RTI (h)(s)
Health Solutions

March 25, 2025

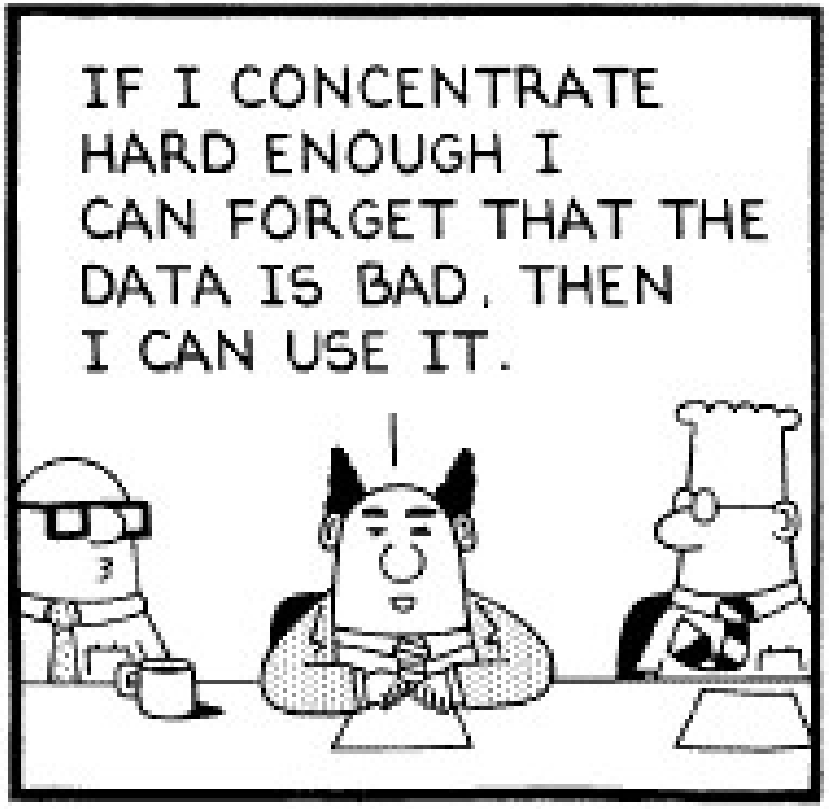
ARS Toscana

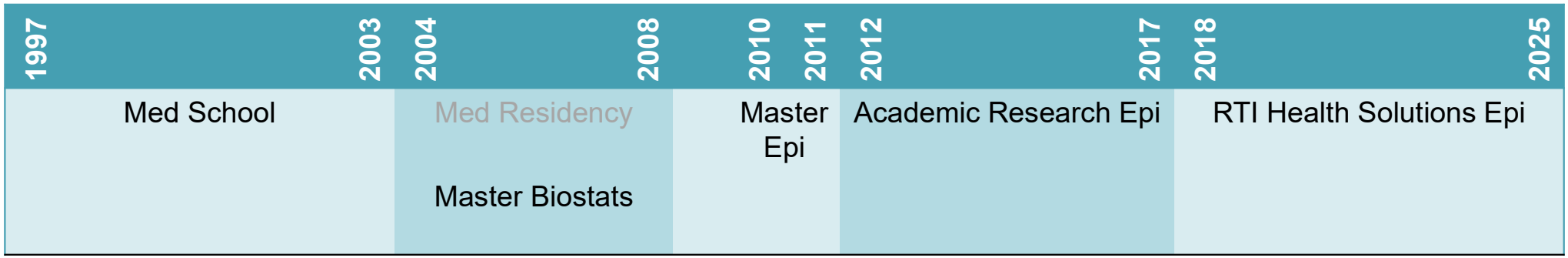
Validate study variables to reduce misclassification bias: recent tools and research needs

