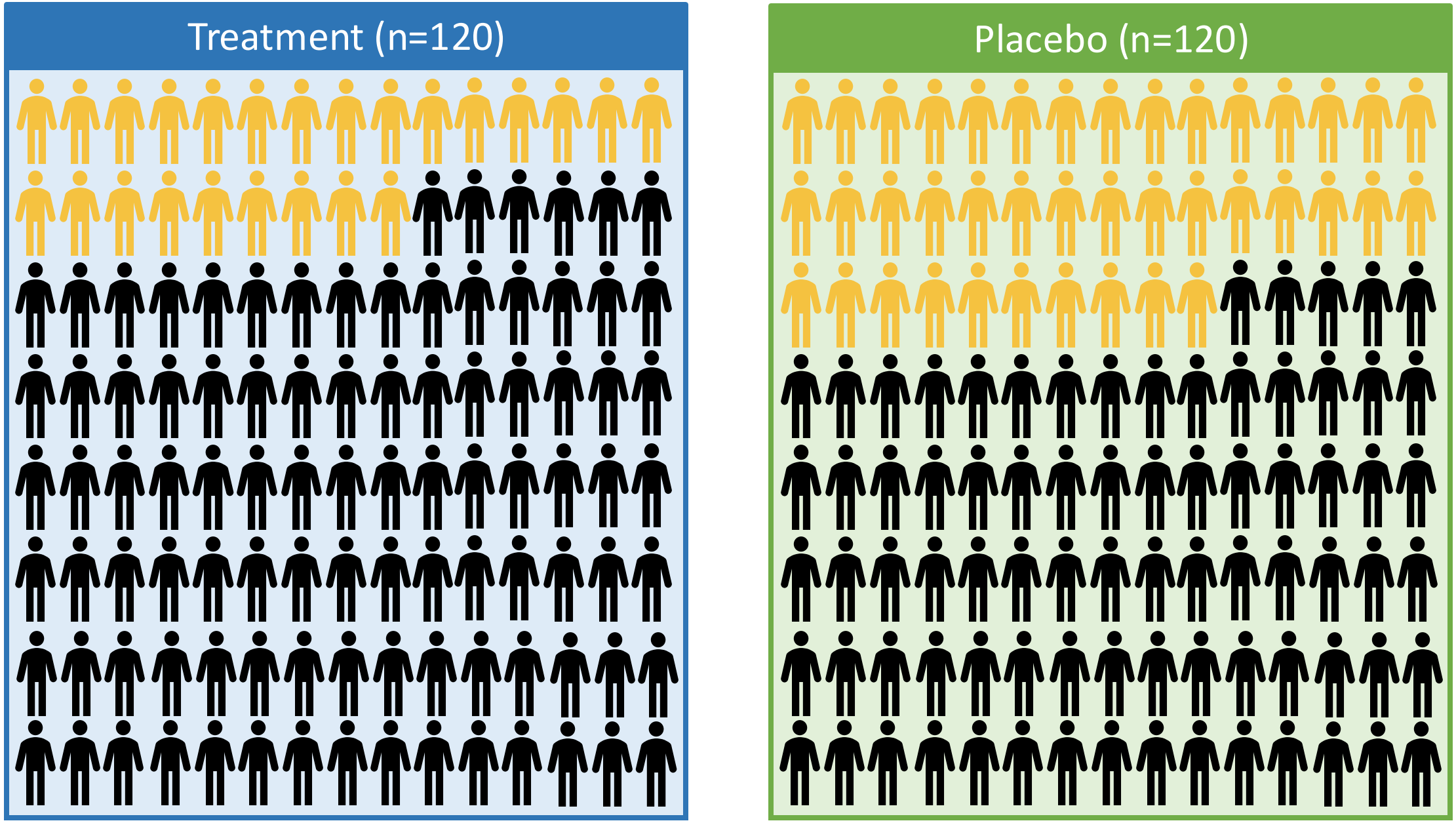
Hypothetical trial

- See how missing data can decrease power and introduce bias

Hypothetical trial



Hypothetical trial

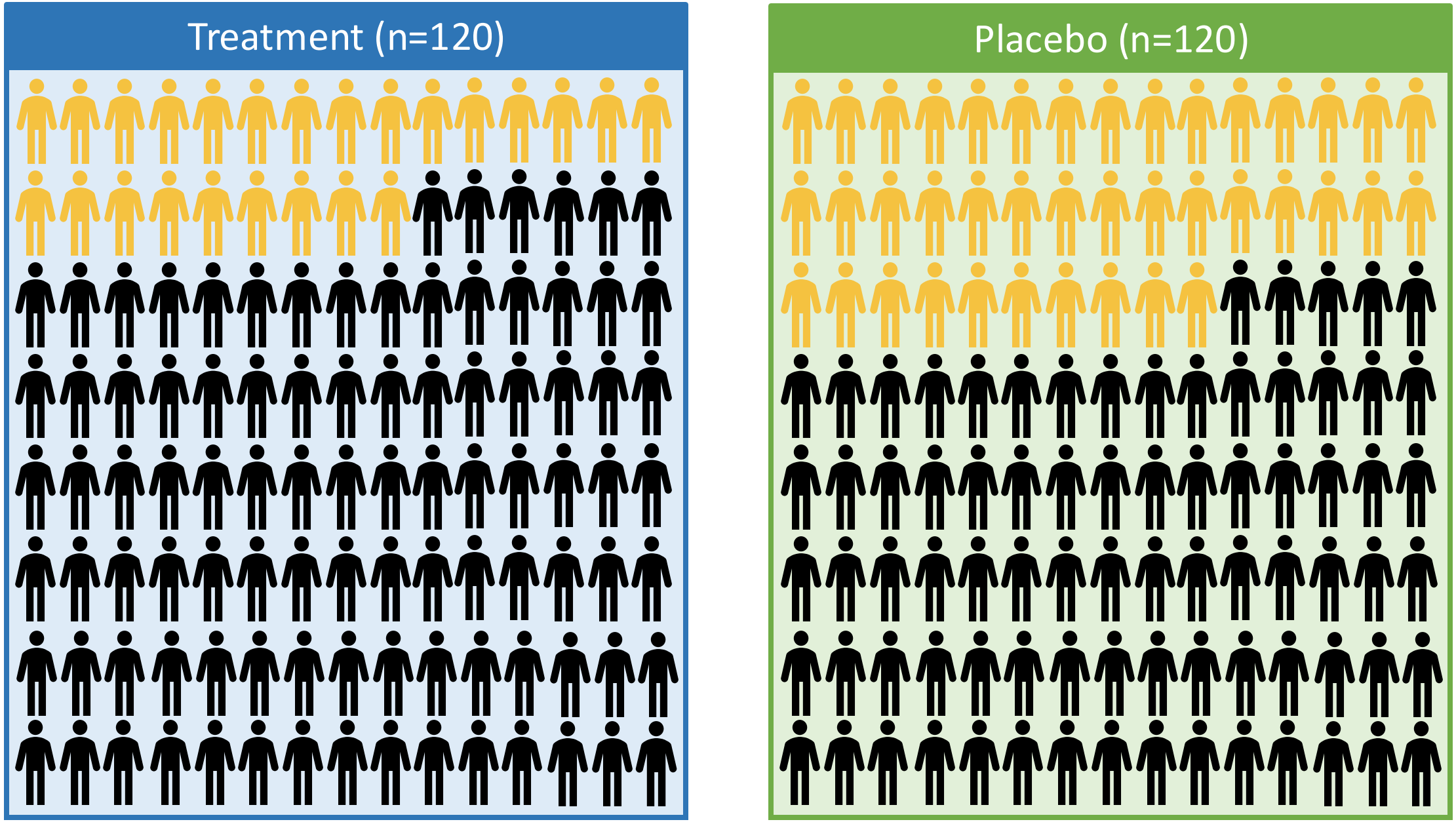


Hypothetical trial

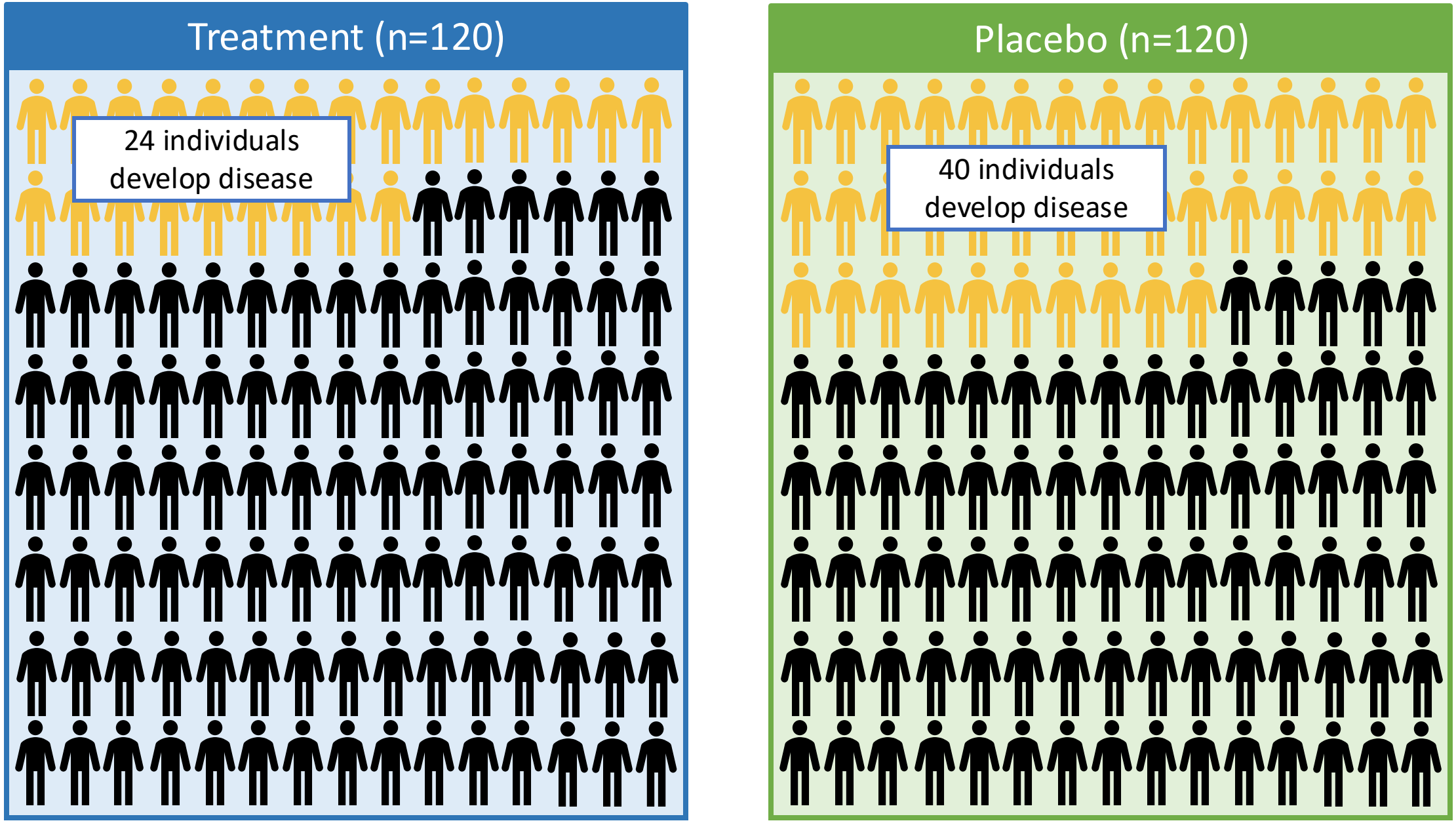
Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%			
B	2%			
C	2%			
D	50%			
E	17%			

