plusproject

AI & ML SIG: PRACTICAL AI AND DEVELOPMENTS IN MACHINE LEARNING

AI & ML SIG – Sam Hadlington, Lesedi Ledwaba-Chapman, Paola Berchialla, Harry Parr, Jason Nicholas

Proudly employee owned.


plusproject
INTRODUCTION TO AI & ML SIG
Proudly employee owned.

plusproject

GOALS

- Aligning machine learning groups across SIGs
- In context of clinical research how and where are AI and ML useful –specifically related to statisticians
  - Where are they not useful
- Longer term goals:
  - Wonderful Wednesday style regular webinars
  - Technical group to develop tools in R
  - Develop common framework for explainable AI in statistics

UPDATES

- Members: 23
- Holding quarterly meetings
- Creation of sub-teams – current active sub-teams:
  - R tools for machine learning
  - SAP generator
  - Using AI for HTA
- New sub-team ideas:
  - AI&ML Regulation
  - Ethics
  - What can AI not do
  - …

plusproject

AI CONSIDERATIONS FOR THE FUTURE

- How can we learn to integrate this technology with our day to day work to stay ahead.
- Is AI a threat for job security?
  - How much of our current work will be done by AI systems? Coding? SAP/Protocol writing? Email correspondence?
  - What can AI not do? Can we identify these areas and become more competent in them to reduce the job risk?
- AI Agents - how effective can these models be?

GETTING INVOLVED

- Please join us for our SIG lounge slot tomorrow lunchtime
- Find me on the conference app and reach out
- Email me at sam.Hadlington@plus-project.co.uk

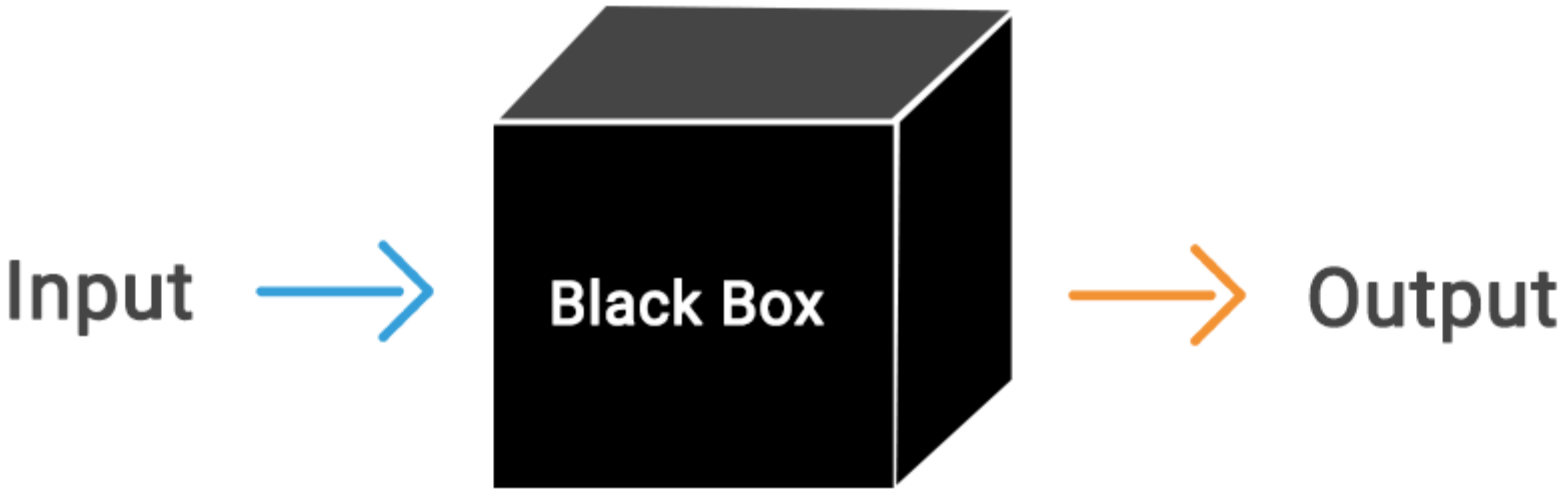
SAP-GPT: USING MODERN AI FOR DOCUMENT GENERATION

CONTENTS

DEVELOPMENT PROCESS

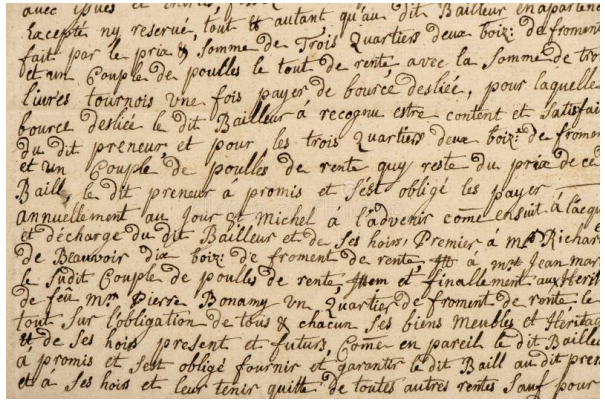
GOALS

- Find an easy to follow and repeatable method to generate a draft SAP using an LLM with a protocol and prompts.
- Attempt to find the easiest and quickest method possible that provides a good enough output to work with.
- Test across several LLMs to see which one works the best.

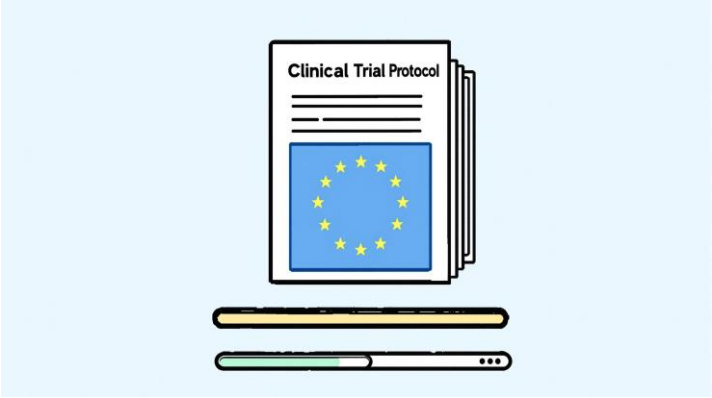


SINGLE PROMPT APPROACH

- One large single editable prompt input to attempt to produce a full SAP draft in one go
- Initial small prompt and protocol – much easier more open approach
- Developed longer more complex and more specific prompt



+

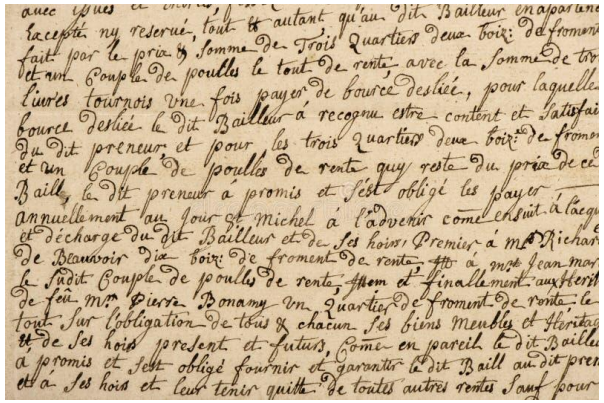


=

SAP

SINGLE PROMPT APPROACH

- Added SAP template to further direct Model
- Added Old SAP to give model idea of expectations from SAP
- Final single prompt process: Prompt added, Protocol added, Template added, old SAP added, confirmation, SAP produced

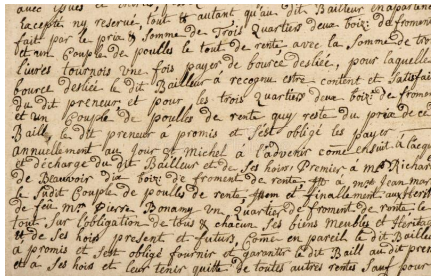


+ & = SAP

+ Old SAP

MULTIPLE PROMPT APPROACH

- Got some more insight after speaking to AI expert
- One large prompt to begin with, then one prompt per section/sub-section.
- We asked the LLM to ask us for a prompt for each section, we then would write what we wanted for each section. Quite time consuming.