Discussion

The power of **knowledge.**
The value of **understanding.**

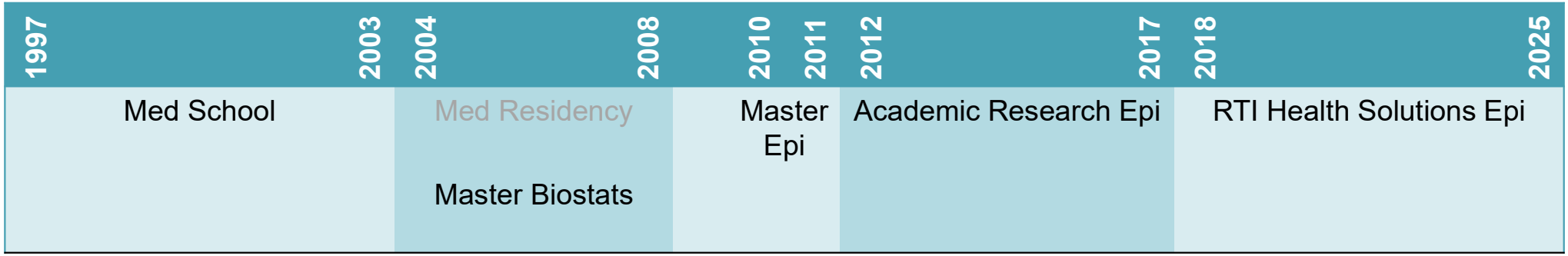
Actual footage of myself as postdoc





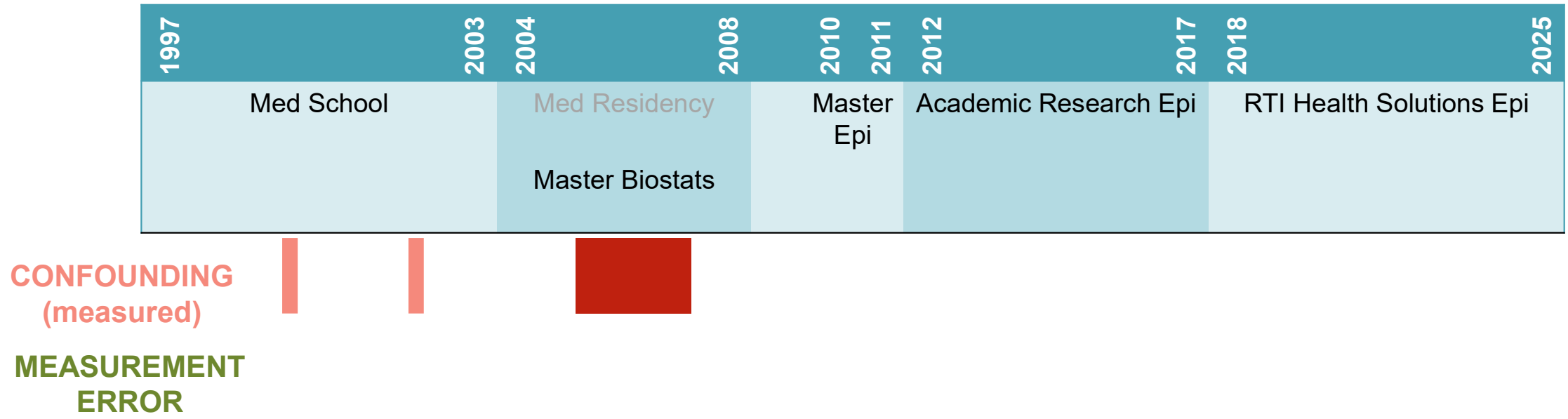
CONFOUNDING (measured)