https://academic.oup.com/ejendo/article/183/4/E7/6653760?login=false

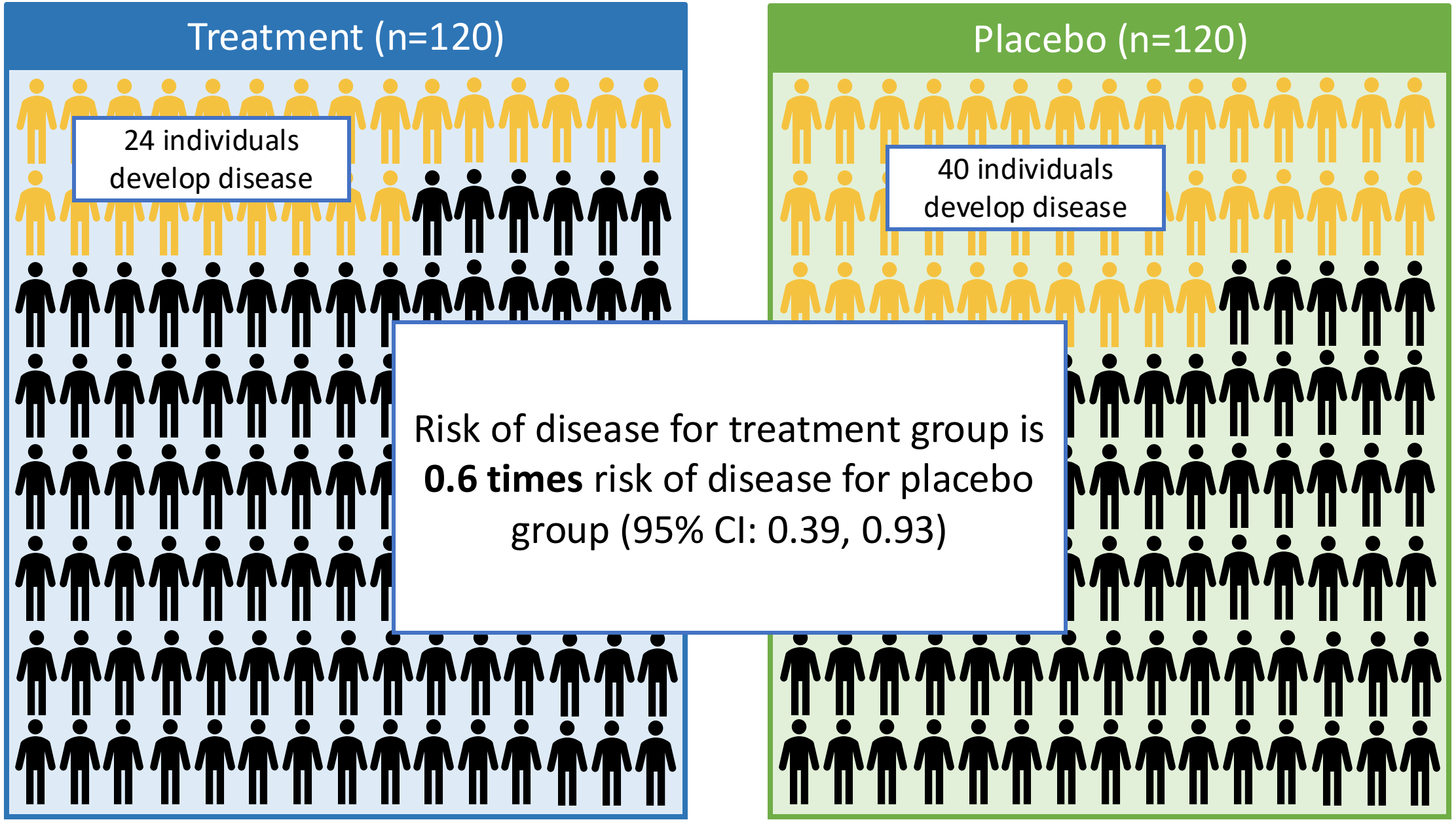
Scenario A: default



Scenario A: default



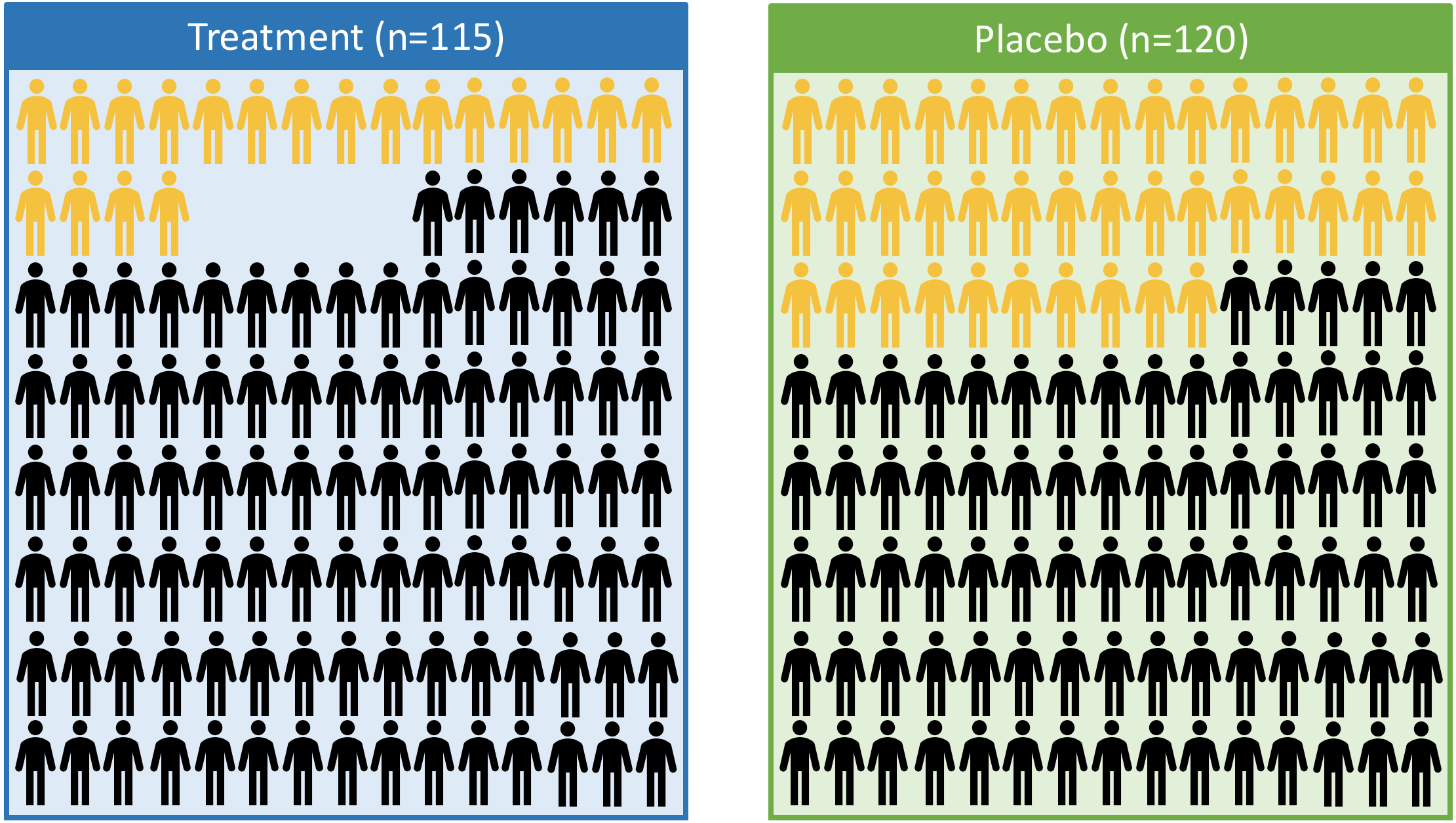
Scenario A: default



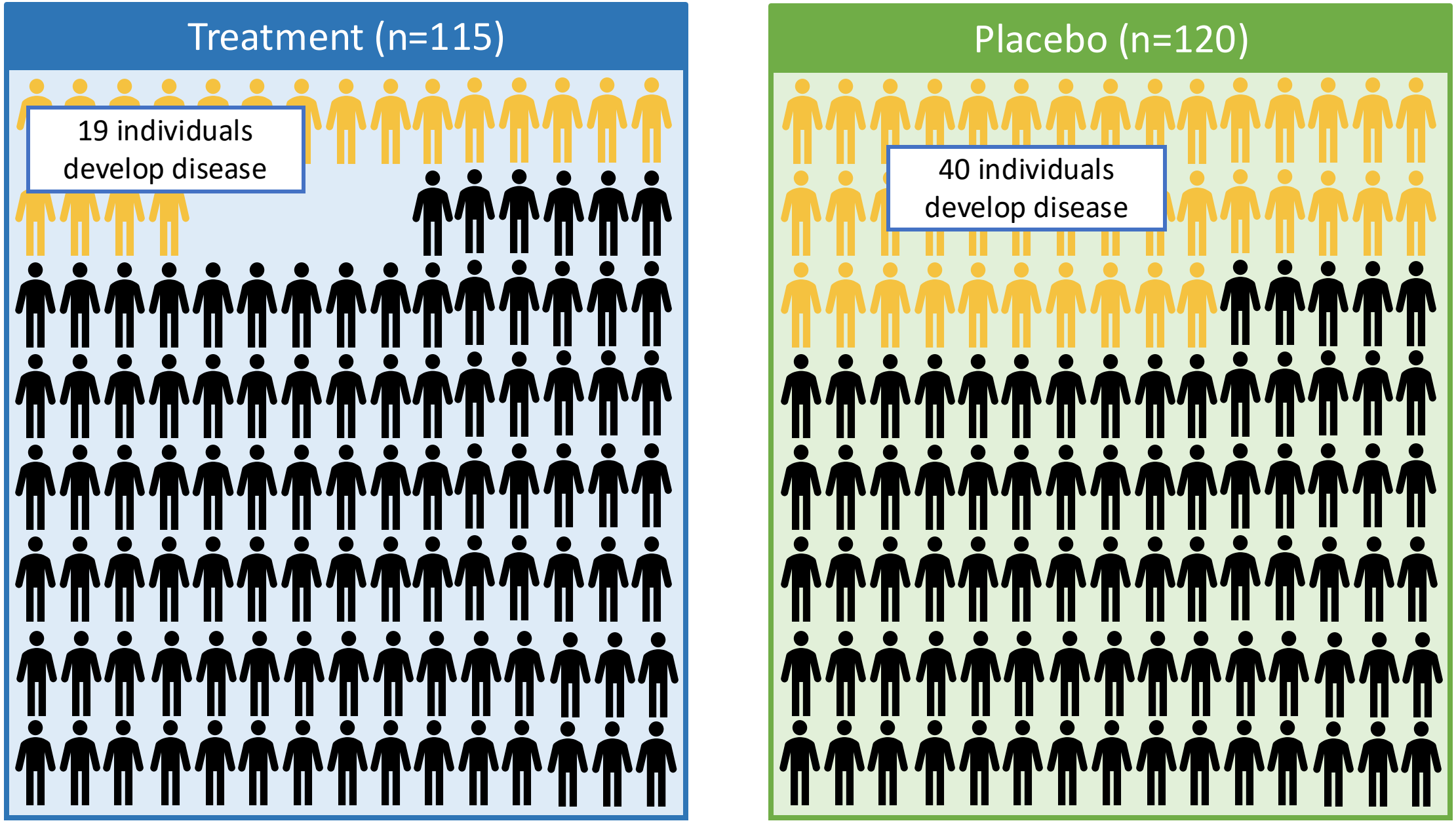
Hypothetical trial

Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%	24/120	40/120	0.60 (95% CI: 0.39, 0.93)
B	2%			
C	2%			
D	50%			
E	17%			

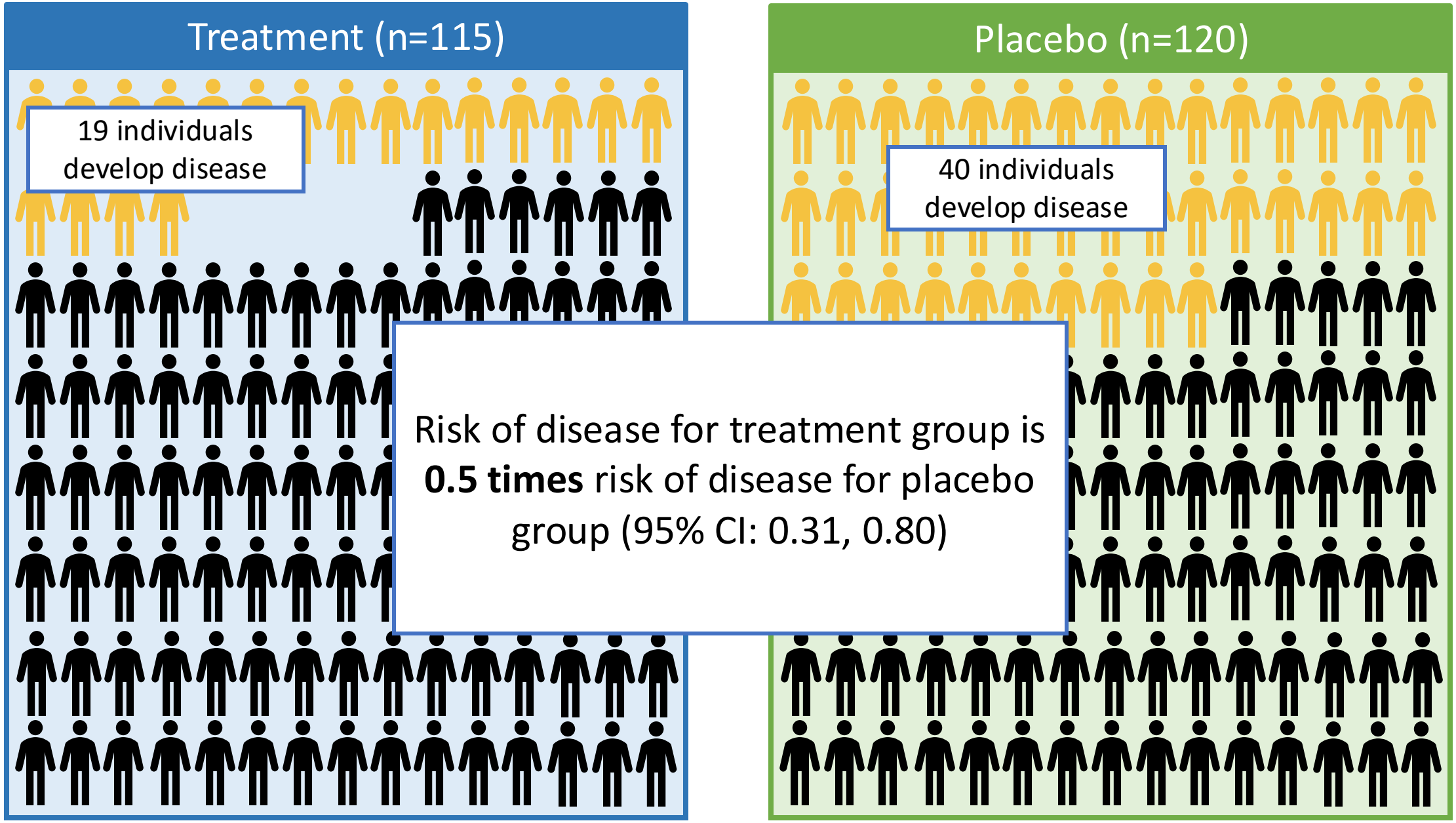
Scenario B (2% missing)