+

&

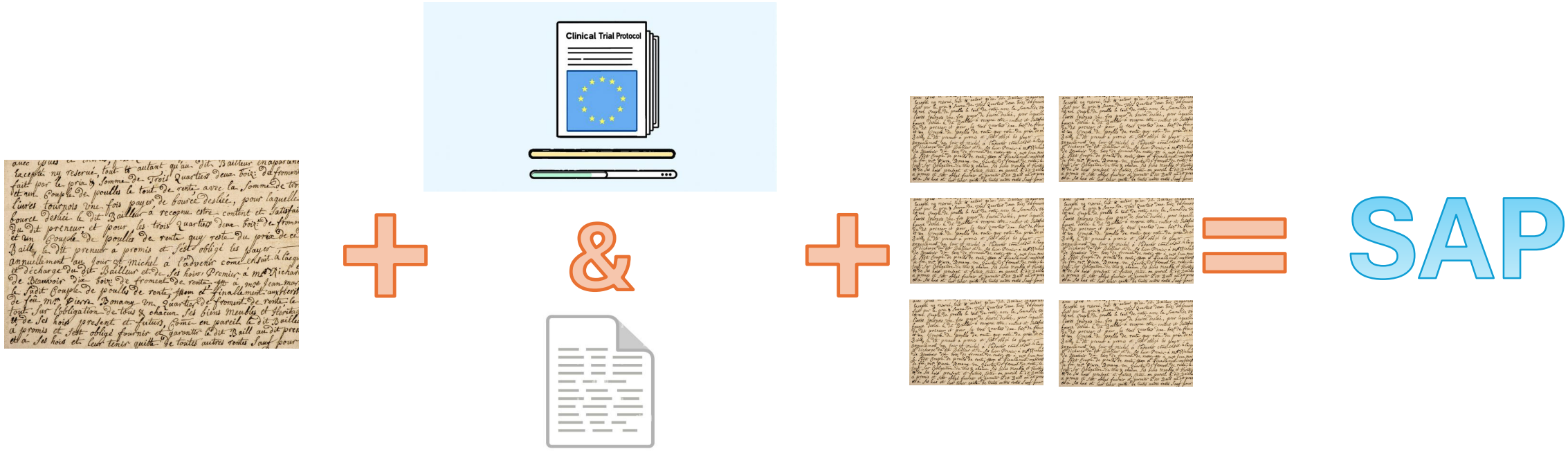
+ Old SAP

+

= SAP

MULTIPLE PROMPT APPROACH

- We then wrote a standard prompt for each section/sub-section throughout the template.
- We also removed the old SAP as the model was drawing data from this into the new SAP.
- This provided a much simpler approach as we would just copy the standard prompt text from the template.
- We feel this approach could be used when producing new SAP templates.



EXCERPT OF PROMPT

You are an expert clinical trial statistician helping to draft an industry-standard Statistical Analysis Plan (SAP) from a study protocol. The user will provide: 1. A new protocol to transform into a SAP. 2. A SAP template outline (legend of required sections in text or in a word document). 3. At each new section the user will provide a section specific prompt. Please ask for this prompt before proceeding with the generation of this section. Workflow: Go through each SAP template section in order. For each SAP template section, identify and present all relevant passages from the new protocol only. Before scanning the full protocol, ask the user for the section specific prompt. If the user does not provide them, restrict your initial search to the table of contents, section titles, and any sections whose titles include the SAP template section keywords (e.g., 'Objectives', 'Endpoints', 'Sample Size'). Only perform a full-document search if the user confirms it's necessary. Draft a candidate SAP section in professional, industry-standard style, mimicking the format of the SAP template and the style of the new protocol. Ask the user whether any protocol sections need to be referenced or dropped and whether they want to accept, edit, or expand the draft. Do not move onto the next section until the user approves the drafted section. At the end of each section generate a word document of that section. Keep track of all approved sections and their numbering in memory during the workflow so that you can reproduce the complete SAP structure on request without re-parsing previous text. Rules: Never invent objectives, endpoints, or methods not present in the protocol or provided examples. The new protocol is the only valid source of factual content. If a section cannot be fully populated, insert a placeholder [MISSING: ...]. Keep the outputs clean and concise (headings, bullet points where appropriate). Always show the relevant new protocol passages before drafting the section. Ask the user if they want more detail or expansion before finalising a section. Present relevant passages and draft sections in Markdown format. If the user asks for the drafted sections to be placed into a word document, replicate the SAP template structure, add the newly drafted section and include previously approved sections, remove extra blank lines but preserve single line breaks within paragraphs, use the Markdown format displayed in ChatGPT (Maintain bold, italics, and bullet formatting) and use Calibri.

2.2 Randomization and Blinding

Create Section 2.2 Randomization and Blinding by describing the approaches to blinding and randomization in the protocol. If there is no randomization or blinding then please state "No randomization or blinding will be used in this study.". Use the protocol as a reference for this but expand upon the text if you think it's appropriate to provide more detail or statistical rationale.

3. OBJECTIVES, ENDPOINTS AND ESTIMANDS

Create Section 3. Objectives, Endpoints and Estimands by briefly describing the objectives and/or endpoints outlined in the protocol. Do not embellish the wording from the protocol, keep it as close to what was described in the protocol as possible. Do not start any subsections yet, they have their own specific prompt.

3.1 Primary Estimand(s)

Create Section 3.1 Primary Estimand(s) by copying the primary estimand from the protocol. It is very important that this is the same here as it is in the protocol. If there is no primary estimand in the protocol then just generate the following text "No primary estimand was produced for this study".

3.2 Secondary Estimand(s)

Create Section 3.2 Secondary Estimand(s) by copying the secondary estimand from the protocol. It is very important that this is the same here as it is in the protocol. If there is no secondary estimand in the protocol then just generate the following text "No secondary estimand was produced for this study".

BEING SPECIFIC WITH PROMPTING

- Very important to tell model **exactly** what you're looking for, don't assume anything.
- When building a prompt make sure to update old text if your approach changes, even a little.
- Something you see as benign or clearly to be ignored in certain situations in a prompt may be read differently by an LLM which can drastically change your output.

USAGE LIMITS

- Protocols have large context windows – a lot of data to process
- Hit usage limits several times, had to wait for this to reset
- We used “Pro” accounts for each LLM which sped things up
- Might be different for “Enterprise” accounts

Findings

Methodology

Benchmarking: The human-authored, finalized SAP is established as the gold standard.

Anonymisation: All identifying markers of which LLM produced the SAP is removed to prevent bias.