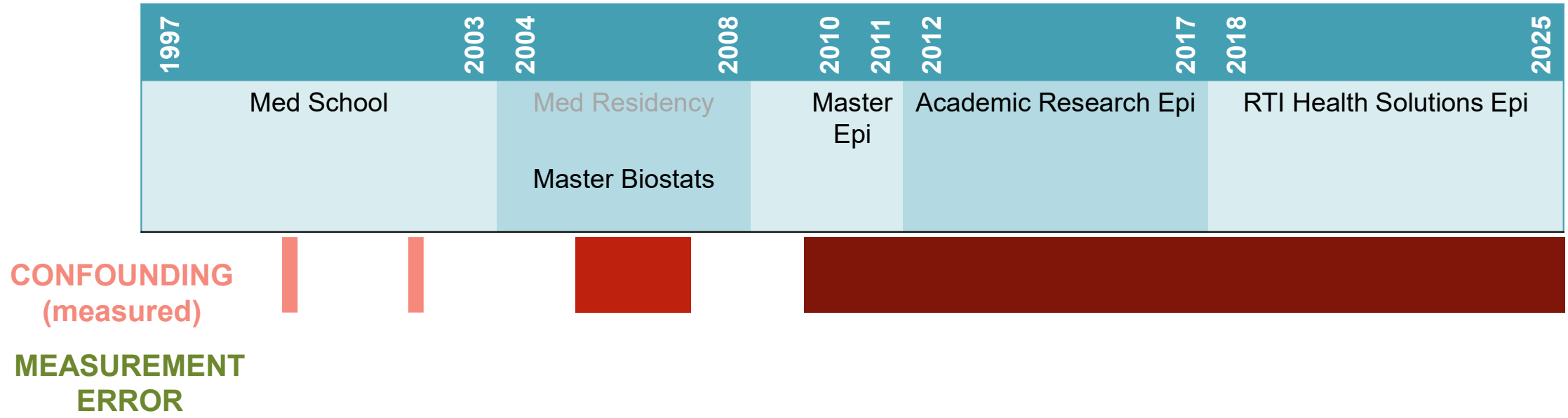
MEASUREMENT ERROR

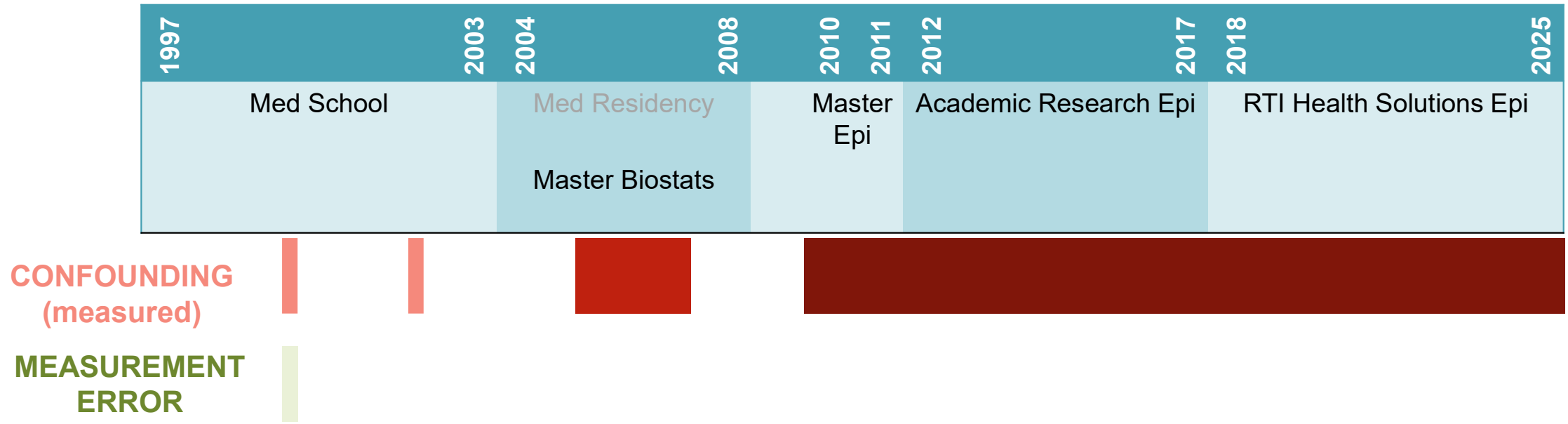


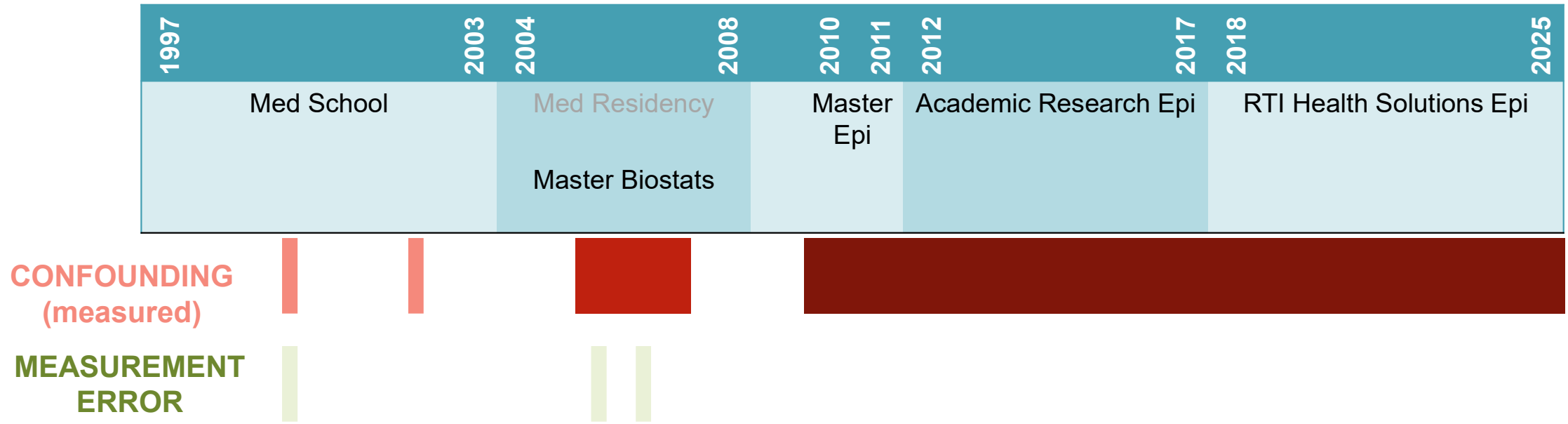
CONFOUNDING (measured)