Scenario B (2% missing)



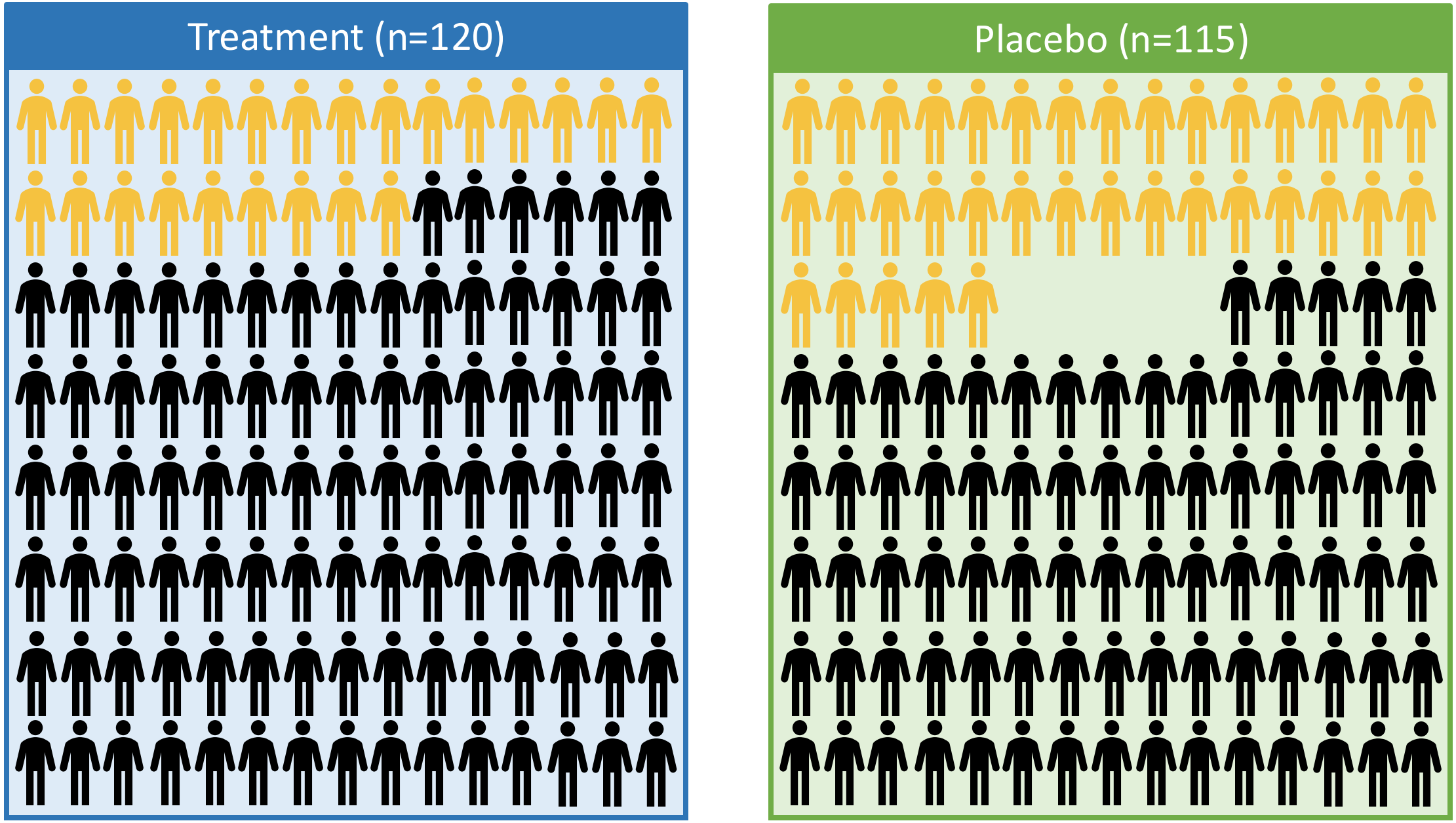
Scenario B (2% missing)



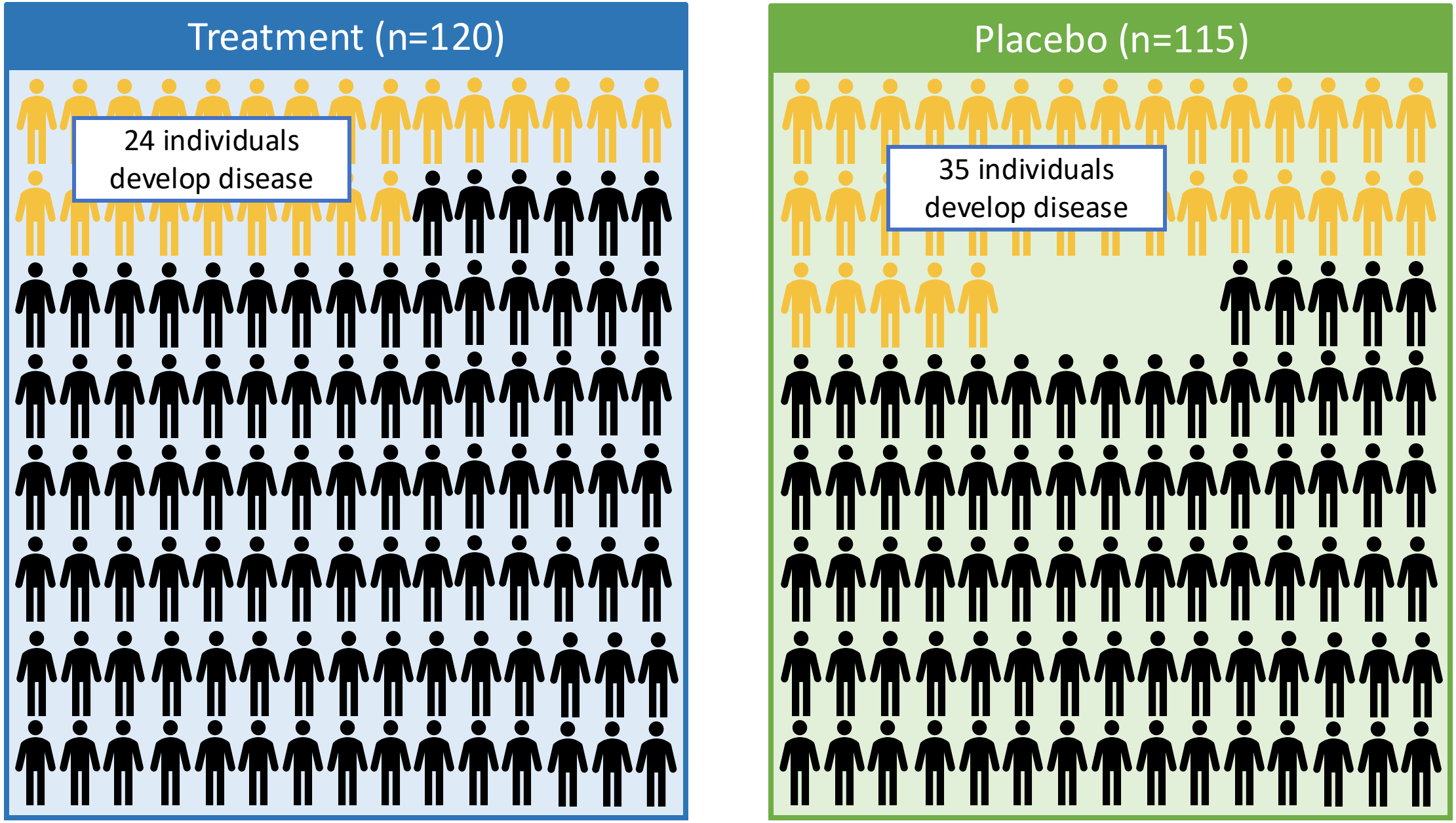
Hypothetical trial

Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%	24/120	40/120	0.60 (95% CI: 0.39, 0.93)
B	2%	19/115	40/120	0.50 (95% CI: 0.31, 0.80)
C	2%			
D	50%			
E	17%			

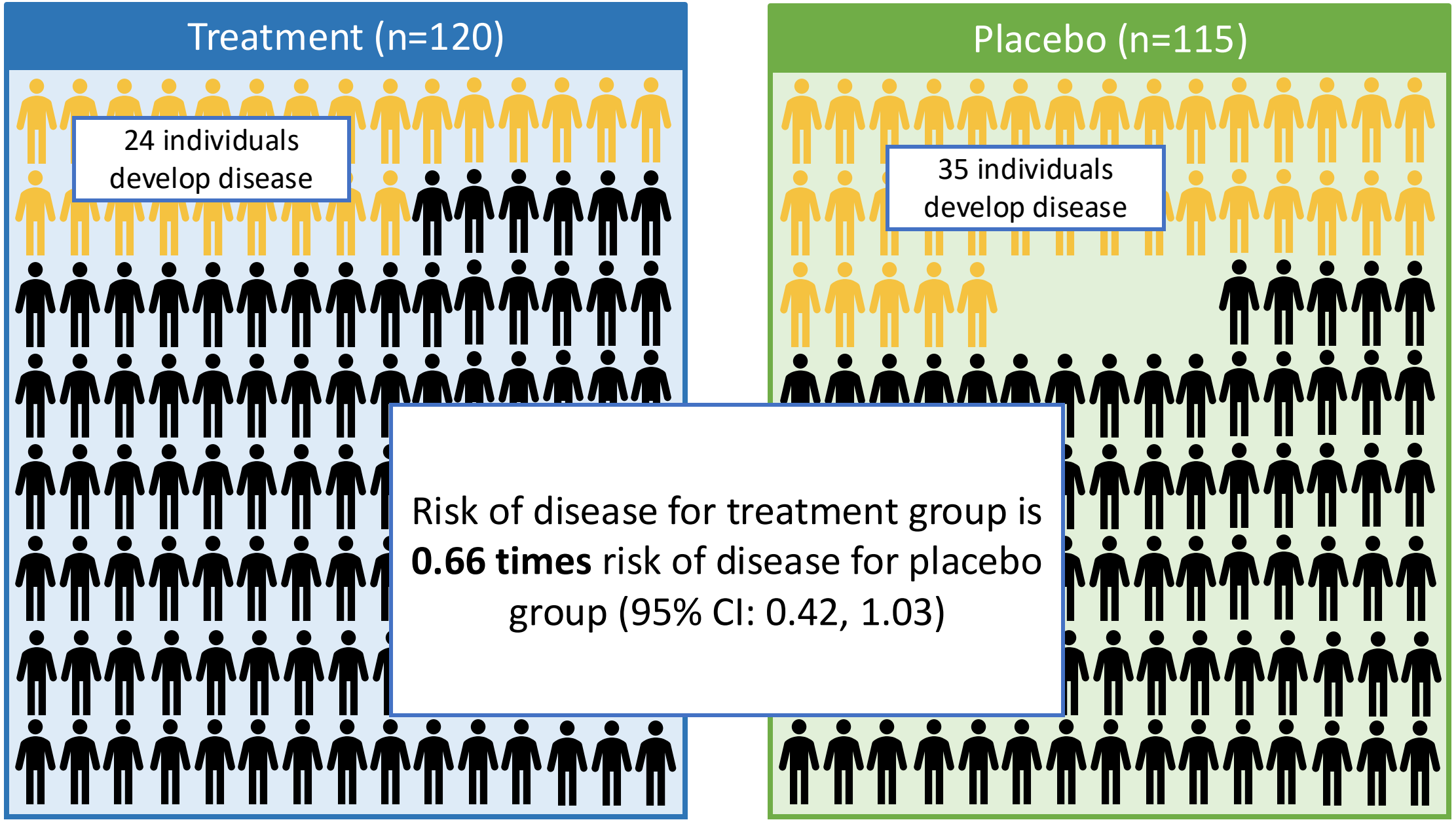
Scenario C (2% missing)