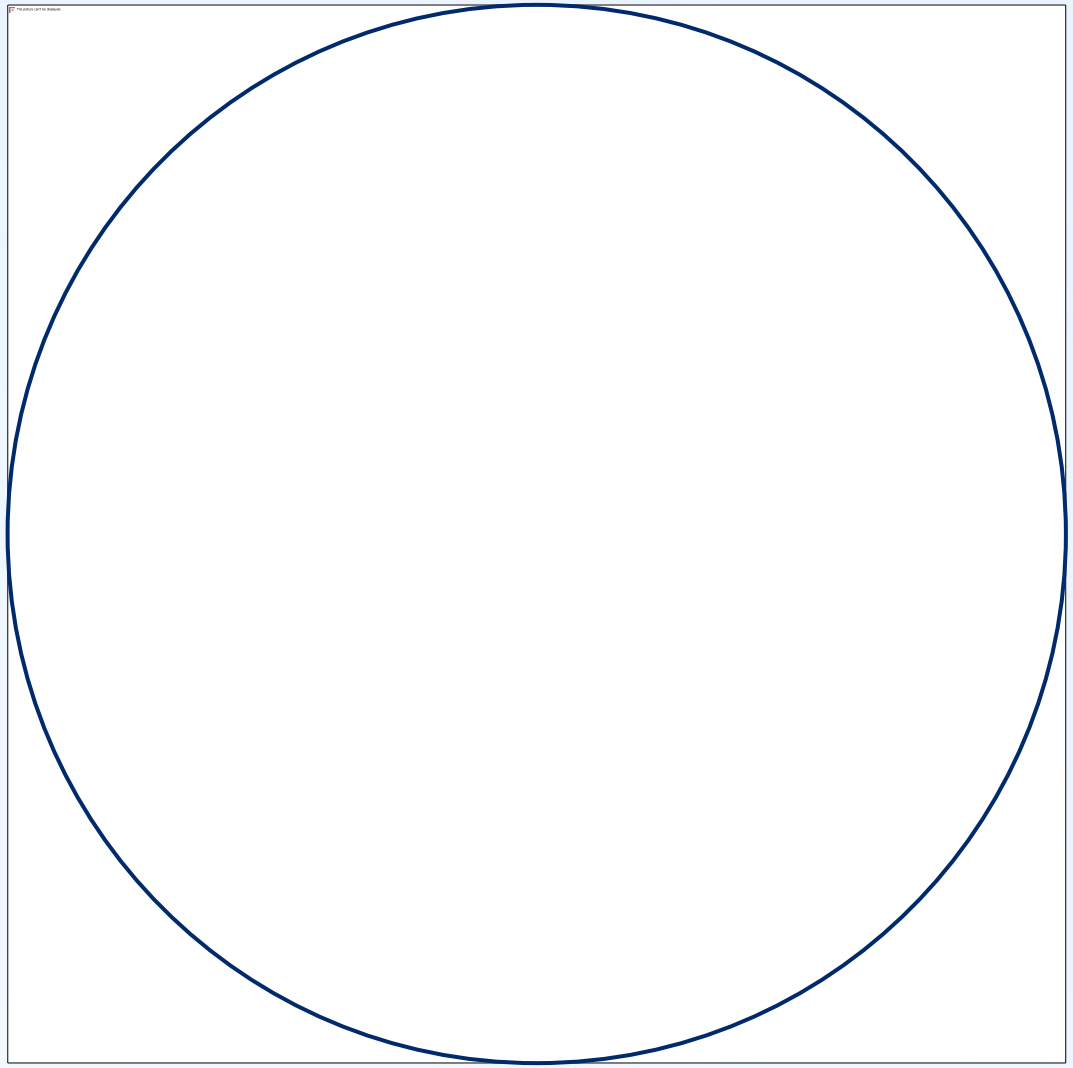
AI Audit: An LLM is used to find discrepancies between the draft SAPs and the source protocol.

Human-in-the-Loop Review: A human assesses if flagged deviations are hallucinations or valid suggestions permitted by the prompt.

Final Review: A human determines suggestion validity based on statistical rationale and the study context.

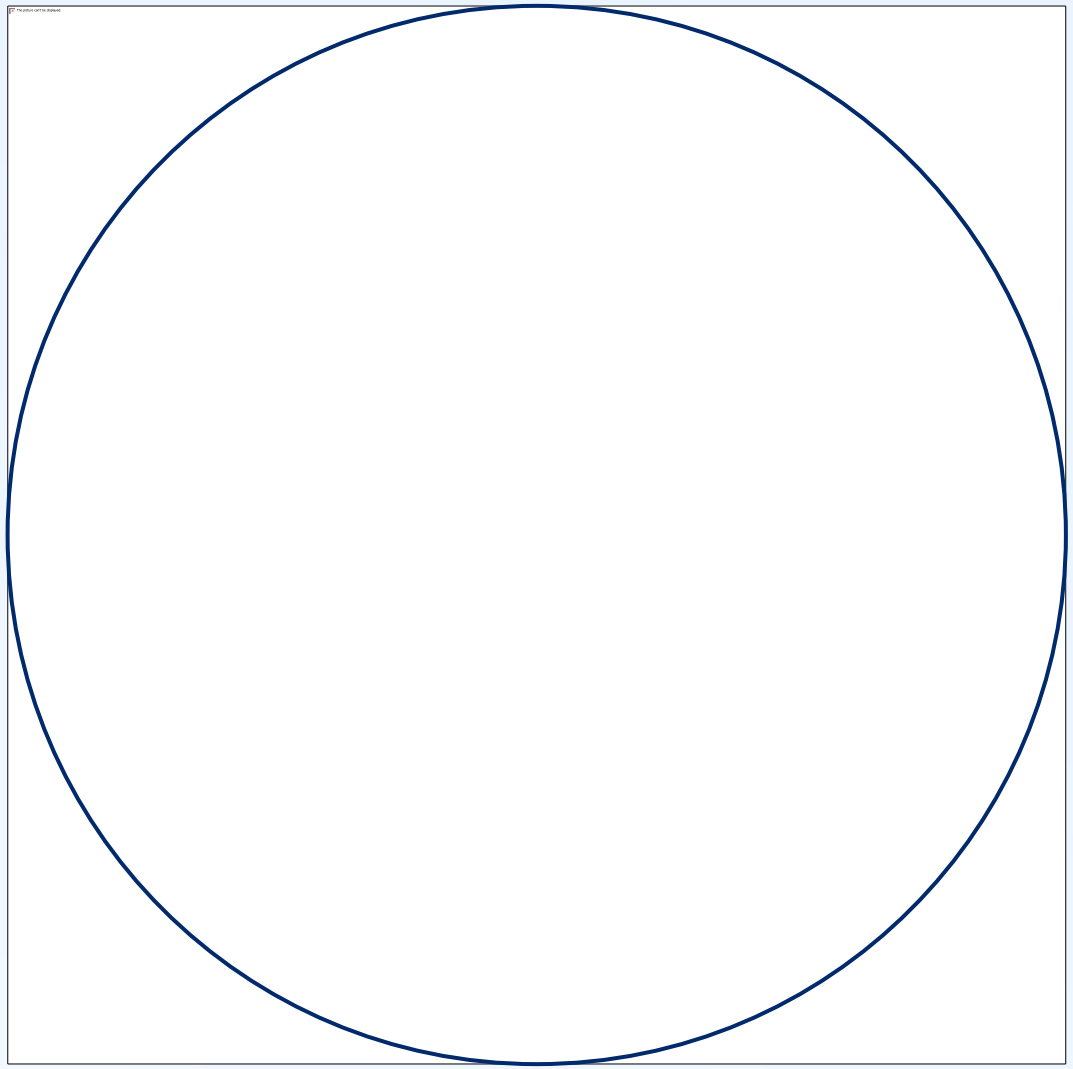
The Good

- **Identified Knowledge Gaps:** Correctly identified areas in the protocol requiring extra detail.
- **Accurate Expansion:** Generated content that closely matched the real SAP.
- **Above and beyond:** Recommended appropriate suggestions absent from the real SAP.
- **Special Mention:** Gemini produced shells and Claude summarised a figure, both unprompted.