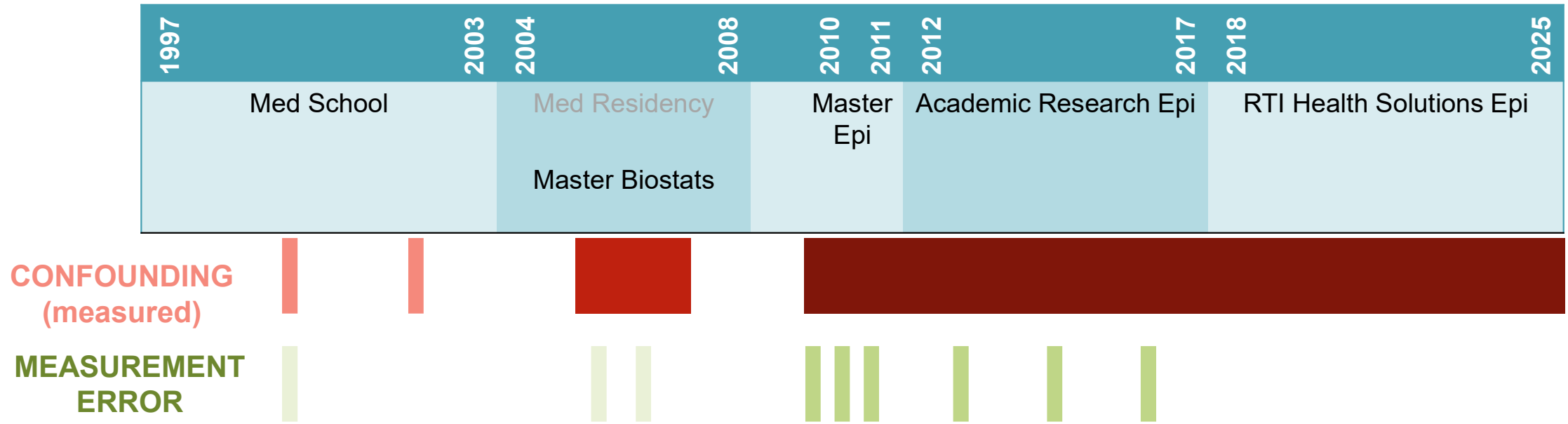
MEASUREMENT ERROR

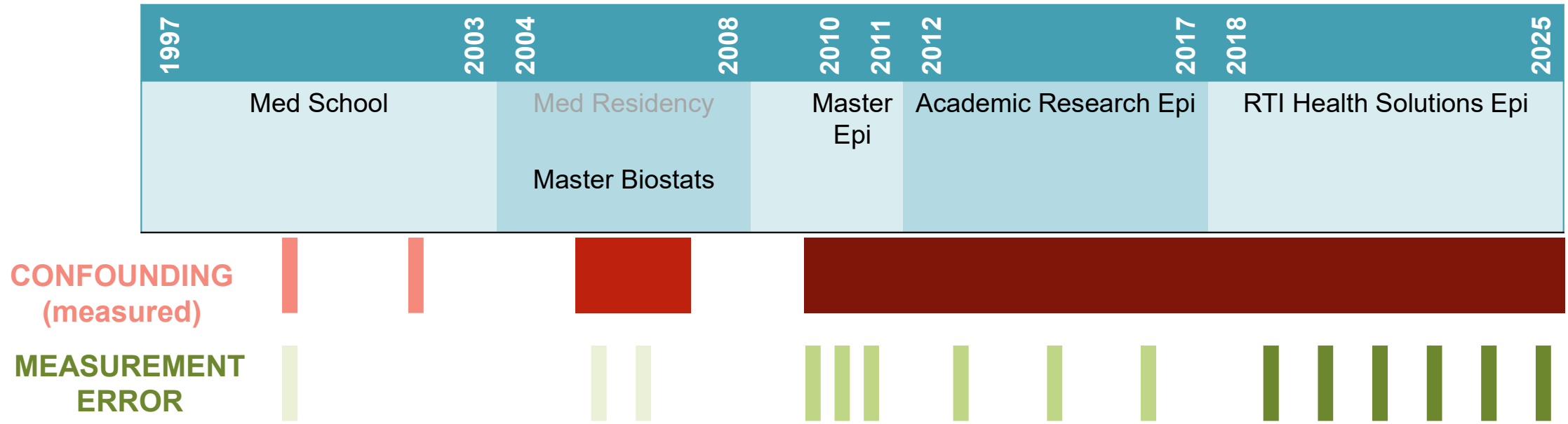












Research question: what is the effect of drug x on the incidence of malignancies compared to no treatment in patients with a rheumatoid arthritis

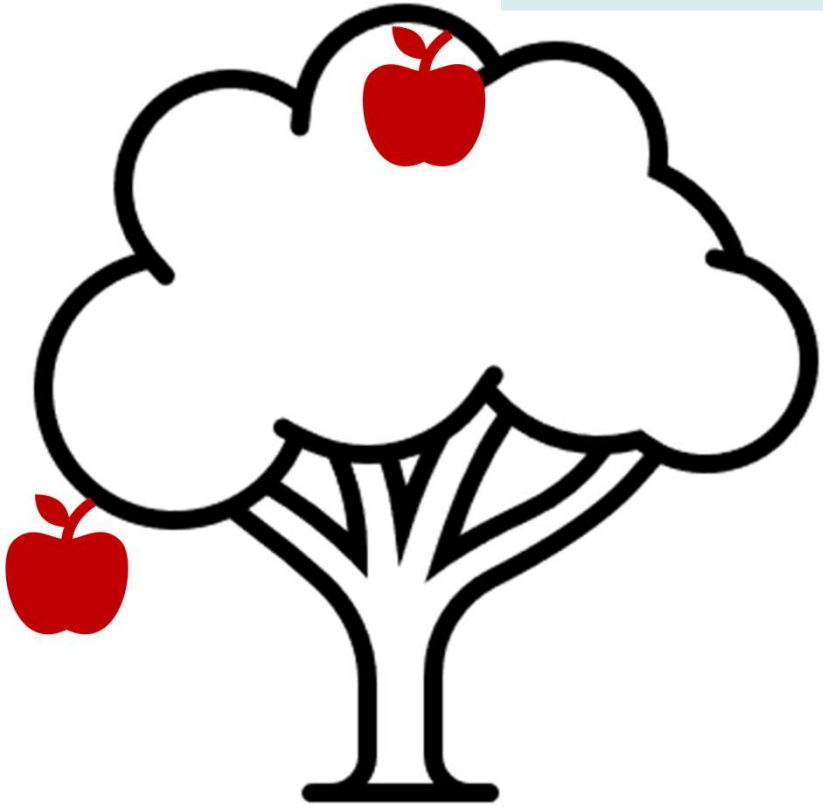
Component	Target Trial Specification	Target Trial Emulation using CPRD
Eligibility	Individuals aged 18+, diagnosed with rheumatoid arthritis Treatment naïve	Same as target trial
Treatment strategies	1. Initiate drug x 2. Do not initiate drug x	Same as target trial
Assignment procedures	Participants are randomly assigned to either strategy at baseline and are aware of the assignment	Patients are classified according to the strategy that their data were compatible with at baseline and attempted to emulate randomizatoin by adjusting for baseline confounders (1:1 matching)
Follow-up	Follow-up starts at randomization and ends at cancer diagnosis, death, loss to follow-up, 5 years after baseline or end of Study period, whichever is first.	Same as target trial
Outcome	Cancer diagnosis within 5 years	Same as target trial
Causal contrast	Intention-to-treat effect Per protocol effect	Observational analogue of the per-protocol effect
Analysis plan	Comparison of 5-year CRC risks among individuals assigned to each treatment strategy. If some baseline variables are unbalanced, the risks are estimated within levels of the baseline variables and subsequently standardized. Will assume no right censoring: 2x2 table	Same per-protocol analyses with sequential emulation and adjustment for baseline variables