Scenario C (2% missing)



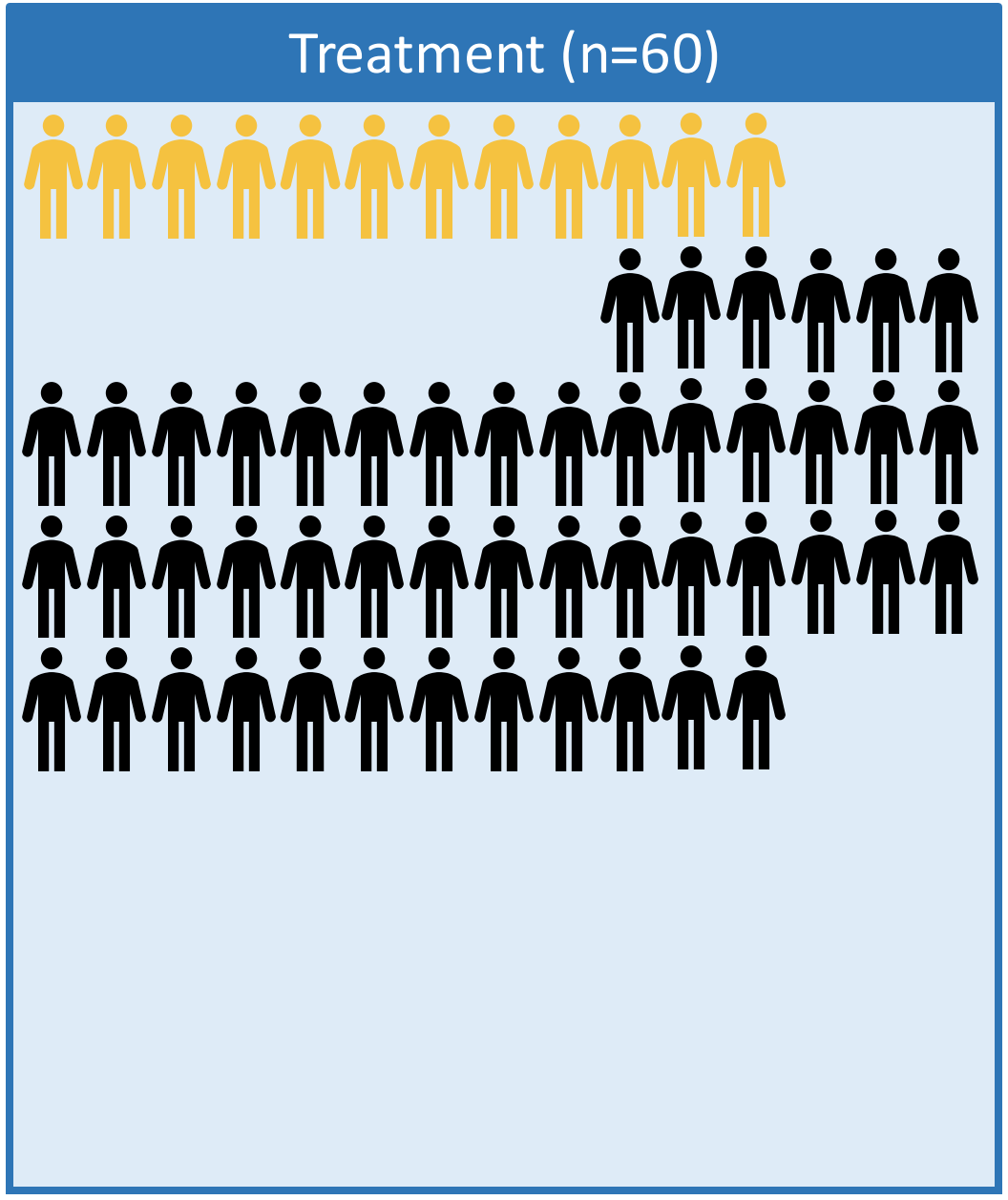
Scenario C (2% missing)

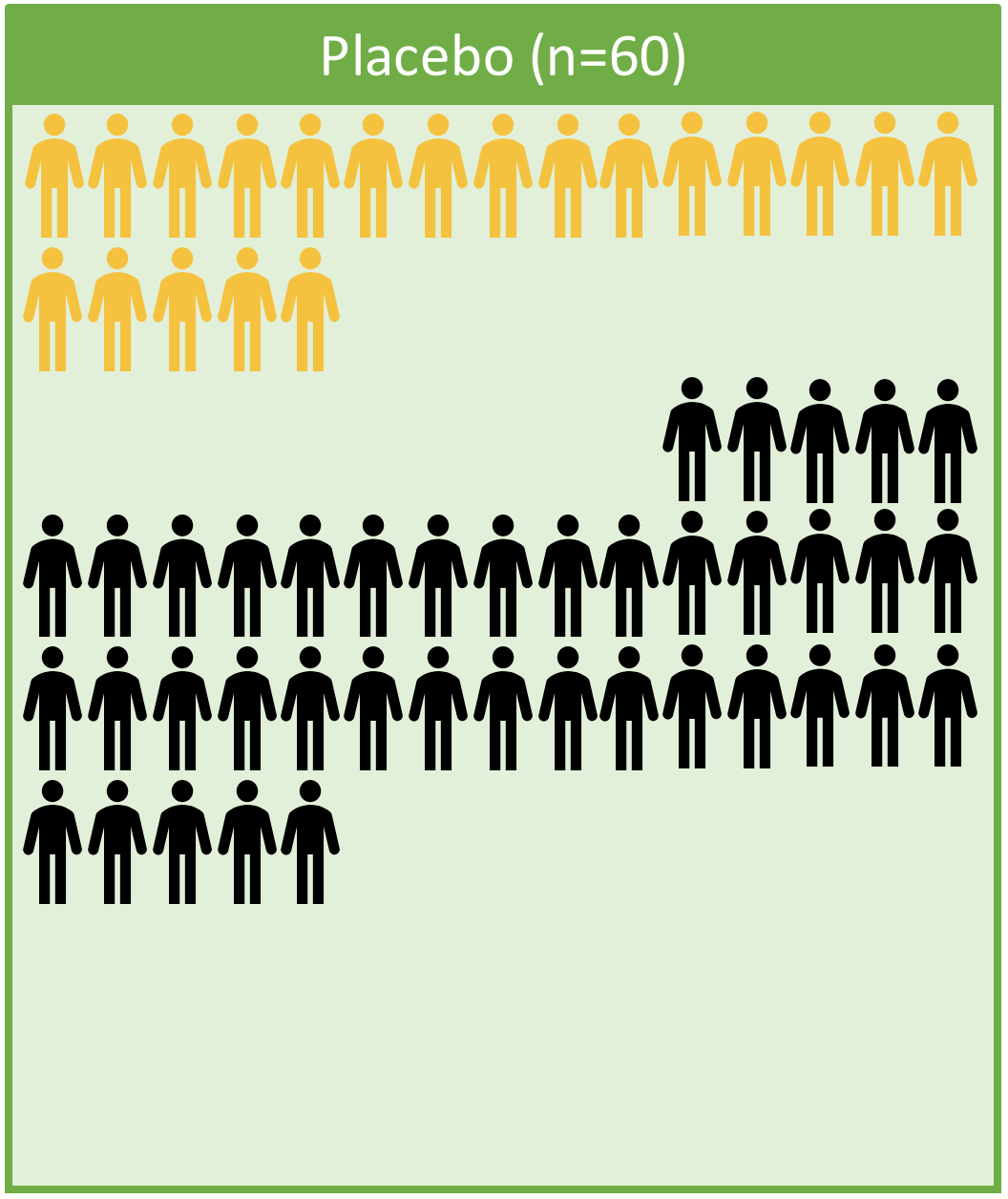


Hypothetical trial

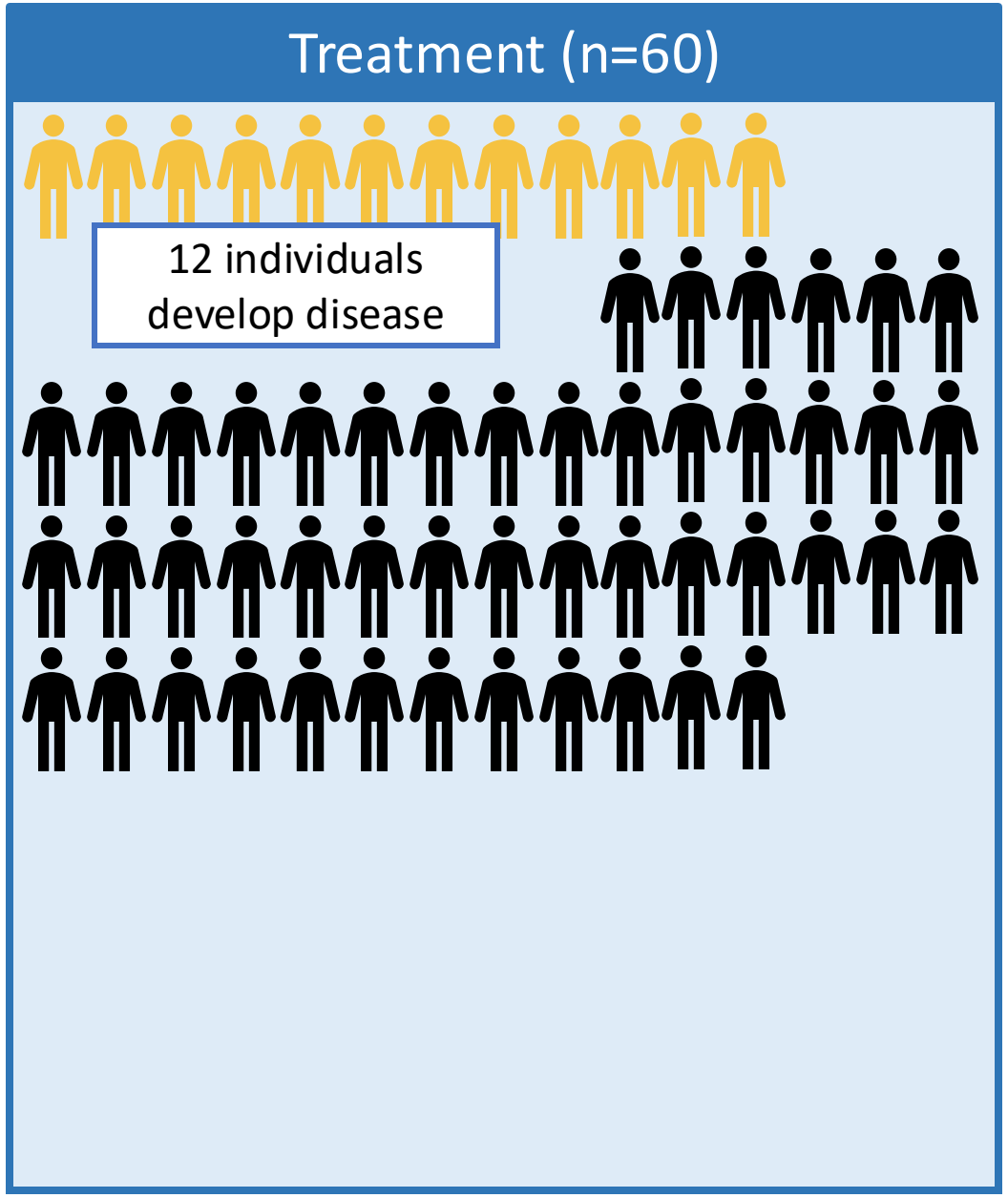
Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%	24/120	40/120	0.60 (95% CI: 0.39, 0.93)
B	2%	19/115	40/120	0.50 (95% CI: 0.31, 0.80)
C	2%	24/120	35/120	0.66 (95% CI: 0.42; 1.03)
D	50%			
E	17%			