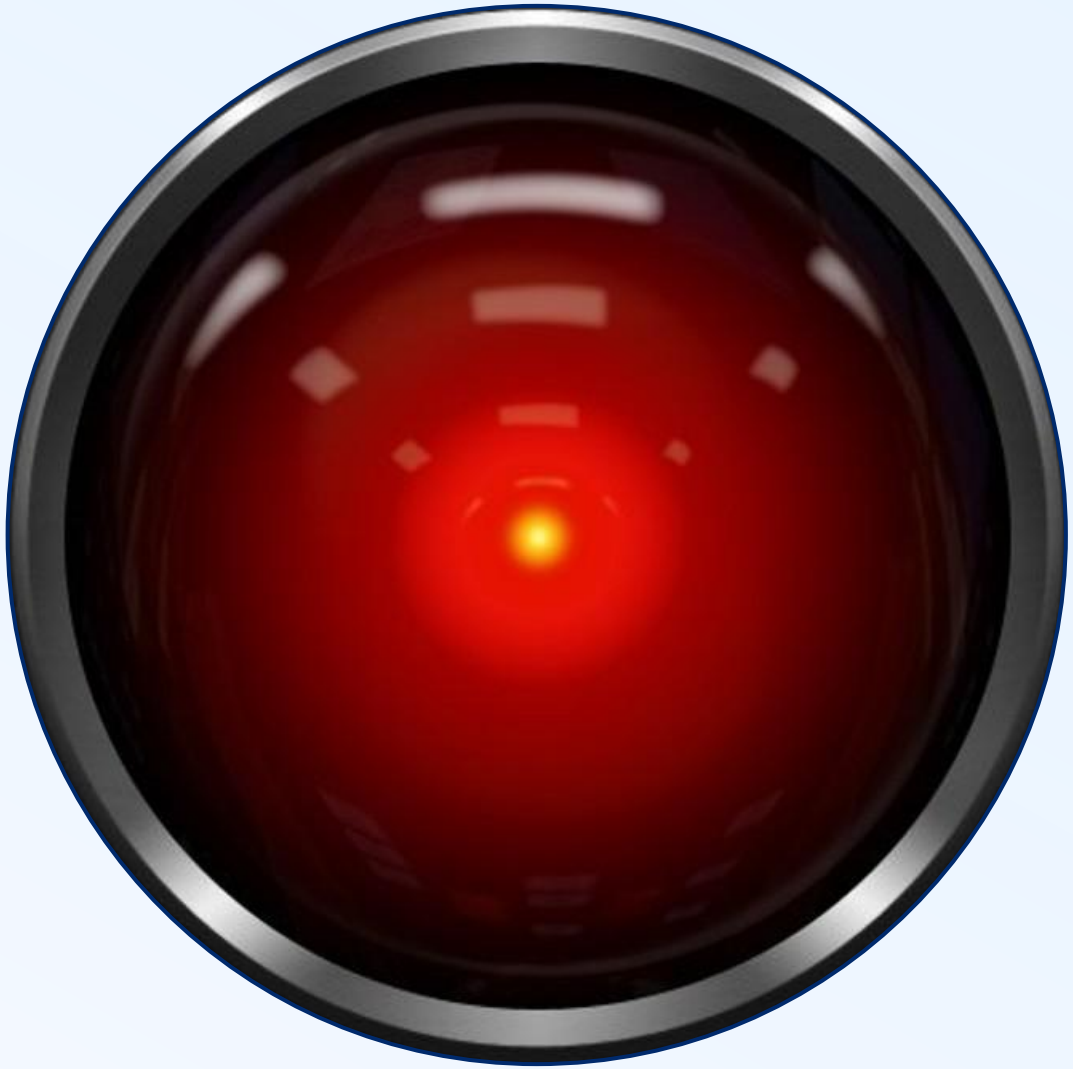
The Bad

- **Contextual Blind Spots:** Contradicted the protocol when information was outside of expected sections.
- **Limited Study Comprehension:** Proposed analysis that are inappropriate for the clinical context.
- **Incorrect Summary:** Generated a description of the disease that contradicted the protocol.



The Ugly

- **ChatGPT** suggested an imputation strategy for AE dates that would underreport TEAE.
- **Claude** suggested a nonsensical sensitivity analysis.
- **Gemini** recommended a flawed correction for heart rate variations.
- **Grok** dropped key covariates from the primary analysis.



How is this useful?

- **Laying Foundations:** You can get a decent "first draft" SAP from each LLM we investigated.
- **Cognitive Reallocation:** Statisticians can focus mental energy on the complex sections of a SAP.
- **Protocol Stress Test:** LLM misinterpretations could highlight areas of improvement in the protocol.
- **The Co-Pilot Mindset:** LLMs can be a valuable assistant when used properly.

ChatGPT (GPT-5)

Pros

- Can produce process and produce work documents
- Correctly identifies areas in the SAP where the information in the protocol requires more detail
    - Accurately expanded on the statistical hypothesis, matching the real SAP
    - Suggested an approach for robust SEs in the MMRM
    - Suggested appropriate summaries of PK data
    - Identified the need for categorical thresholds for compliance
    - Identified the need to impute dates for AE
    - Identified the need for Potential Clinically Significant Abnormalities

Cons

- Couldn’t find an estimand so said none existed
- Suggested categorical thresholds not appropriate in study context
- Inadequate imputation strategy for AE dates could underreport TEAE
- Did not suggest PCSAs for Clinical Laboratory Evaluations

Claude (Opus 4.6)

Pros

- Can produce process and produce work documents
- Correctly understood a figure showing power for different assumptions
- Correctly identifies areas in the SAP where the information in the protocol requires more detail
    - Accurately expanded on the statistical hypothesis, matching the real SAP
    - Suggested protocol deviation examples that were close to the real SAP
    - Conservative imputation approach for missing AE dates was reasonably close to the real SAP and both assumed AEs were treatment emergent
    - MNAR sensitivity analysis was very close to the real SAP
    - Suggested appropriate summaries of PK data
    - Identified thresholds for Potential Clinically Significant values

Cons

- Couldn’t find an estimand so said none existed
- Misunderstood prompt and created mapping for analysis visits that didn’t exist and contradicted the protocol
- Suggested a bad and nonsensical sensitivity analysis for the primary analysis

Gemini (3.1 Pro)

Pros

- Can process word documents & PDFs
- Placed the objectives and endpoints into a similar table to the protocol
- Suggested dummy TFLs
- Correctly identifies areas in the SAP where the information in the protocol requires more detail
    - Correctly defined the study design as a ‘double-dummy’
    - Suggested an approach for robust SEs in the MMRM
    - Added a baseline x study visit interaction to the MMRM
    - Suggested using a plot that regulatory bodies have recommended
    - Accurately expanded on the non-inferiority decision rule

Cons

- Couldn’t find an estimand so said none existed
- Dropped MMRM covariates that were included in the stratified randomisation
- Recommended a flawed correction for heart rate variations (Bazett’s correction)

Grok (Grok 4)

Pros

- Correctly identifies areas in the SAP where the information in the protocol requires more detail
    - Correctly defined the study design as a ‘double-dummy’
    - Correctly expands on blinding and randomisation protocol
    - Suggested a protocol deviation log and appropriate analysis
    - Suggests appropriate sensitivity analyses
    - Suggested appropriate discontinuation reason

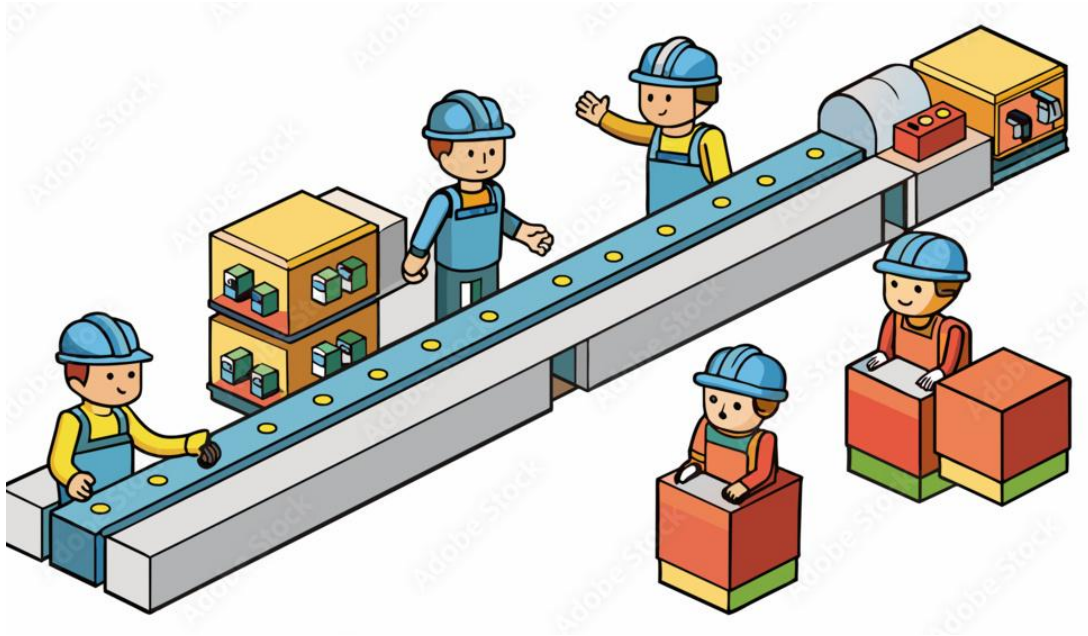
Cons

- Can’t process or produce word documents
- Couldn’t find an estimand so said none existed
- The protocol said chronic hyperglycemia (CH) is a key factor of microvascular complications and to a lesser extent, macrovascular complications. The SAP said CH "remains the primary driver of microvascular and macrovascular complications."
- Couldn’t find the concomitant medication coding dictionary source so used one that contradicted the protocol

FUTURE POSSIBILITIES

AI AGENTS

- It may be possible soon to use AI agents to deliver each new prompt for us once a section has been completed to we don’t have to do that ourselves.
- This may cause issues if the model has specific questions around a prompt for a specific section.
- For the gain in speed here, it might be worth the downside of losing some decision making in the process