Advice from the experts in the room

- Vera Ehrenstein
  - Are there validated algorithms for your variables?
    - If yes, is their validity satisfactory given study aims and variable roles
    - If no, is a bespoke validation study possible?
  - If resources are limited, what is the least that can be done?
- Elisa Martín Merino
  - Estimate PPV/other validation parameters by exposure categories to detect differential misclassification and correct the estimates accordingly (even in Target Trial Emulations)

Motivation

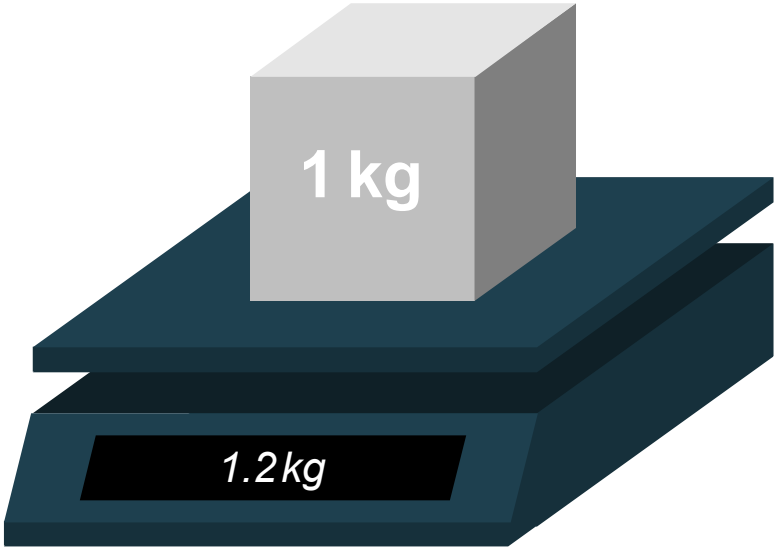
Solving bias due to unmeasured confounding, **measurement error**, missing values



Solving bias due to selection, lack of consistency, measured confounding

Measurement Error

- Measurement error refers to the situation where the measured value and the true value of a variable are substantially different
- When the variables being measured are categorical (e.g., the presence of comorbidities), these errors are referred to as “misclassification”
- We can have misclassification of outcomes, exposure and confounders



Advice from the experts in the room

- Vera Ehrenstein
  - Are there validated algorithms for your variables?
    - If yes, is their validity satisfactory given study aims and variable roles
    - If no, is a bespoke validation study possible?
  - If resources are limited, what is the least that can be done?
- Elisa Martín Merino
  - Estimate PPV/other validation parameters by exposure categories to detect differential misclassification and correct the estimates accordingly (even in Target Trial Emulations)
- Giulia Hyeraci
  - How can we adopt the perspective of coders to improve the identification of the event in the healthcare administrative data? Are there any tools other than interviews that can be used for this purpose?
- Giuseppe Roberto
  - Use the DIVERSE framework to represent and describe data source diversity
  - Apply the component strategy to key study variables (e.g. eligibity event / outcome) to support the generation of approximate estimates of validity

Clinical Practice Research Datalink data

- The CPRD covers about 7% of the U.K. population
- Contains:
  - Issued prescriptions
  - Read codes for diagnoses, signs, symptoms, referrals, test requests, and test results, as well as free-text comments
- About 75% of English practices contributing to the CPRD have consented to have their patients' information linked to other health care data sets:
  - Hospital Episode Statistics
  - National Cancer Data Repository
- In Hospital Episode Statistics and the National Cancer Data Repository, diagnoses are recorded using the International Statistical Classification of Diseases and Related Health Problems, 10th Revision

Malignancies and multiple sources of data, UK CPRD

Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. **Epidemiology. 2018;29(2):308-13**
https://journals.lww.com/epidem/Fulltext/2018/03000/Validation_of_Cancer_Cases_Using_Primary_Care,.19.aspx