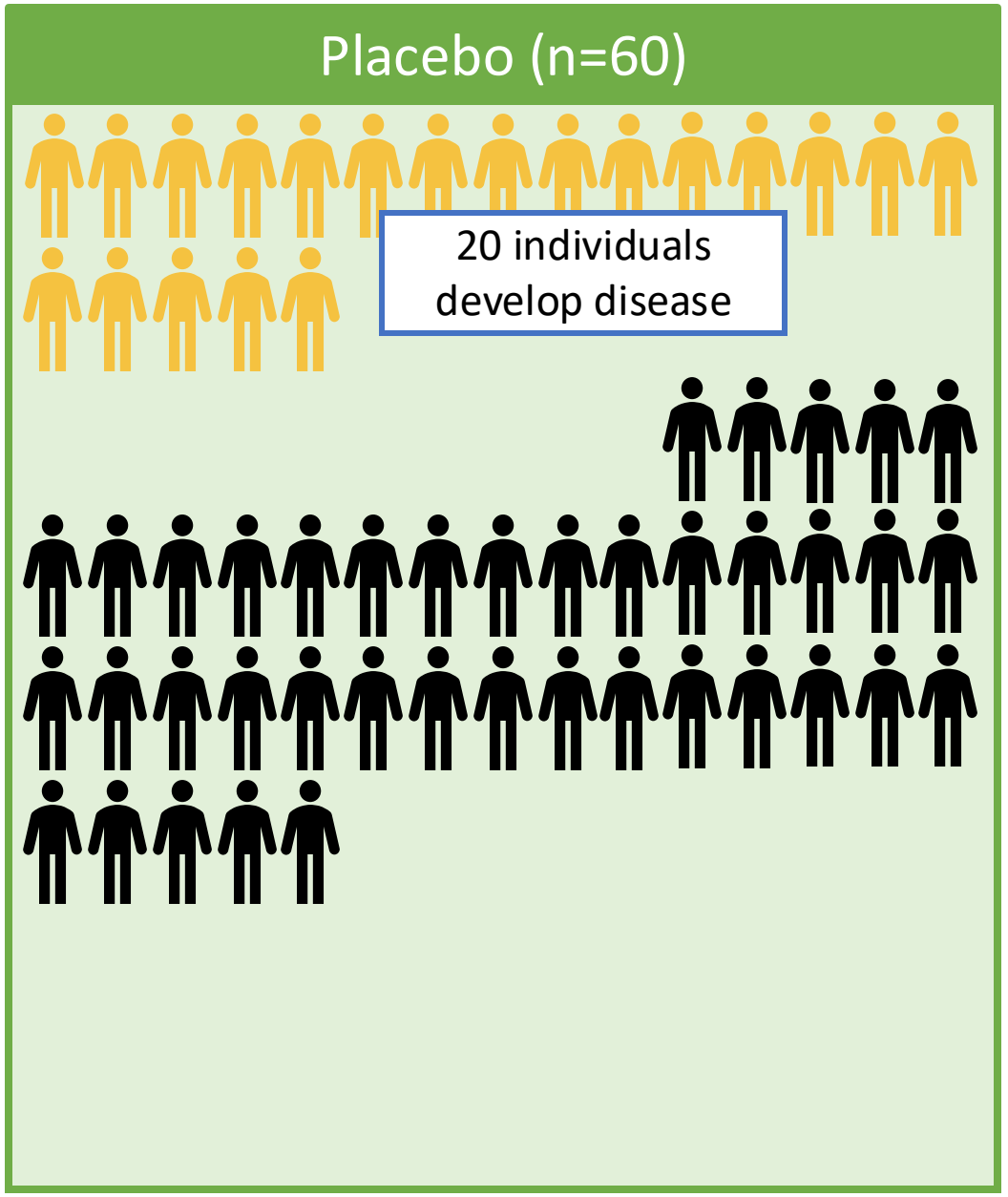
Scenario D (50% missing)



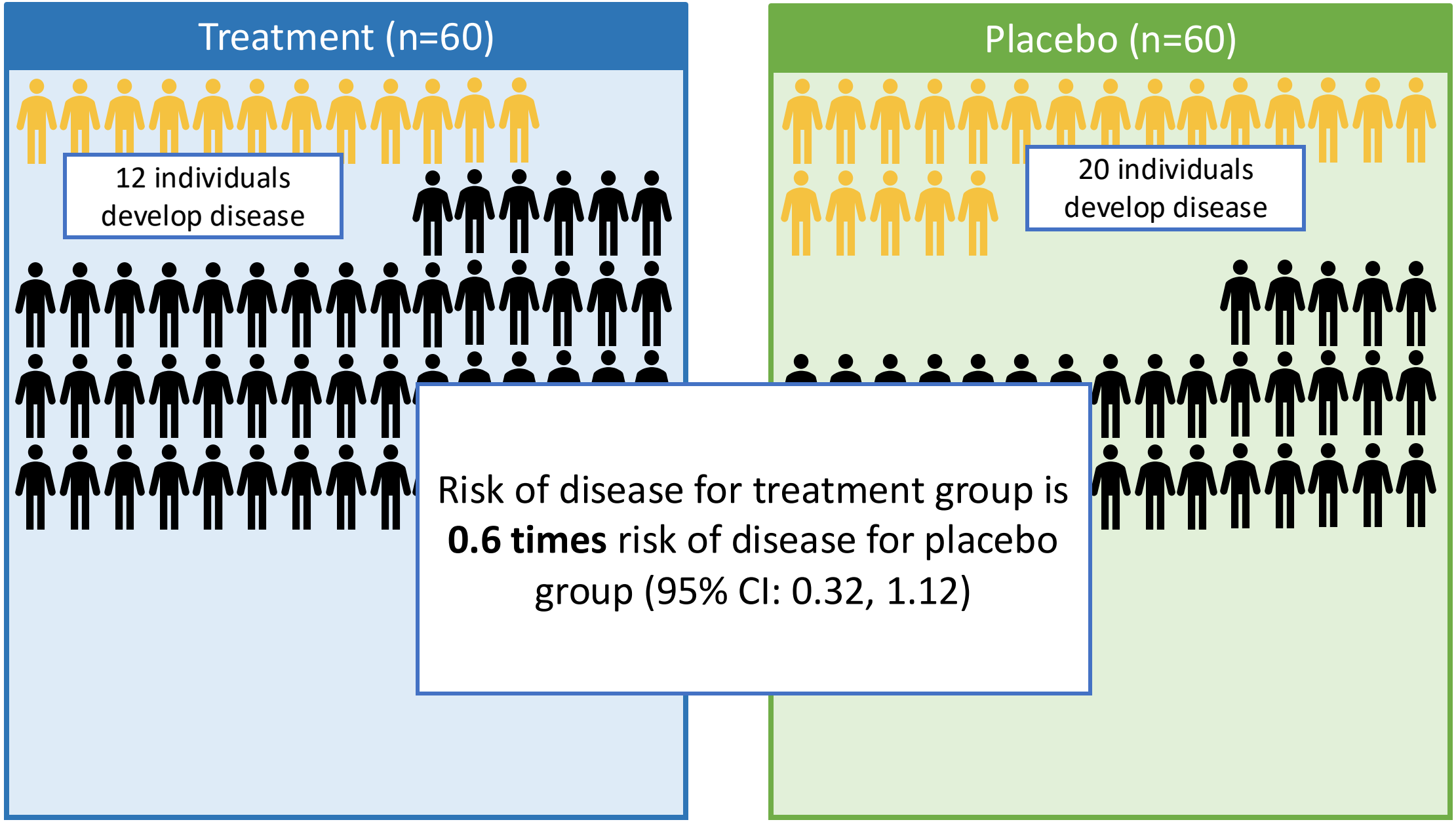


Scenario D (50% missing)





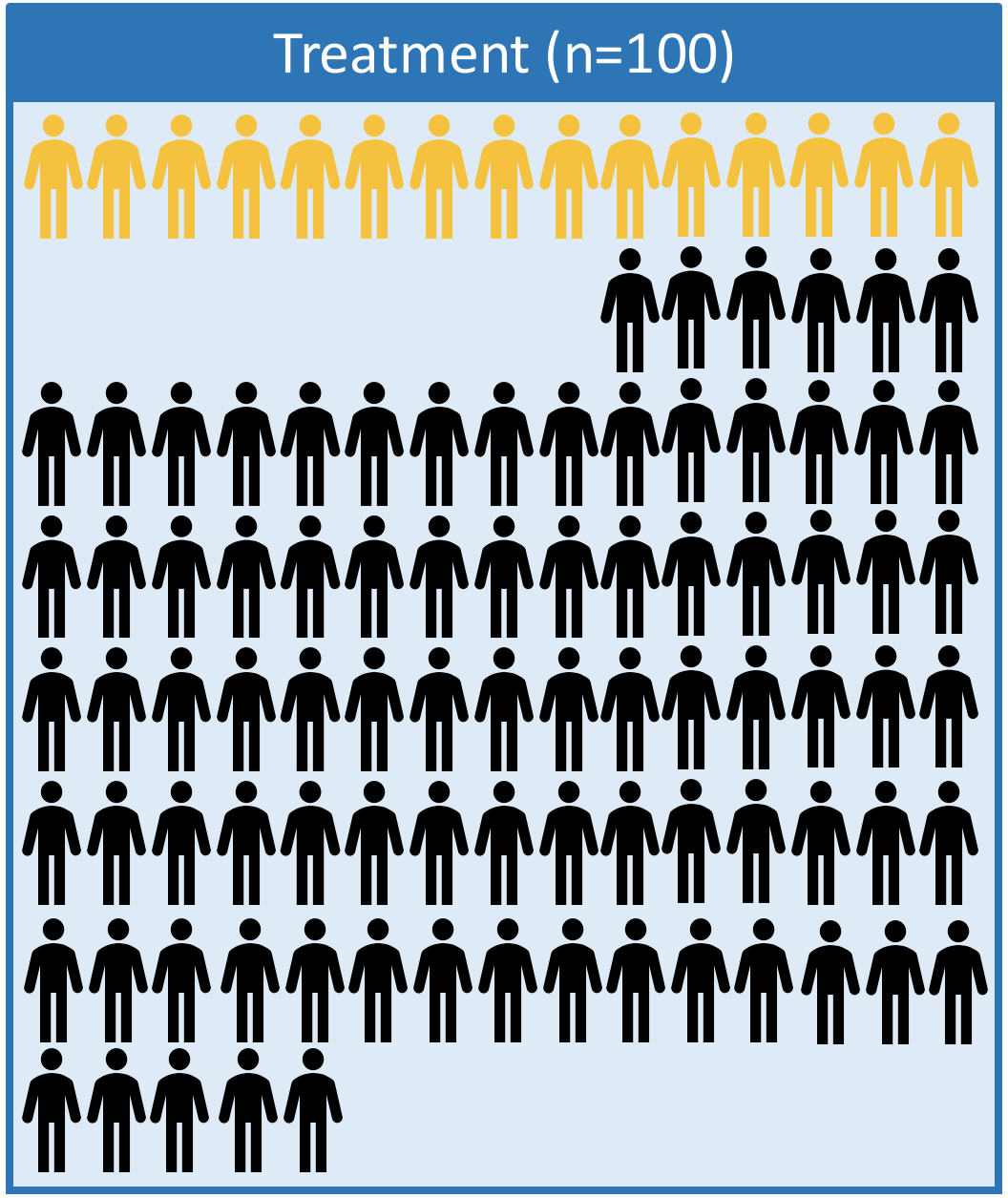
Scenario D (50% missing)



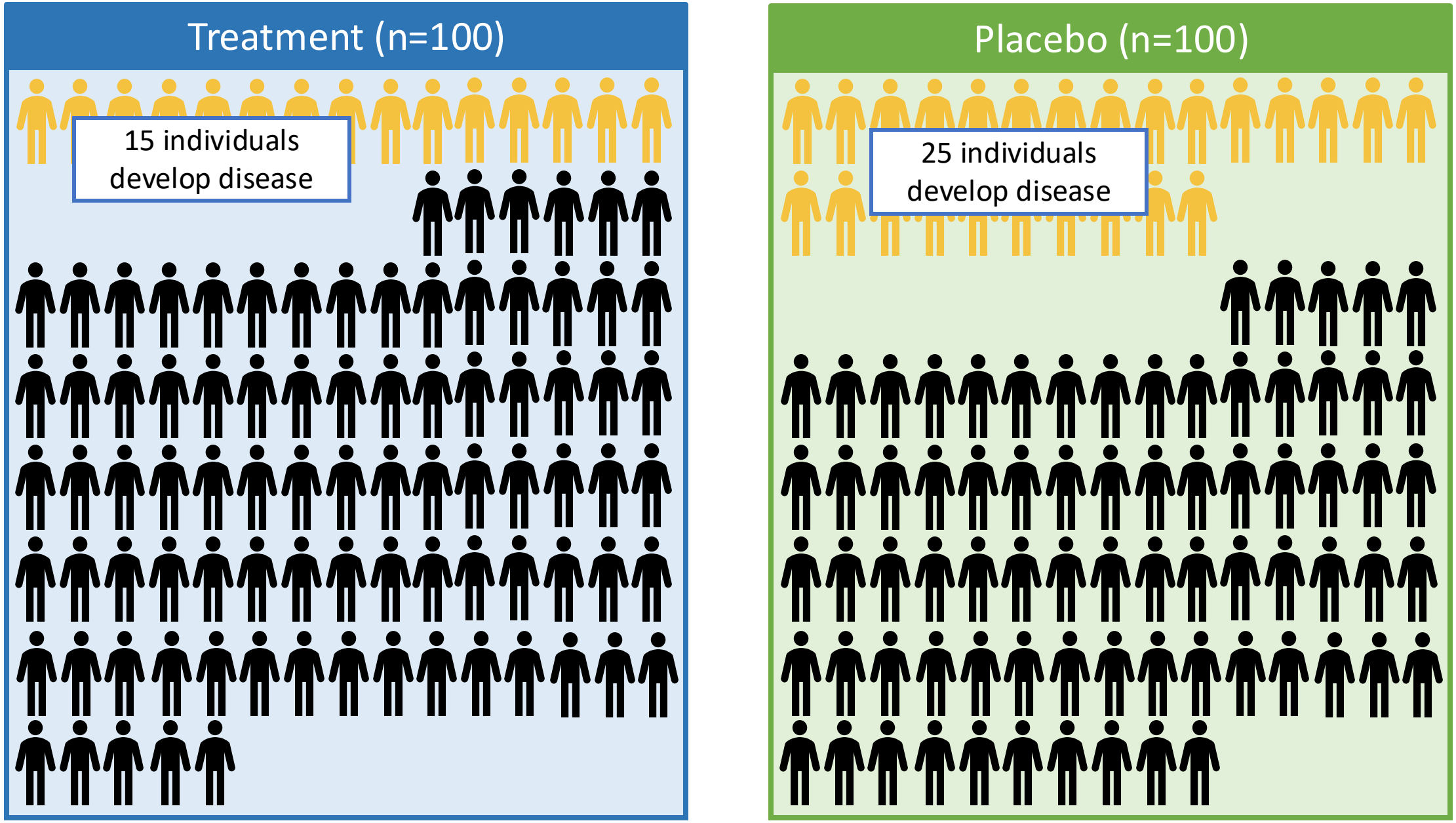
Hypothetical trial

Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%	24/120	40/120	0.60 (95% CI: 0.39, 0.93)
B	2%	19/115	40/120	0.50 (95% CI: 0.31, 0.80)
C	2%	24/120	35/120	0.66 (95% CI: 0.42; 1.03)
D	50%	12/60	20/60	0.60 (95% CI: 0.32, 1.12)
E	17%			

