**CANTRANce:**

Simulation approaches when individual-level data are not available

Original approach:

1. Begin with covariate distributions reported in the literature
2. Generate a **single population** of the desired size with the provided covariate distributions
3. For each simulation, bootstrap from the generated population to get a simulation-specific population for analysis

Problem:

Joint distributions of covariates are rarely available. More often, we have a table giving the proportion of the population with each characteristic, for each covariate separately. Generating a single population and sampling from these data **imposes a single joint distribution** of covariates, which may or may not correctly represent the joint distribution in the actual study population.

Example:

Covariate distributions provided in literature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **N** |  | **Stage** | **N** |
| <65 | 4 |  | Loc/Reg | 7 |
| 65+ | 6 |  | Distant | 3 |

Expanded population

|  |  |
| --- | --- |
| **Age** | **Stage** |
| <65 | Loc/Reg |
| <65 | Loc/Reg |
| <65 | Distant |
| <65 | Loc/Reg |
| 65+ | Distant |
| 65+ | Loc/Reg |
| 65+ | Loc/Reg |
| 65+ | Distant |
| 65+ | Loc/Reg |
| 65+ | Loc/Reg |

Implied joint distribution

|  |  |  |
| --- | --- | --- |
| **Age** | **Stage** | **N** |
| <65 | Loc/Reg | 3 |
| <65 | Distant | 1 |
| 65+ | Loc/Reg | 4 |
| 65+ | Distant | 2 |

Proposed solution:

1. Begin with covariate distributions reported in the literature
2. **Separately** expand each covariate distribution to the desired population size, creating a set of independent covariate datasets
3. For each simulation, **bootstrap** from each of the expanded covariate datasets **individually**, **then combine** the sampled values to get a simulation-specific population for analysis

Example:

Expanded covariate datasets

|  |  |  |
| --- | --- | --- |
| **Age** |  | **Stage** |
| <65 |  | Loc/Reg |
| <65 |  | Loc/Reg |
| <65 |  | Loc/Reg |
| <65 |  | Loc/Reg |
| 65+ |  | Loc/Reg |
| 65+ |  | Loc/Reg |
| 65+ |  | Loc/Reg |
| 65+ |  | Distant |
| 65+ |  | Distant |
| 65+ |  | Distant |

Why it’s better:

The proposed approach **doesn’t impose any joint distribution** of covariates on the sampled data, but rather randomly samples from each covariate separately according to its frequency in the actual study population but with no relation to other covariates.

It should **more accurately reflect uncertainty** in estimates due to uncertainty in the joint distribution of covariates

Question:

Is there any difference between the proposed approach with

* a moderate sample size and a lot of simulations; or
* a huge sample size and only one simulation?