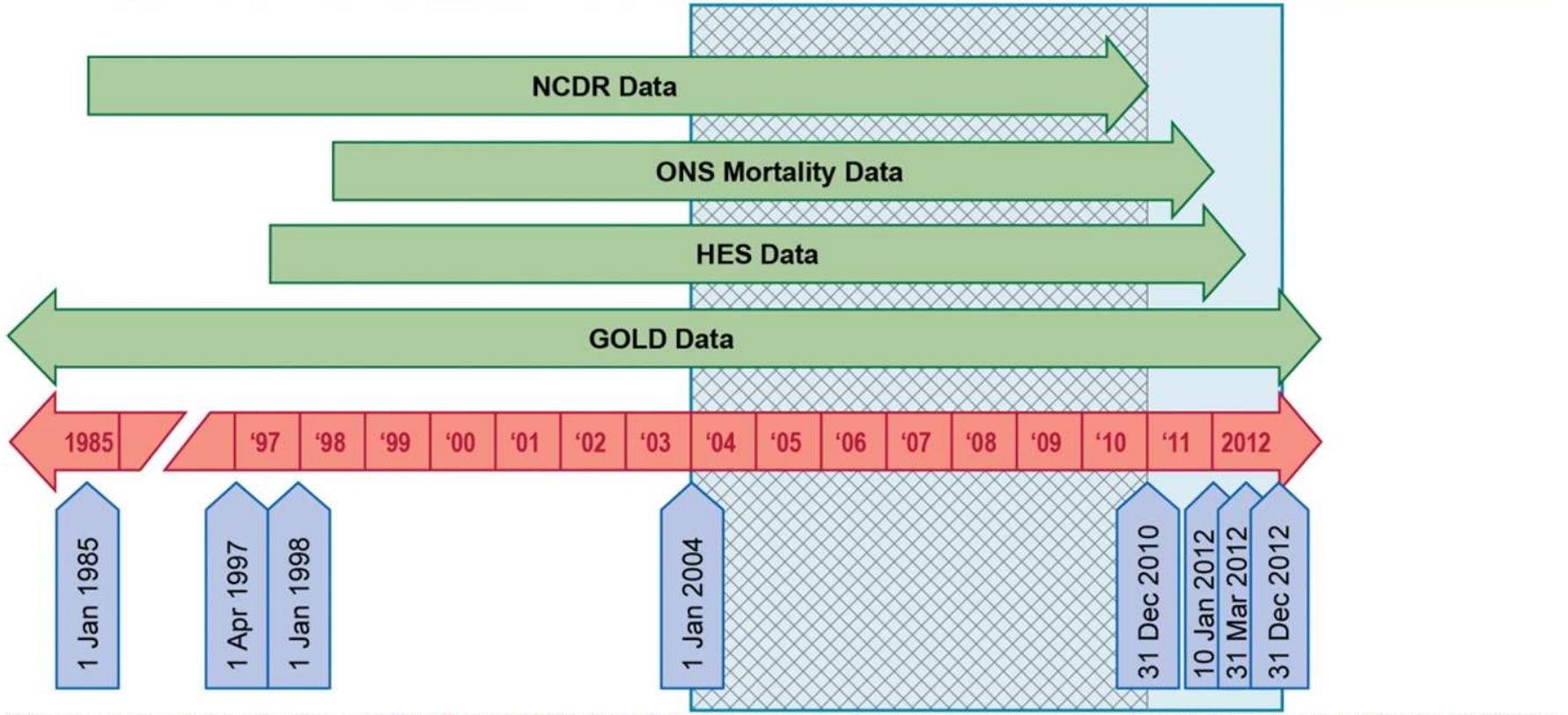
- GOLD data only (online general practice database)
- GOLD data plus linkages to Office for National Statistics (ONS) for mortality data, Hospital Episode Statistics (HES) for data on hospitalizations, and/or cancer registries (NCDR)

Malignancies, CPRD - Methods

- Clinical Practice Research Datalink data (2004-2012)
- Patients treated with overactive bladder medications
- Identified provisional cases of 10 common cancers in General Practitioner Online Database (GOLD)
- Validated them by medical profile review.
- In practices with linkage to Hospital Episodes Statistics and National Cancer Data Repository (2004-2010),
  - Also validated provisional cancer cases against these sources
  - Allowed to identify additional cancer diagnoses not recorded in GOLD

Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018;29(2):308-13

Data Source Coverage in Relation to the Study Period



Data source coverage in relation to the study period. GOLD indicates General Practitioner Online Database; HES, Hospital Episode Statistics; NCDR, National Cancer Data Repository; ONS, Office for National Statistics.

Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018;29(2):308-13

Malignancies, CPRD - Results

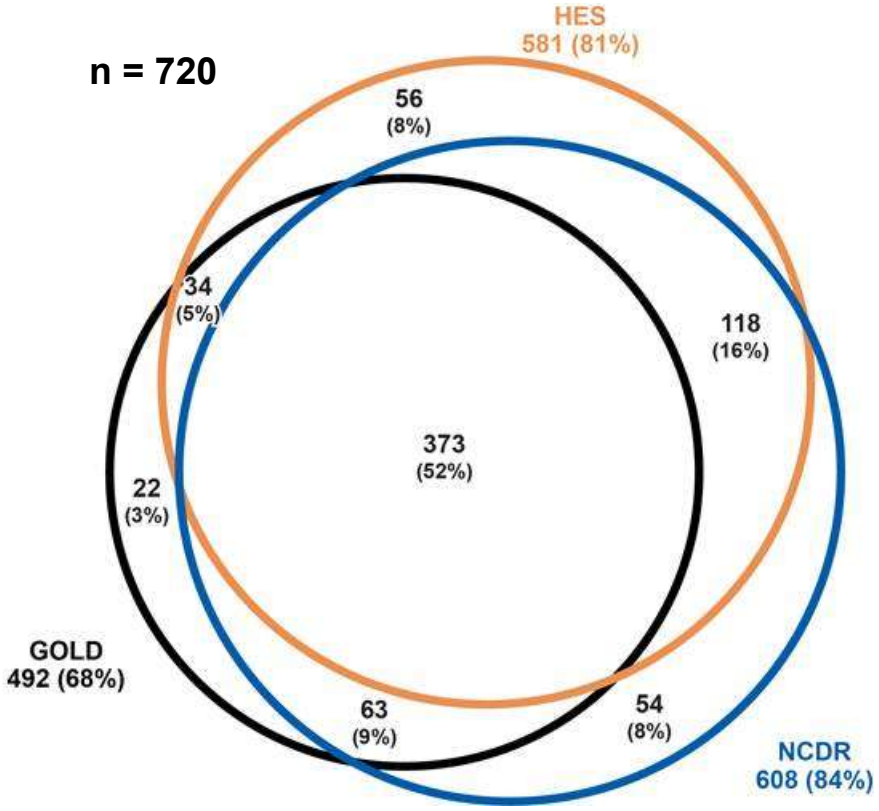
- Among 50,840 patients, 1,486 provisional cancer cases were identified
  - Nonlinked practices 93% of 661 cases confirmed in medical review
    - Range, 100% of non-Hodgkin lymphomas and uterine cancer to 77% of skin melanomas
  - Linked practices 96% of 825 cases confirmed in medical review
    - Range: 100% of kidney and uterine cancers to 92% of melanomas
- Most cases of cancer identified electronically in the GOLD were confirmed.
- A substantial proportion of cases, especially of cancer types not typically managed by general practitioners, would be missed without Hospital Episodes Statistics and National Cancer Data Repository data (and are likely missed in nonlinked practices).