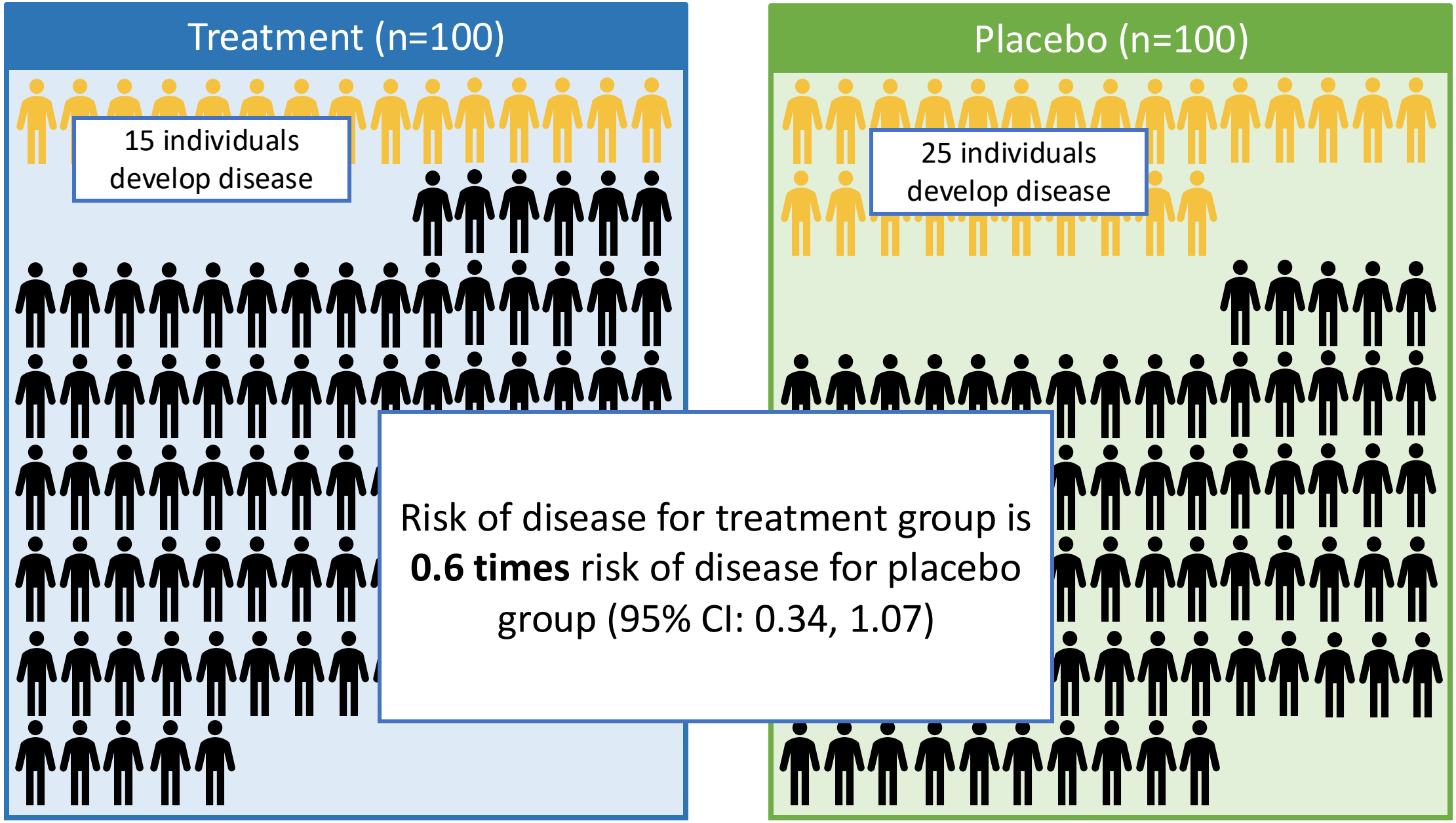
Scenario E (17% missing)



Scenario E (17% missing)



Scenario E (17% missing)



Hypothetical trial

Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%	24/120	40/120	0.60 (95% CI: 0.39, 0.93)
B	2%	19/115	40/120	0.50 (95% CI: 0.31, 0.80)
C	2%	24/120	35/120	0.66 (95% CI: 0.42; 1.03)
D	50%	12/60	20/60	0.60 (95% CI: 0.32, 1.12)
E	17%	15/199	25/100	0.60 (95% CI: 0.34, 1.07)

Hypothetical trial

Scenario	Percent Missing	Treatment	Placebo	Estimated Risk Ratio
A (default)	0%		/120	0.60 (95% CI: 0.39, 0.93)
B	2%	19/115	40/120	0.50 (95% CI: 0.31, 0.80)
C	2%	24/120	35/120	0.66 (95% CI: 0.42; 1.03)
D	50%	12/60	20/60	0.60 (95% CI: 0.32, 1.12)
E	17%	15/199	25/100	0.60 (95% CI: 0.34, 1.07)

Introduce Bias

Loss of Power

Preventing Missing Data

Preventing missing data

- Missing data is unavoidable, but we can work to minimize the amount of missing data

- We can do this through:

  - **Trial design:** study population, outcome measurement, and way treatment is administered

  - **Trial conduct:** limiting the burden on participants

https://www.nejm.org/doi/full/10.1056/NEJMsr1203730

Trial design

Table 1. Eight Ideas for Limiting Missing Data in the Design of Clinical Trials.

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study.
Include a run-in period in which all patients are assigned to the active treatment, after which only those who tolerated and adhered to the therapy undergo randomization.
Allow a flexible treatment regimen that accommodates individual differences in efficacy and side effects in order to reduce the dropout rate because of a lack of efficacy or tolerability.
Consider add-on designs, in which a study treatment is added to an existing treatment, typically with a different mechanism of action known to be effective in previous studies.
Shorten the follow-up period for the primary outcome.
Allow the use of rescue medications that are designated as components of a treatment regimen in the study protocol.
For assessment of long-term efficacy (which is associated with an increased dropout rate), consider a randomized withdrawal design, in which only participants who have already received a study treatment without dropping out undergo randomization to continue to receive the treatment or switch to placebo.
Avoid outcome measures that are likely to lead to substantial missing data. In some cases, it may be appropriate to consider the time until the use of a rescue treatment as an outcome measure or the discontinuation of a study treatment as a form of treatment failure.

https://www.nejm.org/doi/full/10.1056/NEJMsr1203730

Trial conduct

Table 2. Eight Ideas for Limiting Missing Data in the Conduct of Clinical Trials.

Select investigators who have a good track record with respect to enrolling and following participants and collecting complete data in previous trials.
Set acceptable target rates for missing data and monitor the progress of the trial with respect to these targets.
Provide monetary and nonmonetary incentives to investigators and participants for completeness of data collection, as long as they meet rigorous ethical requirements.[15,16]
Limit the burden and inconvenience of data collection on the participants, and make the study experience as positive as possible.
Provide continued access to effective treatments after the trial, before treatment approval.
Train investigators and study staff that keeping participants in the trial until the end is important, regardless of whether they continue to receive the assigned treatment. Convey this information to study participants.
Collect information from participants regarding the likelihood that they will drop out, and use this information to attempt to reduce the incidence of dropout.
Keep contact information for participants up to date.

Sample inflation is not prevention

- Inflating the sample size to reach the originally desired sample size is NOT a good method!
  - Does not account for **why data is missing**

- Does not factor in that missing data **may be fundamentally different than observed data**

- Will **further strengthen any bias due to missing data**
  - Think back to Scenario B and C in hypothetical trial

Defining missing data

Defining missing data

- Three ways to define missing data

- Missing data patterns
- Missing data types/mechanisms
- Ignorability

- Important definition
  - Imputation: assigning a value to missing data

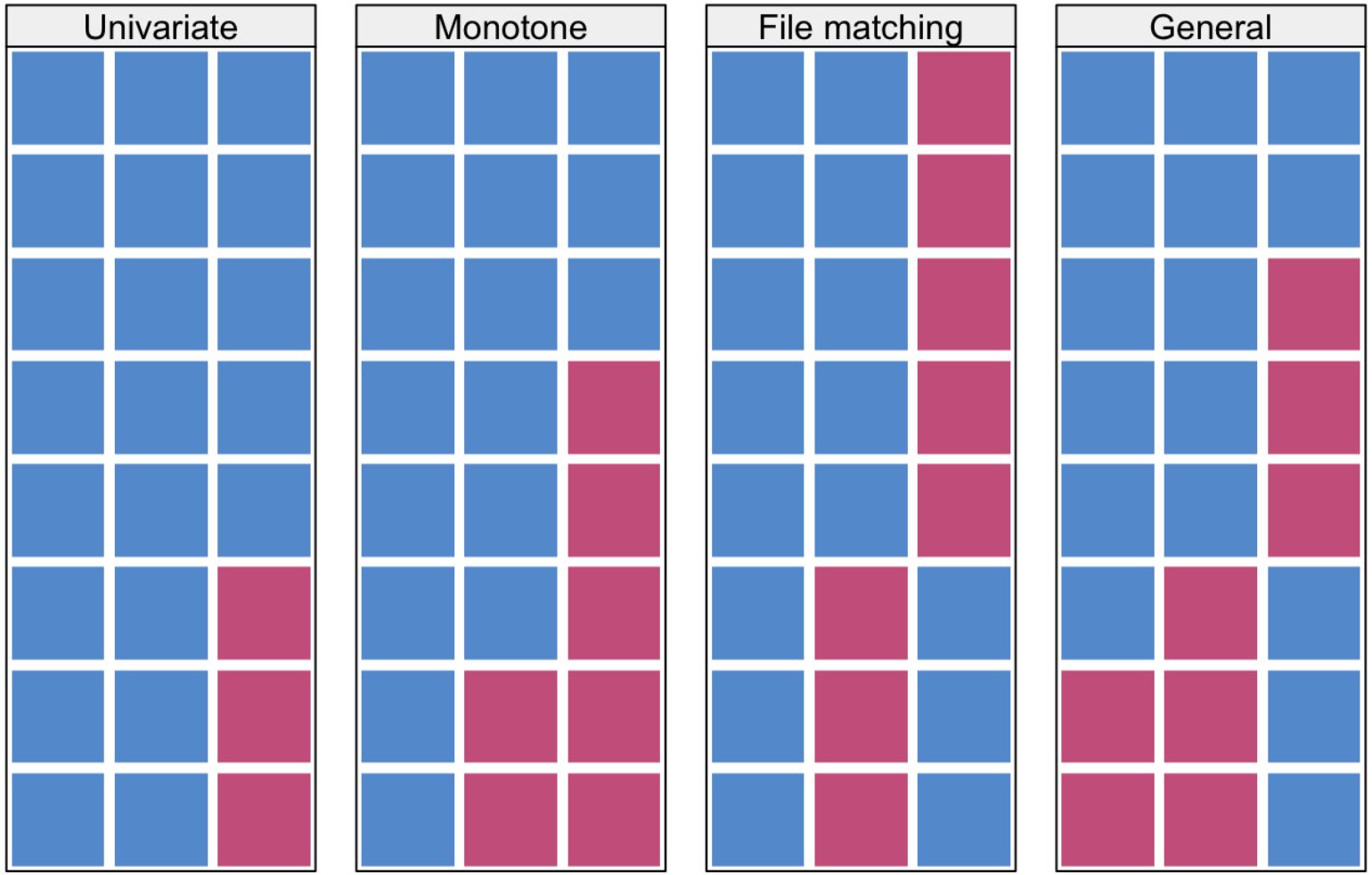
Missing data patterns

- “Describes where the gaps in data are” (Koptur site)