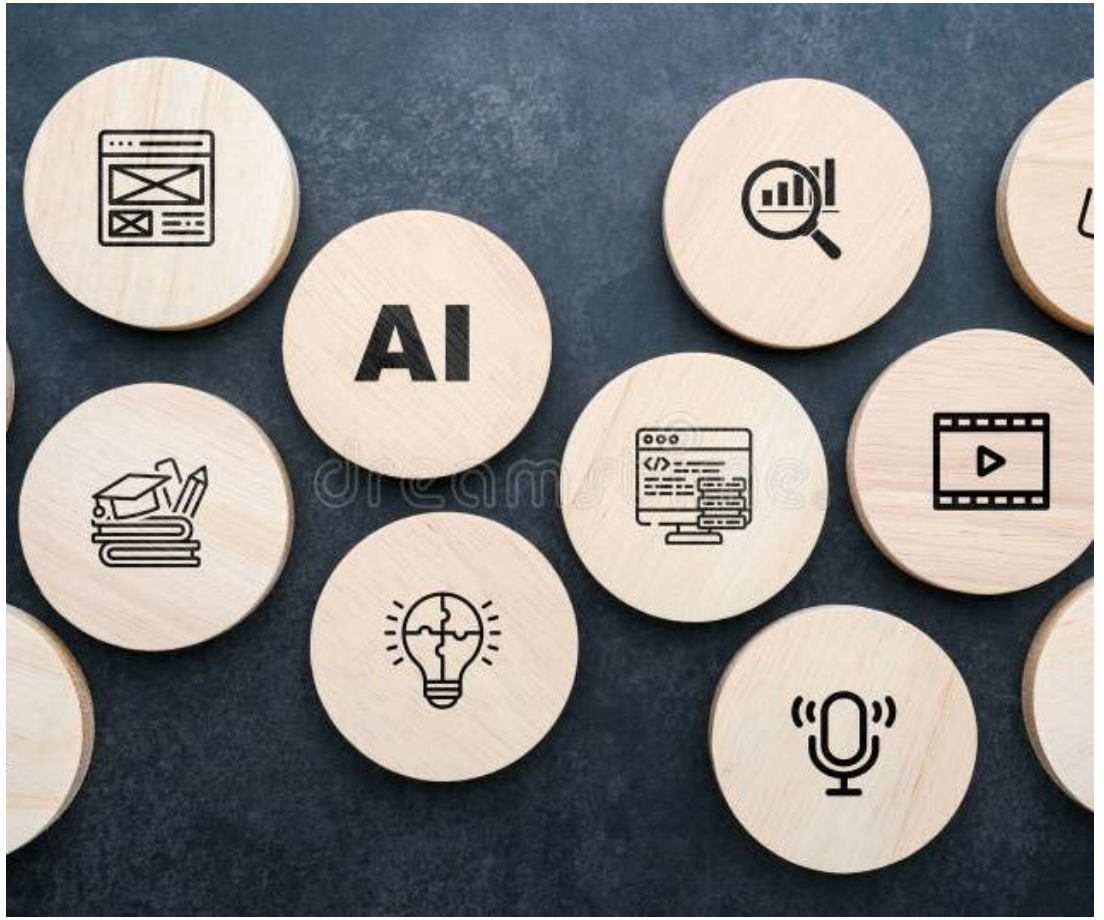


MORE CAPABLE LLMS

- LLMs are getting more capable and more powerful
- How will this affect our process, can we get away with less prompts?
- Drafts will gradually get better and better, though human oversight will probably be necessary for sometime.

DIFFERENT USE CASES

- Protocols – Potentially quite a difficult approach as we don't have other study documents to work from.
- Shells – May work drawing from shell template and SAP.
- Specs – May work drawing from CDISC documentation and SAP
- CSRs – I'm aware this is already being done, this makes sense as you have a swathe of documents from throughout the study to draw data from as well as the template.



JOB SECURITY

- In our opinion this is not replacing us any time soon.
- It will make us more efficient however, I suspect in future we will be expected to work more quickly on some things with the help of AI.
- It’s much harder to pin accountability if relying on AI and I think this will continue to be key within clinical research and therefore it cannot replace skilled staff.

ANY QUESTIONS?

REFERENCES

- Leveraging generative AI to transform statistical analysis plan authoring in clinical trials – PubMed
- Shutterstock image - 43 Ai Superpower Royalty-Free Images, Stock Photos & Pictures | Shutterstock
- Alamy images - Get creative with stock photos and videos from Alamy
- WALL-E image
- Ex Machina image
- 2001: A Space Odessey image
- R2-D2 image
- LucotIC's Cam Newton Template (12#) - Imgflip

PSI 2026 CONFERENCE

15th - 17th June

ICC BELFAST

Northern Ireland

visit Belfast

TOURISM

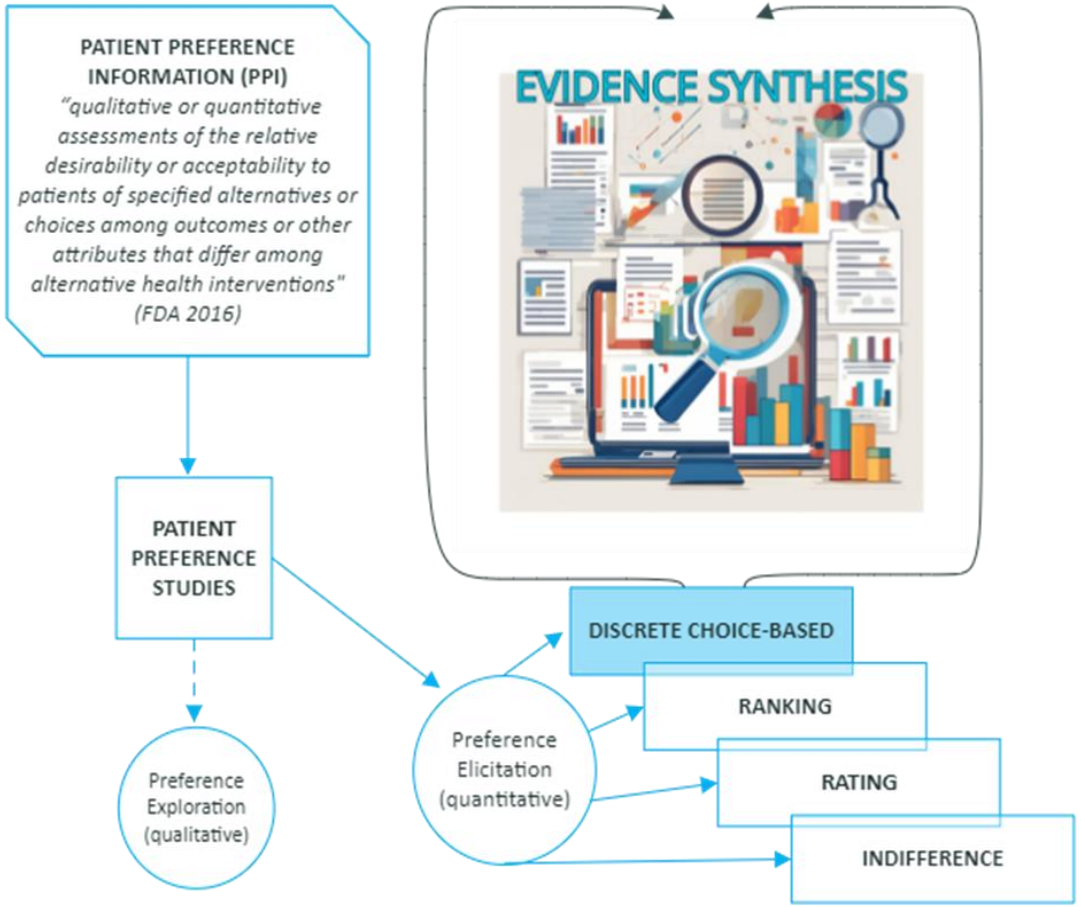
Repertor.IO: Patient Preference Research with AI-Powered Evidence Generation

E. Pietropaolo[1], M. Bracco[2], A.a Visconti[2], Ileana Baldi[3], Paola Berchialla[2]

[1] Repertor.IO s.r.l, Italy; [2] University of Torino, Italy; [3]University of Padova, Italy

A community call

- HPR community sees an urgent need for more methodological research on synthesis and transferability of PPI
- Transfers on the basis of meta-analyses combine effect sizes from multiple studies while adjusting for characteristics that are relevant to the new decision context(s)