Margulis AV, Fortuny J, Kaye JA, Calingaert B, Reynolds M, Plana E, et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018;29(2):308-13

Figure 2. Origin of Cancer Cases Diagnosed During Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices, by Data Source, All Study Cancers Combined

CPRD = Clinical Practice Research Datalink; GOLD = primary care data from the CPRD; HES = Hospital Episode Statistics; NCDR = National Cancer Data Repository.

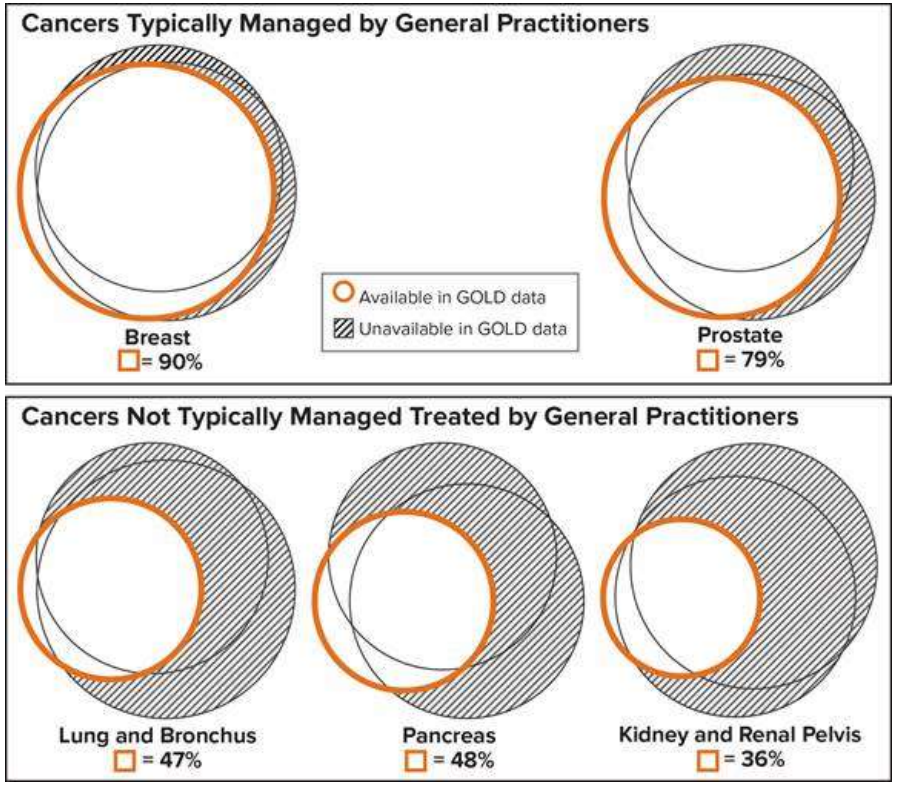
Note: This figure represents the 720 confirmed cases in linked practices, regardless of the data source in which the cases were initially identified. Percentages were calculated using 720 as the denominator.



Margulis AV et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018;29(2):308-13

Figure 3. Selected Cancers by Main Treating Physician: Percentage of Cases Identifiable in GOLD During the Period of Complete Overlap of Data Sources (2004-2010) in Linked Practices

CPRD = Clinical Practice Research Datalink;
GOLD = primary care data from the CPRD;
GP = general practitioner;
HES = Hospital Episode Statistics.



Margulis AV et al. Validation of Cancer Cases Using Primary Care, Cancer Registry, and Hospitalization Data in the United Kingdom. Epidemiology. 2018;29(2):308-13

Research question: what is the effect of drug x on the incidence of malignancies compare to no treatment in patients with a rheumatoid arthritis

Component	Target Trial Specification	Target Trial Emulation using CPRD
Eligibility	Individuals aged 18+, diagnosed with rheumatoid arthritis Treatment naïve	Same as target trial
Treatment strategies	1. Initiate drug x 2. Do not initiate drug x	Same as target trial
Assignment procedures	Participants are randomly assigned to either strategy at baseline and are aware of the assignment	Patients are classified according to the strategy that their data were compatible with at baseline and attempted to emulate randomizatoin by adjusting for baseline confounders (1:1 matching)
Follow-up	Follow-up starts at randomization and ends at cancer diagnosis, death, loss to follow-up, 5 years after baseline or end of Study period, whichever is first.	Same as target trial
Outcome	Cancer diagnosis within 5 years	Same as target trial
Causal contrast	Intention-to-treat effect Per protocol effect	Observational analogue of the per-protocol effect
Analysis plan	Comparison of 5-year CRC risks among individuals assigned to each treatment strategy. If some baseline variables are unbalanced, the risks are estimated within levels of the baseline variables and subsequently standardized. Will assume no right censoring: 2x2 table	Same per-protocol analyses with sequential emulation and adjustment for baseline variables **Validate the outcome, provide validity measures**