- Will help us determine what statistical methods are appropriate for imputing missing data

Missing data patterns

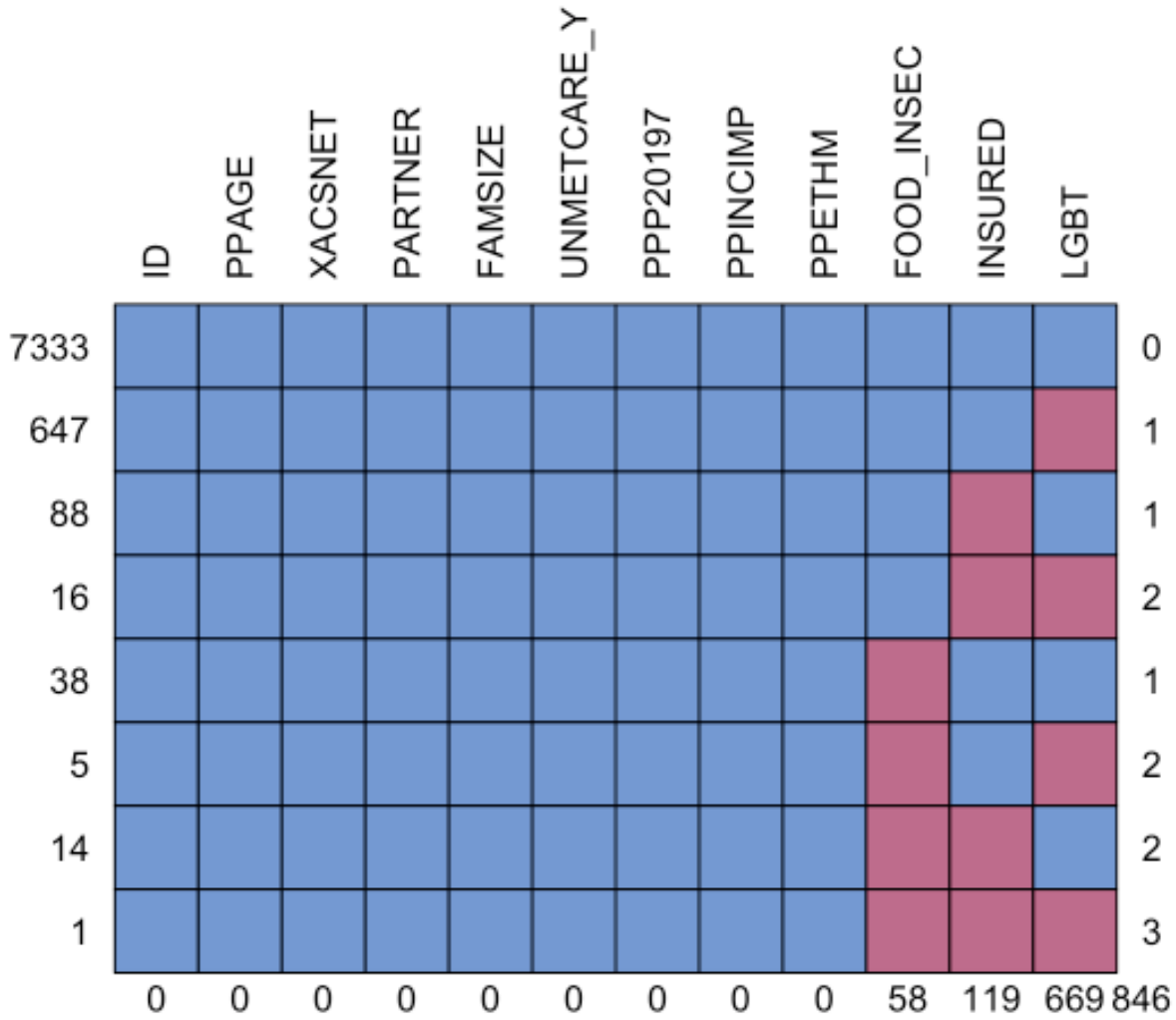
- **Univariate:** only one variable has missing data
- **Monotone:** drop out in longitudinal studies
- **File matching:** variables are not observed together
- **General:** multiple variables and at any time point



https://stefvanbuuren.name/fimd/missing-data-pattern.html

Missing data pattern

```
md.pattern(wbns2, rotate.names = T)
```



2. Missing data types/mechanisms

- **Missing completely at random** (MCAR)
  - Probability of missing data is *independent* of observed and unobserved data

- **Missing at random** (MAR)
  - When we condition on the observed data, the probability of missing data is independent of unobserved data

- **Missing not at random** (MNAR)
  - Probability of missing data is *NOT independent* of unobserved data (whether we condition on the observed data or not)

white

light grey brown black

orange black

dark grey light grey calico

I’m collecting data on the color of my cats…

Possible colors: white, black, brown, tan, light grey, dark grey, calico

Missing completely at random

I randomly picked up whatever cats I could find, and recorded their color.

I'm collecting data on the color of my cats…

Possible colors: white, black, brown, tan, light grey, dark grey, calico

Missing at random

I also observed and recorded their eye colors, and (for some reason??) didn't pick up cats with green eyes.

I’m collecting data on the color of my cats…

Possible colors: white, black, brown, tan, light grey, dark grey, calico

Missing not at random

I picked up my favorite cats, but I’ll never record which ones are my favorite.

3. Ignorable vs. Non-ignorable

- Type of missingness dictates (statistically) how we need to address the missing data
- We often **group the types into ignorable and non-ignorable** missing data

- **Ignorable:** Measures that are missing can be inferred from the observed data
- **Non-ignorable:** Measures that are missing cannot be inferred from observed data

- Note: ignorable does **NOT** mean ***we do not need to address*** the missing data
  - We still need to address ignorable missing data
  - Non-ignorable means we **cannot proceed with analysis**

Why is it important to determine the type?

- For each type/mechanism, there is a specific way to address the missing data

- **MCAR (ignorable):** we *do not need to impute* the missing data

- **MAR (ignorable):** we can *use the observed data to impute the missing data*

- **MNAR (non-ignorable):** we ***cannot*** *use observed data to impute* and we need to conduct further research on what we can measure to determine why data are missing

How to determine the missing mechanism

- When in doubt between **MCAR** and **MAR**, just assume **MAR**

- So I'd say the biggest decision is **ignorable** vs. **non-ignorable**
  - And this is NOT something we can test for statistically
  - So we need to use field knowledge

- If **ignorable** (MCAR or MAR), we can use imputations

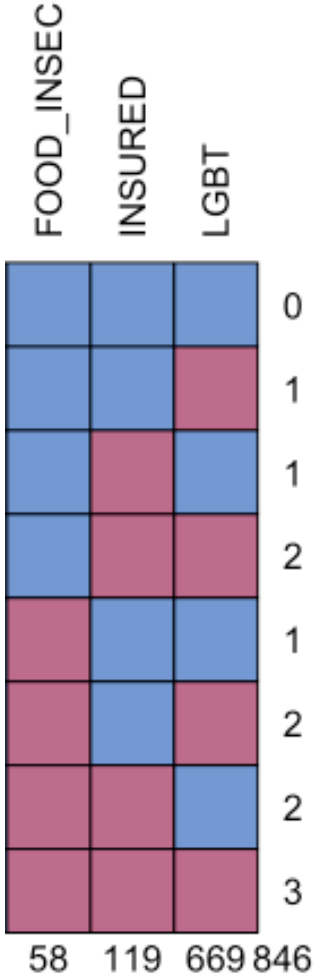
- If **non-ignorable**, we need to incorporate more variables for more information

Example: LGBT variable

- Why is it missing?
- Look at the User Guide

- This makes me think it might be **MNAR**
  - Seems like options for sexual orientation and gender identity are lacking
  - Missing LGBT observation may be linked to queer identities
  - Are missing values more likely queer folks not answering?

- Two options
  - Assume **MAR**, impute, and then perform a sensitivity analysis
  - Missing indicator method: missing values become their own category



respondent did not answer the question. The variable for sexual orientation and gender identity (lgbt) is coded as missing if the questions were not asked, the respondent did not answer one or both questions, or the respondent provided insufficient information to determine if they are gay, lesbian, bisexual, or transgender.

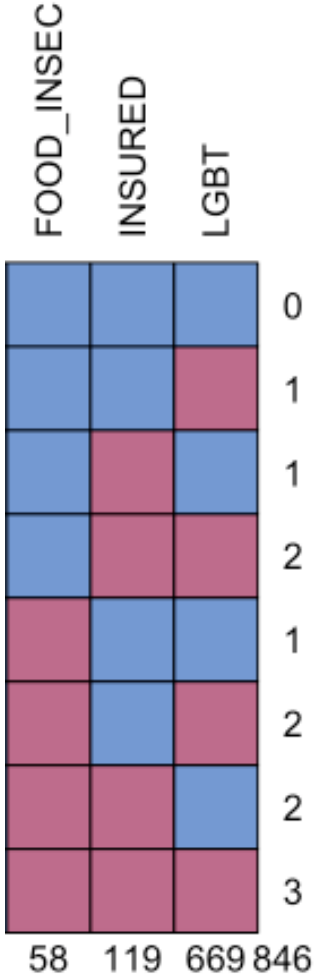
Example: insurance

- If **MNAR**: underlying insurance status may determine if missing
  - Any other unobserved traits may determine

- If **MAR**: then we can impute
  - Maybe missing insurance is related to unmet care?

- Three options
  - Assume **MAR**, impute, and then perform a sensitivity analysis
  - Missing indicator method: missing values become their own category
  - Use the 5 imputed sets in the dataset

- Check User Guide, and looks like they used additional info not available to us to generate 5 imputations (page 6 of User Guide)



5 min break + me giving your missing data mechanism

When you come back, you will receive a paper from me with your assignment for first activity

Break for activity part 1

We will split into groups of 3-4. Each person in the group will be assigned a missing type (MCAR, MAR, or MNAR)

15 minutes

Handling missing data

Handling missing data

Table 1. Handling missing data: an overview

Missing data mechanism	Analysis	Imputation
MCAR	Complete case analysis	No imputation necessary
MAR	No complete case analysis	Single imputation methods not valid Multiple imputation needed
MNAR	No complete case analysis	All imputation methods not valid

https://www.sciencedirect.com/science/article/pii/S0895435622002189

Statistical methods to handle missing data

- Leave data as is (**analysis**):
  - Only works for **MCAR** (does not introduce bias)
  - We use the observed data to perform our analysis

- Generate observations for missing data (**imputation**):
  - Works for **MCAR** or **MAR**
  - We use the observed data to impute the missing data

Analysis

- We use the observed data to perform our analysis

- Two common approaches
  - Complete case analysis (Listwise deletion)
    - Use only the observations with complete information
  - Pairwise deletion
    - Use incomplete observations
    - Missing variables within each incomplete obersvation will not be analyzed

Imputation

- Works to fill in missing data with the best guess value

- **Structures to imputation:** number of imputations performed on data

  - *Single imputation:* One imputation, one analysis
    - For methods with best guess that never changes
    - Generally not advised

  - *Multiple imputation:*
    - More than one imputation with their own analyses
    - Pool together the results from all the analyses
    - Typically better because it includes the uncertainty around our best guess value