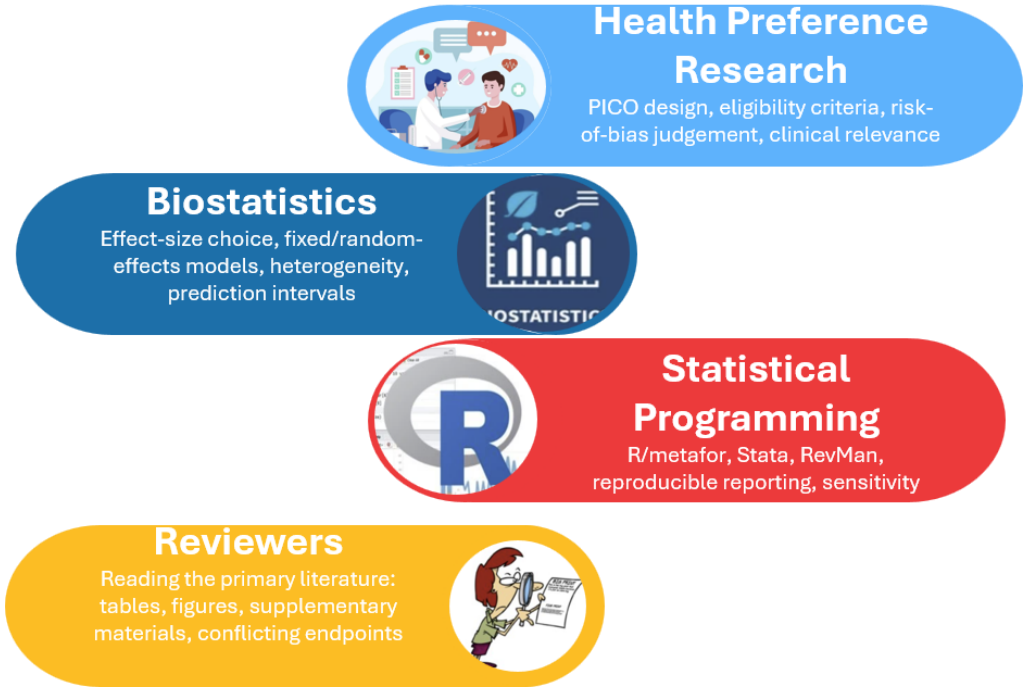
Systematic Literature Review

Do We Always Need a New Preference Study? A Scoping Review of Promising Research Areas for Meta-Analyses and Benefit Transfers of Patient Preference Studies

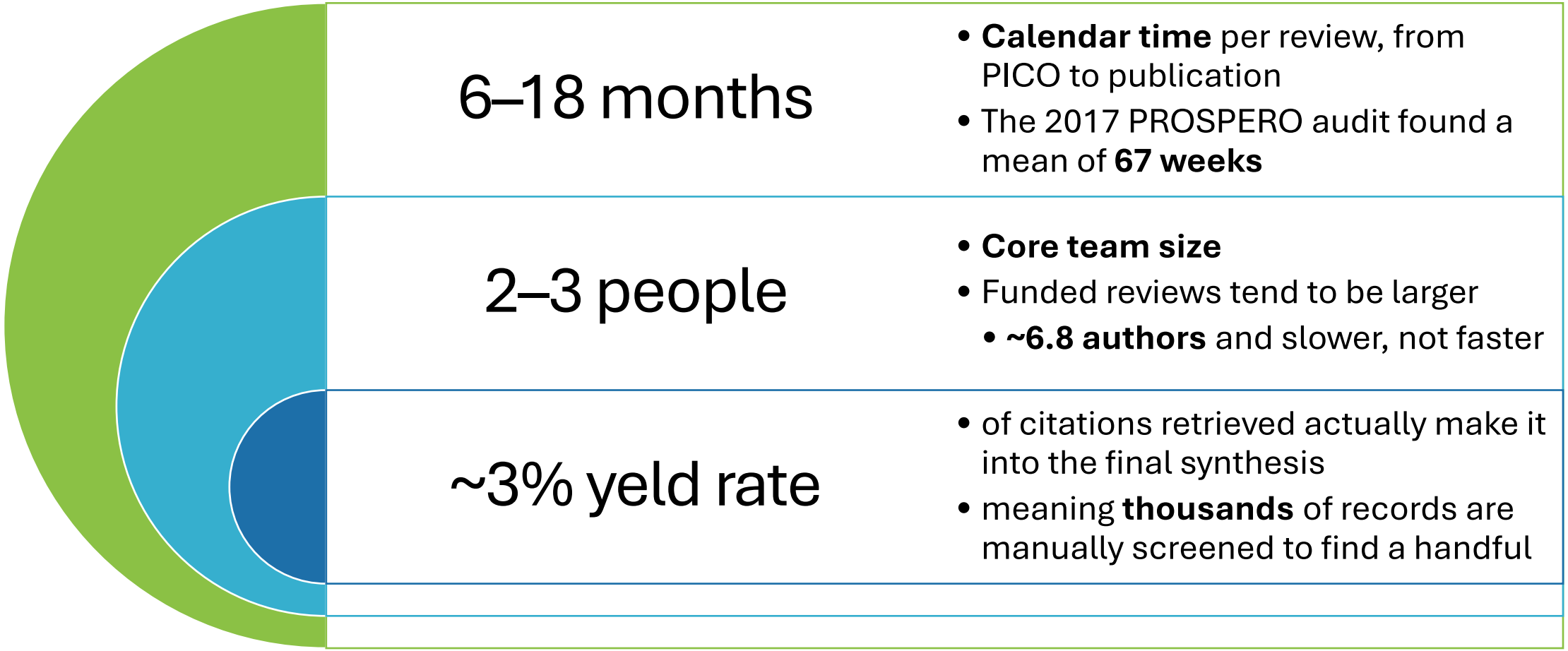
Michael Bui, MSc, Catharina G.M. Groothuis-Oudshoorn, PhD, A. Cecilia Jimenez-Moreno, PhD, Byron Jones, PhD, Conny Berlin, Dipl-Math, Janine A. van Til, PhD

Know-how: the rare integration of three disciplines

- Conducting a meta-analysis is a *multidisciplinary* exercise: the integrated expertise is rare and errors are common where it is missing