Advice from the experts in the room

- Vera Ehrenstein
  - Are there validated algorithms for your variables?
    - If yes, is their validity satisfactory given study aims and variable roles
    - If no, is a bespoke validation study possible?
  - If resources are limited, what is the least that can be done?
- Elisa Martín Merino
  - Estimate PPV/other validation parameters by exposure categories to detect differential misclassification and correct the estimates accordingly (even in Target Trial Emulations)
- Giulia Hyeraci
  - How can we adopt the perspective of coders to improve the identification of the event in the healthcare administrative data? Are there any tools other than interviews that can be used for this purpose?
- Giuseppe Roberto
  - Use the DIVERSE framework to represent and describe data source diversity
  - Apply the component strategy to key study variables (e.g. eligibity event / outcome) to support the generation of approximate estimates of validity
- Anna Schultze
  - If you are doing a validation study, report validity measures by exposure / outcome
  - Pre-specify when you will run a QBA, and your bias parameters
  - To incorporate uncertainty, implement a PBA

Research question: what is the effect of drug x on the incidence of malignancies compare to no treatment in patients with a rheumatoid arthritis

Component	Target Trial Specification	Target Trial Emulation using CPRD
Eligibility	Individuals aged 18+, diagnosed with rheumatoid arthritis Treatment naïve	Same as target trial
Treatment strategies	1. Initiate drug x 2. Do not initiate drug x	Same as target trial
Assignment procedures	Participants are randomly assigned to either strategy at baseline and are aware of the assignment	Patients are classified according to the strategy that their data were compatible with at baseline and attempted to emulate randomizatoin by adjusting for baseline confounders (1:1 matching)
Follow-up	Follow-up starts at randomization and ends at cancer diagnosis, death, loss to follow-up, 5 years after baseline or end of Study period, whichever is first.	Same as target trial
Outcome	Cancer diagnosis within 5 years	Same as target trial
Causal contrast	Intention-to-treat effect Per protocol effect	Observational analogue of the per-protocol effect
Analysis plan	Comparison of 5-year CRC risks among individuals assigned to each treatment strategy. If some baseline variables are unbalanced, the risks are estimated within levels of the baseline variables and subsequently standardized. Will assume no right censoring: 2x2 table	Same per-protocol analyses with sequential emulation and adjustment for baseline variables **Validate the outcome, provide validity measures** **Implement a probabilistic bias analysis**

Advice from the experts in the room

- Giorgio Limoncella
  - Is it possible to conduct a validation study for the outcome variable?
  - Can this be done stratified by exposure status?
  - And if so, could a screening algorithm be defined that, even with limited specificity, would still identify some true cases missed by the primary algorithm?
  - And if so, can you **assume** that sensitivity is non-differential?
- Marco Lippi
  - Is it safer to have the LLM stick with the information in the clinical record?
  - Would it be suitable to run multiple models in order to mitigate variance?
  - Would it be suitable to run different models for different clinical cases?
  - Would it be reasonable to know in advance whether the patient belongs to the set of narrow or possible patients, or not?

Research question: what is the effect of drug x on the incidence of malignancies compare to no treatment in patients with a rheumatoid arthritis

Component	Target Trial Specification	Target Trial Emulation using CPRD
Eligibility	Individuals aged 18+, diagnosed with rheumatoid arthritis Treatment naïve	Same as target trial
Treatment strategies	1. Initiate drug x 2. Do not initiate drug x	Same as target trial
Assignment procedures	Participants are randomly assigned to either strategy at baseline and are aware of the assignment	Patients are classified according to the strategy that their data were compatible with at baseline and attempted to emulate randomizatoin by adjusting for baseline confounders (1:1 matching)
Follow-up	Follow-up starts at randomization and ends at cancer diagnosis, death, loss to follow-up, 5 years after baseline or end of Study period, whichever is first.	Same as target trial
Outcome	Cancer diagnosis within 5 years	Same as target trial
Causal contrast	Intention-to-treat effect Per protocol effect	Observational analogue of the per-protocol effect
Analysis plan	Comparison of 5-year CRC risks among individuals assigned to each treatment strategy. If some baseline variables are unbalanced, the risks are estimated within levels of the baseline variables and subsequently standardized. Will assume no right censoring: 2x2 table	Same per-protocol analyses with sequential emulation and adjustment for baseline variables **Validate the outcome, provide validity measures** **Implement a probabilistic bias análisis** **Be explicit about the assumptions behind the bias análisis** **(stay curious about the role of LLM)**

Further points of discussion

- What about treatment?
  - Are there challenges specific to the validation of the exposure? (e.g., a drug is prescribed but not dispensed, or dispensed but not swallowed…)
  - How can we handle the validation of a treatment that is sustained over time?

- What about confounders?
  - Is there measurement bias if treatment decision is based only on the mismeasured variable (not the true variable)?