Roche Case Study in MHRA sandbox:
Using Synthetic Data to Enable Robust Subgroup Analysis in a Pivotal RRMM Trial

Federico Mattiello and Francesco Brizzi on behalf of the team (Gracy Crane, Felipe Castro, Ken Wang, Antoine Soubret, Laurent-Philippe Albou)

20 August 2026 | European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) 2026 Meeting

The Justification of Using Synthetic Data

Scientific Justification

Ensure treatment access for high-unmet-need of RRMM

Operational Justification

Overcome data access obstacles

The challenge to address

- **RRMM features a poor prognosis** and a complex, rapidly evolving treatment landscape.
- **Lack of a universally defined SoC** causes significant regional variation.
- This variation leads to fragmented enrollment and **underrepresentation of SoC B** in the control arm.
- Resulting imbalances **risk compromising** SoC B data interpretability, labeling discussions, and ultimately, **patient access.**

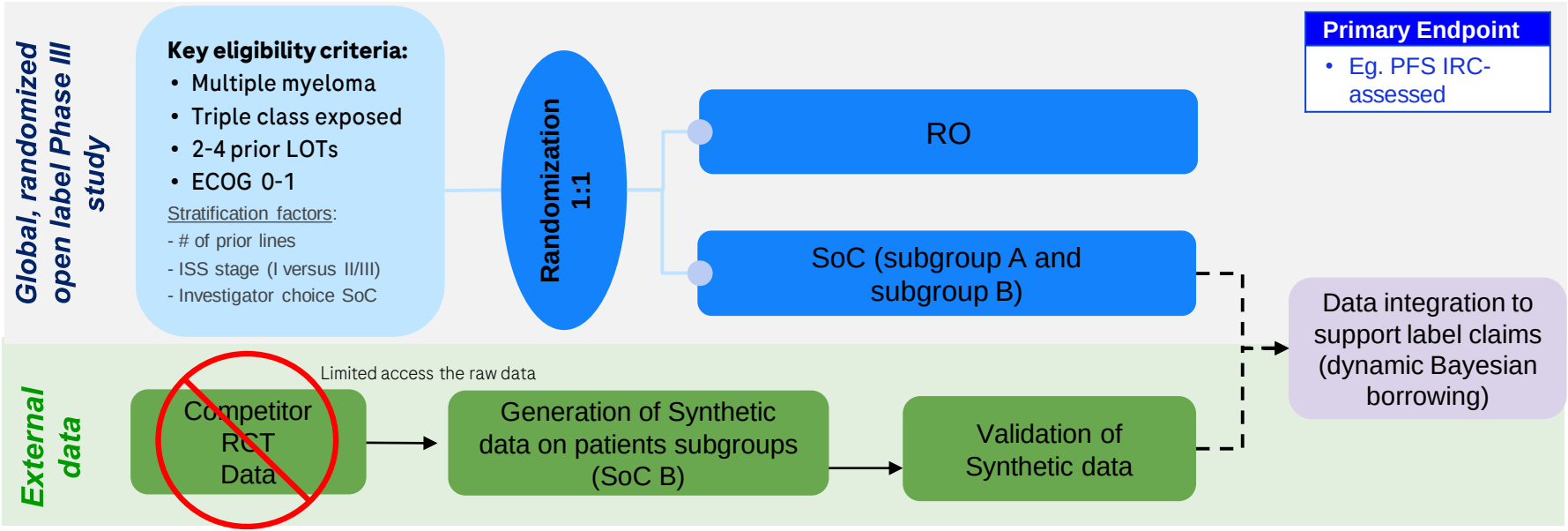
- **Access** to raw competitor data (SoC B) is **restricted** by IP and privacy regulations (GDPR/HIPAA).
- **Historical Phase 3 competitor data is the ideal** external source due to matching I/E criteria, PFS definitions, and assessment schedules.

Proposed solution

- **Augment** the limited **SoC B** population using **high-fidelity synthetic data** to decrease subgroup analysis uncertainty.
- This approach **compensates for underrepresented populations**, improving the overall representativeness of the evidence package.

- Create a **privacy-preserving synthetic dataset** that **retains** the **key propertie**s (e.g., multivariate distribution) of the raw data

(Hypothetical) Pivotal Ph3 Global Study Design Outline



The Challenge: As this pivotal trial anchors the market authorization package, **underrepresentation of SoC B** (driven by regional SoC differences) in the control arm leads to **high uncertainty in the RO to SoC B subgroup analysis**, risking a restricted final label and/or future reimbursement limitations.