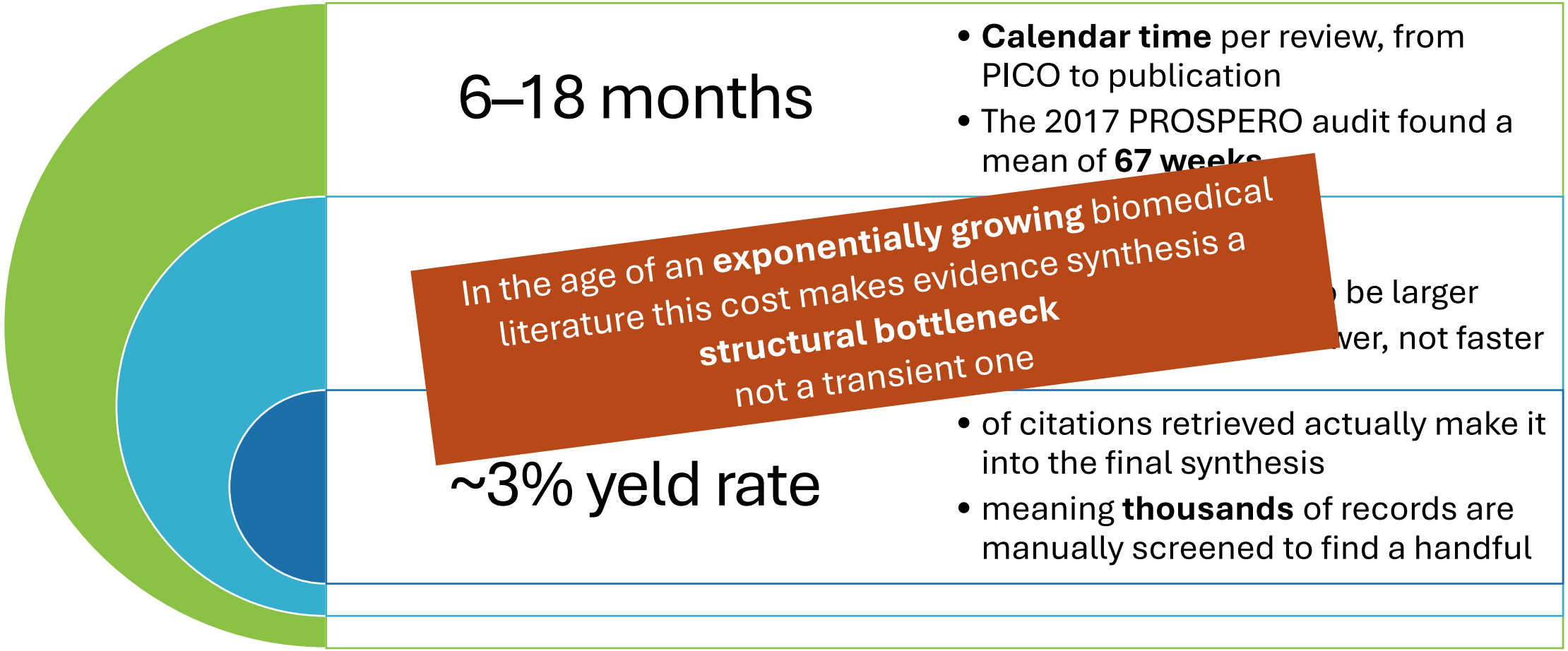


Time and people: meta-analyses are expensive to produce



Borah, Brown, Capers & Kaiser (2017) *BMJ Open*
Allen & Olkin (1999) *JAMA*

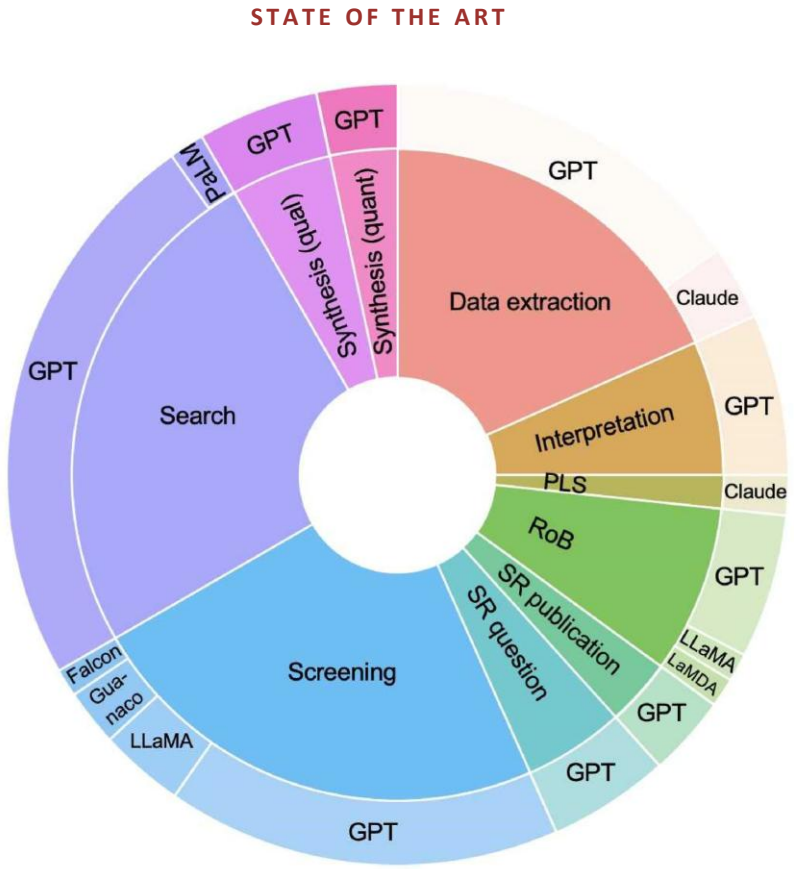
Time and people: meta-analyses are expensive to produce



Borah, Brown, Capers & Kaiser (2017) *BMJ Open*
Allen & Olkin (1999) *JAMA*

LLM for each Systematic-Review step

- Pie-donut chart depicting proportions of systematic review (SR) steps (inner layer pie) and associated large language model (LLM) applications (outer layer donut)
- The inner layer represents the frequencies and proportions of all individual SR steps (n = 60) extracted across the 37 studies included (multiple counts per study were possible); the outer layer provides a breakdown of the percentage distribution of LLM types used for each SR step:
  - Search
  - Screening
  - Data extraction
  - Risk of bias assessment (RoB)
  - Interpretation
  - SR question
  - Synthesis (qualitative)
  - Synthesis (quantitative)
  - SR publication
  - Plain language summary (PLS)



Lieberum et al. 2025 Journal of Clinical Epidemiology

Standardize steps

PRISMA & Cochrane guidelines define a rigid, sequential workflow. Reproducibility is a feature; the cost is that every step is a manual lift

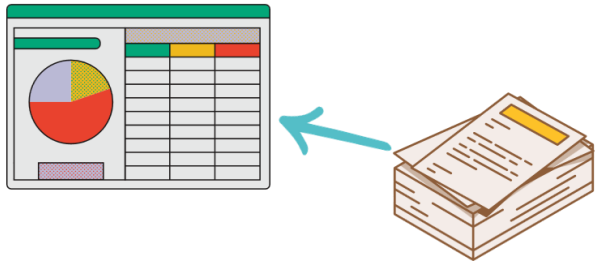
Research question → **Literature Search** →

"What are patients' preferences regarding disease X?"