Methods of imputation: single imputation

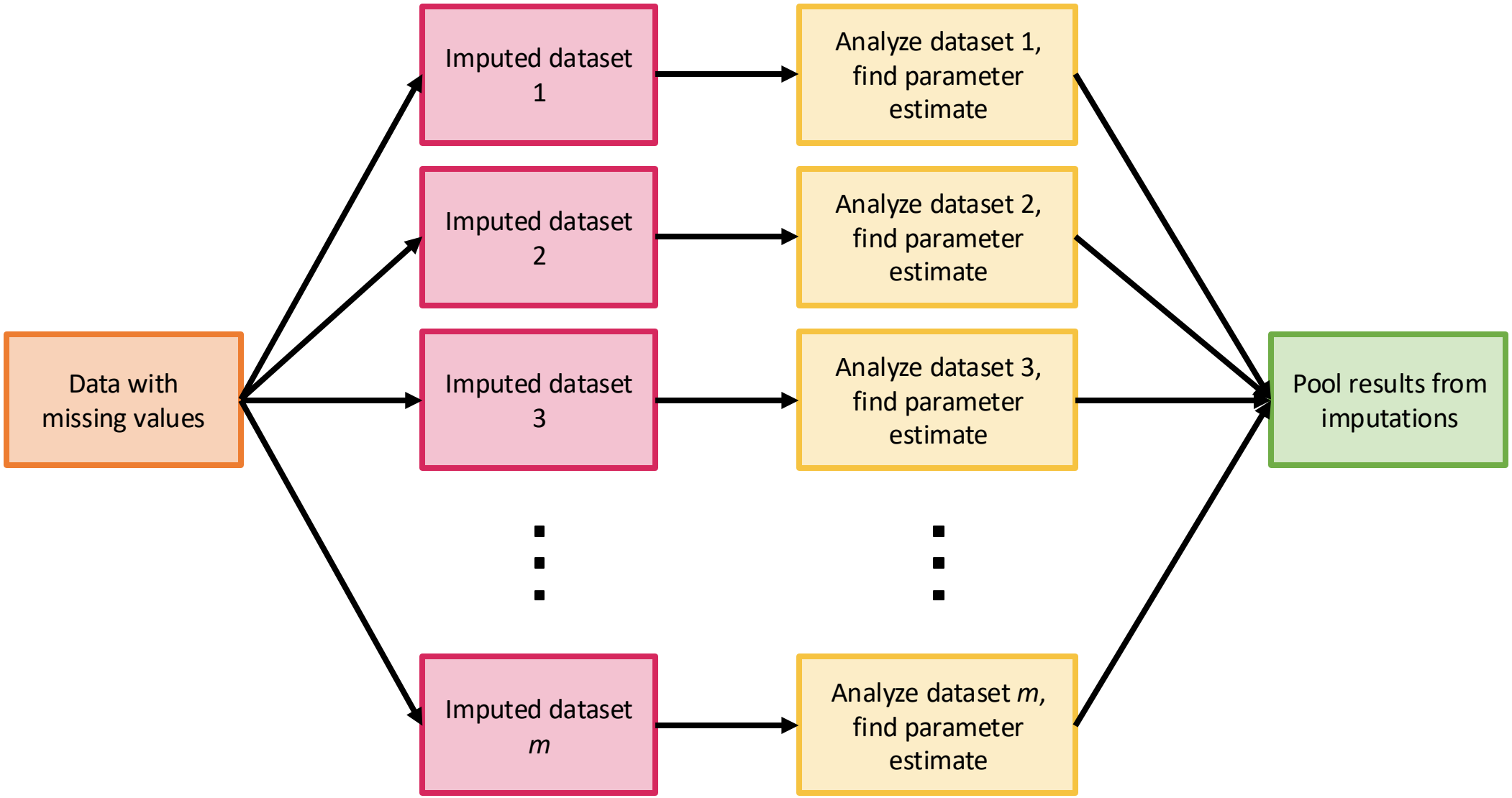
- **Again, generally not advised**

- Mean imputation
  - For number values, take the mean of the observed values
  - Fill missing data with mean value

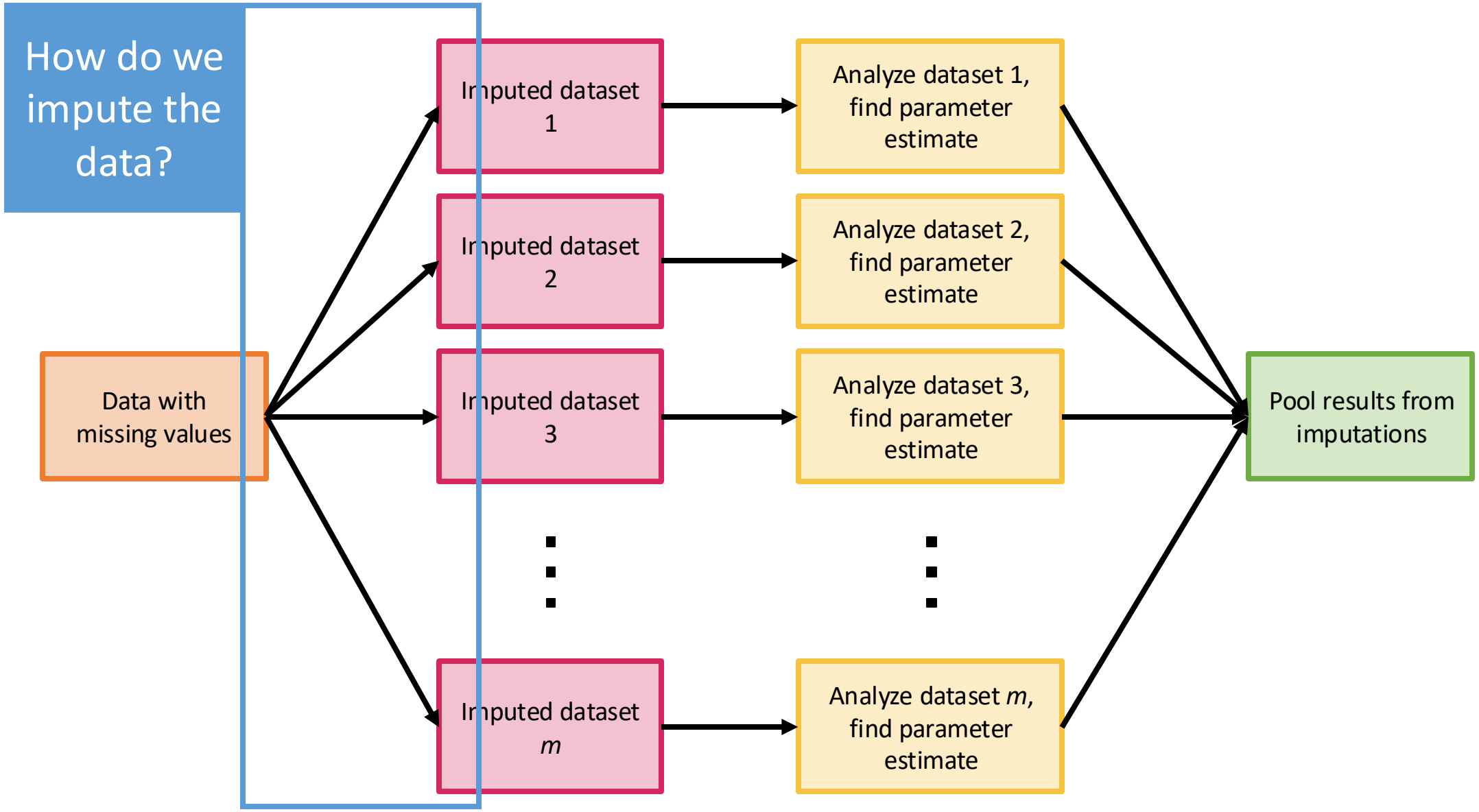
- Last observation carried forward (LOCF)
  - For longitudinal studies: replace missing data with previous observation
  - Example: height in child's growth trajectory

- Baseline observation carried forward (BOCF)
  - Similar to LOCF except you take the baseline value

Multiple imputation framework



Multiple imputation framework



Methods of imputation: multiple imputation

- Regression imputation/joint modeling
  - Use observed data to construct model to predict another variable
  - Then we use the prediction to impute data

- Hot-deck
  - Create a donor pool of observations that are similar to observation with missing data
  - Randomly pick from donor pool

- Each of the above can be used as a single imputation, but it does not capture to the uncertainty in the imputations!

How to pool results?

Rubin's Rules!!

- This applies to parameter estimates in regression
- Use parameter estimate and sampling variance for parameter estimate

Rubin's Rules

$$\bar{Q}_m = \frac{1}{m}\sum_{k=1}^{m}\hat{Q}_k$$

Number of imputations

Parameter estimate for imputation k

$$\bar{U}_m = \frac{1}{m}\sum_{k=1}^{m}U_k$$

Sampling variance of parameter estimate for imputation k

$$B_m = \frac{1}{m-1}\sum_{k=1}^{m}(\hat{Q}_k - \bar{Q}_m)^2$$

Variance of combined parameter estimate

$$T_m = \bar{U}_m + (1 + m^{-1})B_m$$

Combined variance

Butera et al, 2018

Reporting missing data

Report information on missing data!!

- It is important to be transparent about missing data

- At the very least, make sure the following is presented on missing data:
  - How much
    - Percentage for each variable
    - Percentage with incomplete data
  - Why
  - Potential consequences
  - Method to handle (can be complete case only)

Reporting guidelines from Van Buuren textbook

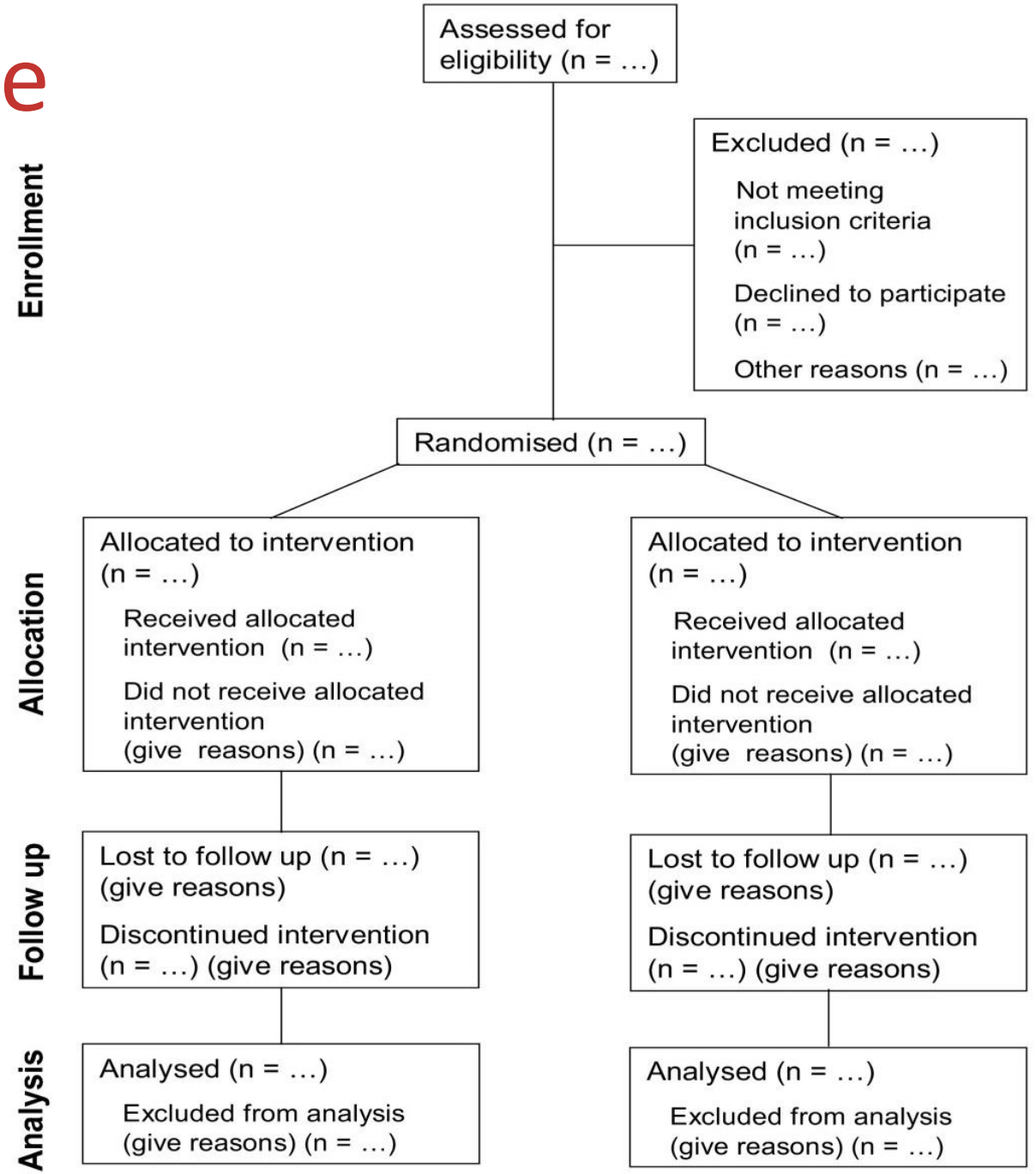
- Amount of missing data
- Reasons for missingness
- Consequences
- Method to account of missing data
- If imputing:
  - Software
  - Number of imputed datasets
  - Derived variables
  - Diagnostics
  - Pooling
  - Compare to complete case analysis
  - Sensitivity analysis

Example from Van Buuren textbook

> The percentage of missing values across the nine variables varied between 0 and 34%. In total 1601 out of 3801 records (42%) were incomplete. Many girls had no score because the nurse felt that the measurement was “unnecessary,” or because the girl did not give permission. Older girls had many more missing data. We used multiple imputation to create and analyze 40 multiply imputed datasets. Methodologists currently regard multiple imputation as a state-of-the-art technique because it improves accuracy and statistical power relative to other missing data techniques. Incomplete variables were imputed under fully conditional specification, using the default settings of the `mice` 3.0 package (Van Buuren and Groothuis-Oudshoorn 2011). The parameters of substantive interest were estimated in each imputed dataset separately, and combined using Rubin’s rules. For comparison, we also performed the analysis on the subset of complete cases.

Presenting your sample size

- **Good way to report reasons and amount**
- Based on CONSORT (Consolidated Standards of Reporting Trials)
- Present loss to follow-up or any other reasons that may result in smaller n for analysis
- May need figure out the reasons



https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-12-253

Break for activity part 2

Closing words

- If you take one thing from today, I hope it is that missing data is complicated
- If you're not sure how to address missing data, at least be transparent about the data and your procedure!
- Work with statisticians if you need help
  - Design a study that limits missing data
  - Address any missing data
- Great online textbook: Flexible Imputation of Missing Data