The Goal: Use synthetic data to augment the trial data, **reduce uncertainty in subgroup effects**, and enable a robust statistical assessment of the RO vs. SoC B comparison.

Our Proposed Approach: Data Flow

Data Custodian

1. Identify the suitable source data *

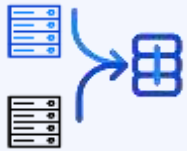
2.Generate synthetic data: *MIIC-SDG * or Synthpop*

3.Synthetic data quality assessment *

4.Transfer the synthetic data to Roche

Roche

5. Data integration: match covariates *via propensity score weighting* *****



6. Data integration: match outcomes *via dynamic Bayesian borrowing* *****

7. Present totality of information including subgroup (SoC B) analysis

**Dedicated slide to explain the approach in details*

Source Data Suitability

SOURCE DATA SELECTION

Phase 3 Competitor Clinical Trial Data

- Same RRMM patient population
- Similar inclusion/exclusion criteria (I/E)
- Same endpoints, and treatment definitions

PATIENT SELECTION

1. Defining critical variables:

Tier 1: Mandatory: match Phase 3 protocol exactly

Prior lines (2–4), class exposure (PI, IMiD, anti-CD38), ECOG (0–1)

Tier 2: Critical Covariates: propensity score weighting

Age, cytogenetic risk (del 17p, t(4;14)), R-ISS stage

Tier 3: Exploratory: context; no data exclusion

Race/ethnicity, country, specific prior drug names

2. Patient Selection Process

- Pass all 'mandatory' criteria
- Use Propensity Score Weighting to match study population distribution

RISK MITIGATION

Bias Mitigation

Propensity Weighting:

Match synthetic arm covariates to Roche trial population.

Sensitivity Analysis:

Accounting for previous treatment effects.

Dynamic Borrowing:

Integration of validated evidence.

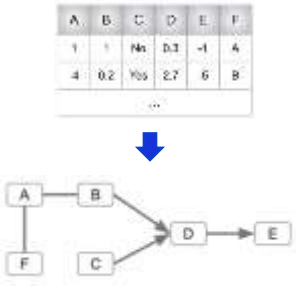
Privacy Preservation

MIIC-SDG: Algorithm proposed to balance data congruence and privacy preservation.

Synthetic Data Generation: MIIC-SDG

STEP 01

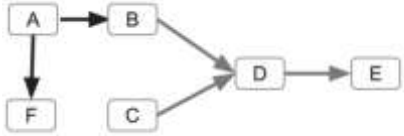
MIIC network reconstruction



- ❑ Infers a graph from raw data using information theory (MIIC algorithm)
- ❑ Nodes represent the variables of the data and edges represent direct associations between variables

STEP 02

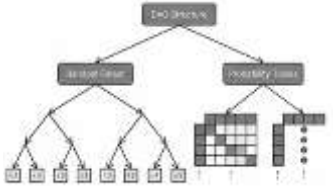
MIIC-DAG



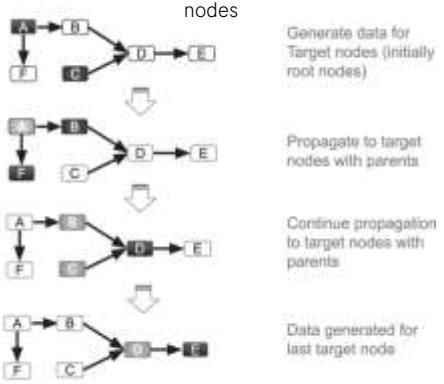
- ❑ Transforms the graph into a Directed Acyclic Graph (DAG), by determining cause-and-effect pathways
- ❑ Orient edges, removing cycles while preserving V structure
- ❑ Defines the exact order of operations

STEP 03

Generating Synthetic Data



Use random forest or probability table depending on the variable type of the target and parents nodes



Source: Sella, N. et al. (2025) npj Digital Medicine

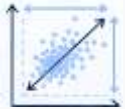
Synthetic Data Quality Assessment

Congruence

Assessing the alignment of marginal, bivariate and multivariate distributions

Coverage

Assessing the diversity captured: *variance, entropy, and range*



Constraints

Assessing the adherence to clinical constraints: *constraint violation rates*

Completeness

Assessing presence of lab tests and missing data rate

Compliance

Assessing the re-identification risk: *Identifiability, membership Inference score*

Consistency

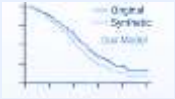
Check congruence and coverage, constraints across subpopulations

Comprehension

Clarity of data generation method and process documentation

Outcome Validation

Cox proportional hazard models fit on original and synthetic data. Assess similarity of model outcomes and estimates



Following 7Cs scorecard: Zamzmi, G. et al. (2025) Communications Engineering

Synthetic and RCT Data Integration

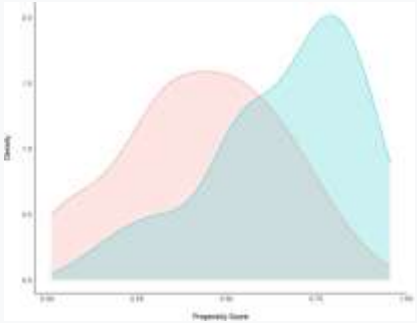
The proposed approach:

Potential issue: synthetic SOC_B population might not match the RCT population in terms of: important baseline covariates and/or outcomes.
Solution: Combine synthetic and RCT data, with a two-step approach to mitigate potential biases.

Step 1: For baseline covariates

Propensity score weighting with *Inverse Probability of Treatment Weight (ATT)* which uses a logistic regression model to estimate the probability of being in the RCT SOC_B population (p_i)

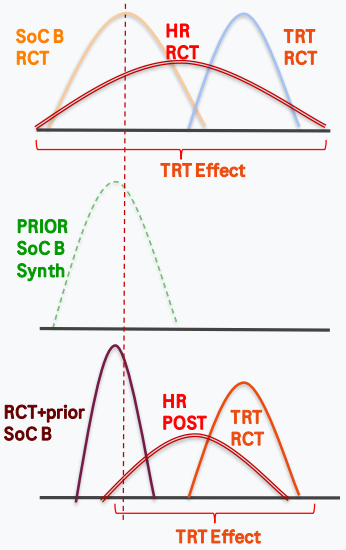
- Each patient in the synthetic data gets a weight of $p_i/(1-p_i)$. RCT patients gets a weight of 1



Step 2: For outcomes

Bayesian dynamic borrowing to combine synthetic and RCT data to estimate trt-effect

- An **informative power prior** assigned to the SOC_B-effect
  - **Mean: "best-guess"** of SOC_B-effect obtained by fitting the same model to weighted synthetic data
  - **Standard deviation:** chosen to cap maximum Type I error inflation at a pre-specified limit (e.g. 20%)
- The **dynamic scaling factor** controlling the amount of borrowing (between 0-1) is estimated using the empirical bayes approach



Evaluation of New treatment vs SOC B

Success: if high confidence that new treatment reduces the risk of progression compared to SOC_B on integrated analysis of synthetic and RCT data.
Statistically: Posterior probability of treatment effect (new treatment - SOC_B) > alpha (e.g. 80%).

Propensity score (Bottigliengo, D. et al., BMC Med Res Methodol. 2021, https://doi.org/10.1186/s12874-021-01454-z) | Dynamic borrowing (Wiesenfarth M. and Calderazzo S., Biometrics 2020, https://doi.org/10.1111/biom.13124)

Conclusions / discussion points

Some questions are not fully answered by RCT data alone (subgroups, rare populations…)
These questions are still being asked and are pivotal to drug development.

Regulatory opportunities:

- Importance of guidance on how innovative approaches can be implemented within a regulatory framework
- Co-creation and continuous global stakeholder engagement – dialogue is fundamental

Synthetic data opportunities:

- Synthetic data is a broad term: irrespective of how it is built (AI, stats, PK/PD modelling) the application is key.
- Could ease privacy / data-access constraints and "common" models could be developed and leveraged

Statistical opportunities:

- Importance of developing E2E approach (method agnostic) to answer these (hard) questions appropriately
- Definition of: acceptable quality for SD, pre-specification for integration of SD, success criteria

How to get the ball rolling:

- Gauge interest from regulators as to which agencies would consider these types of approaches
- Run the pre-specified synthetic analysis, but still collect the real data – would we reach the same conclusion?
- Collect tangible examples