Scopus
Web of Science™
PubMed

Manually reading thousands of titles/abstracts

← **Manual Data Extraction** ←



← **Statistical Synthesis**

Risk of Bias
Meta-analysis
Reporting

Comprehensive literature search

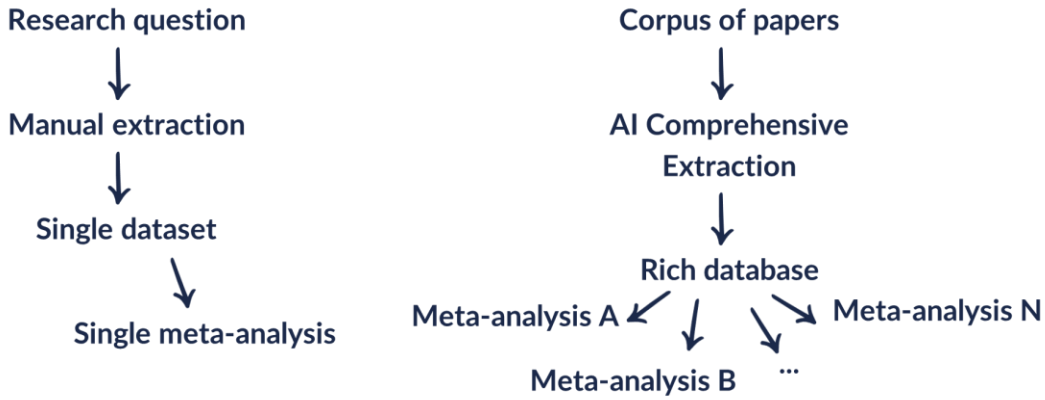
Scopus®

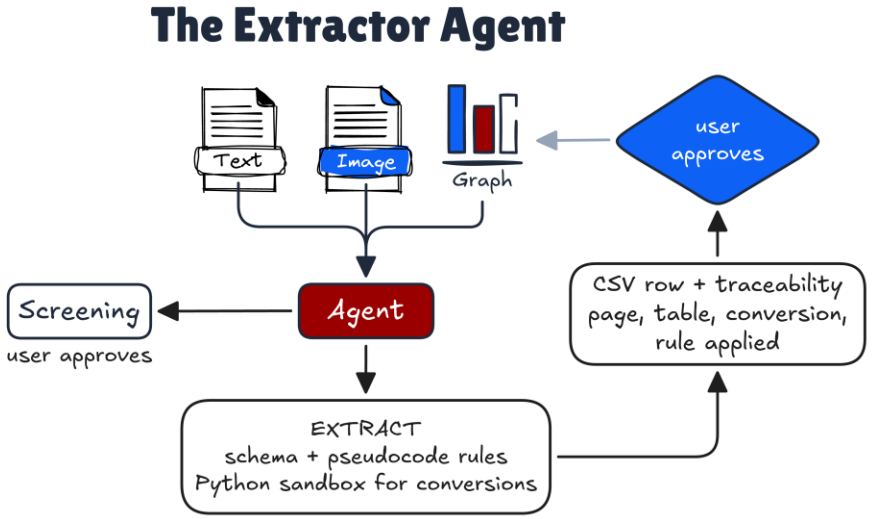
Web of Science™

PubMed

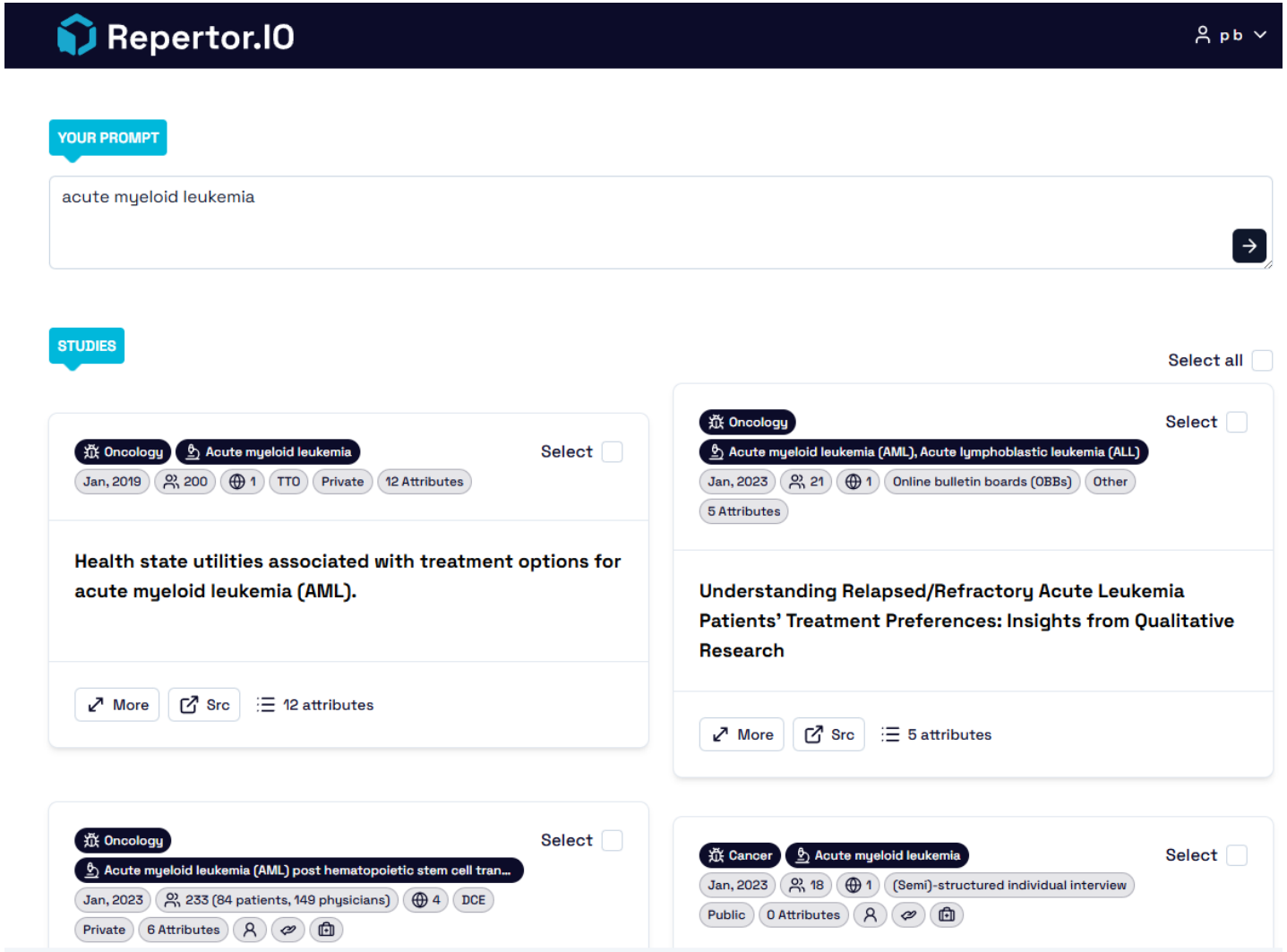
- Patient preference studies are not easy to retrieve
- We use a very broad research string to collect as many PPS as possible, even if this means collecting a lot of irrelevant data
- This string is applied to major literature databases

AI powered platform

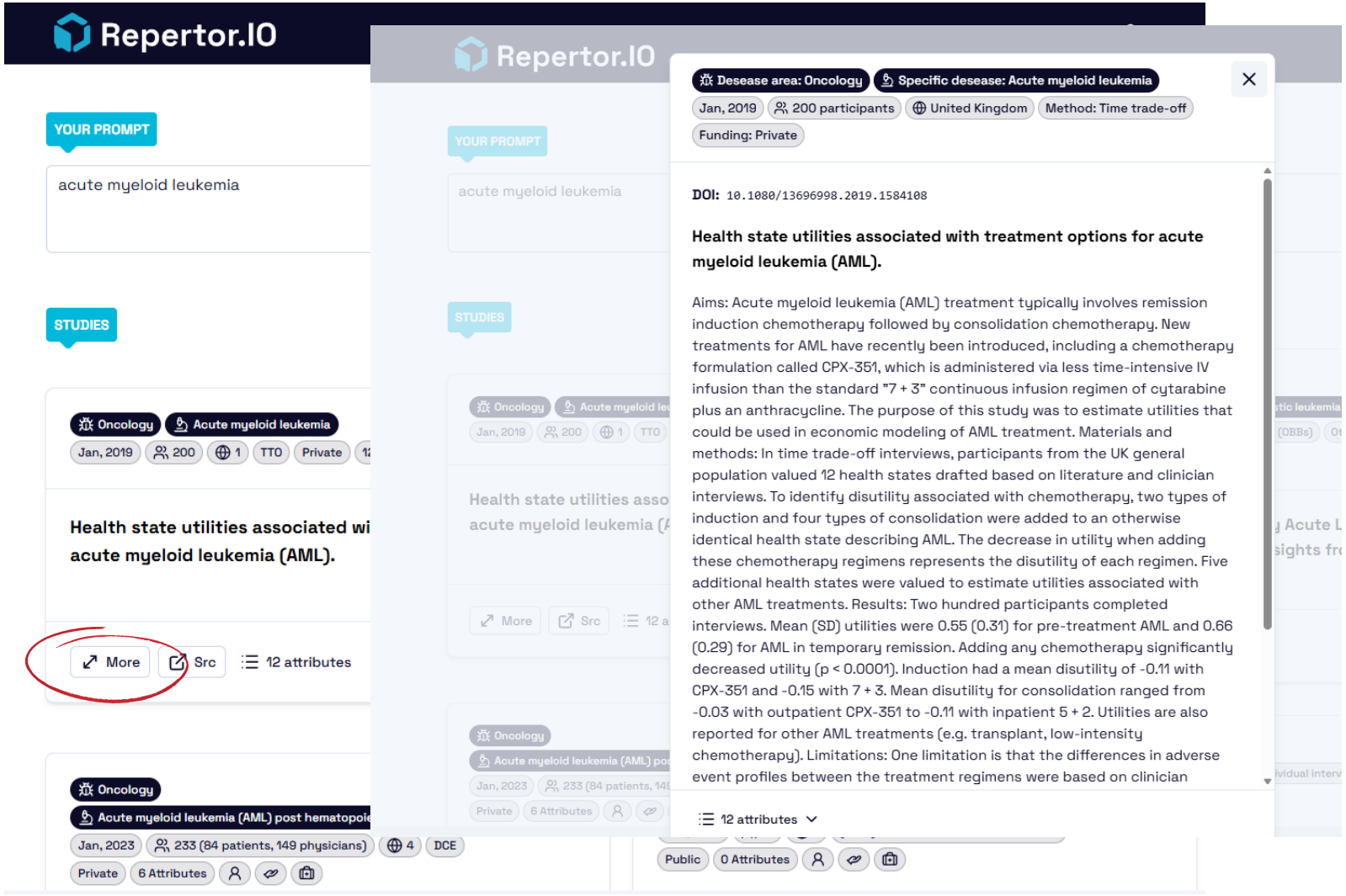




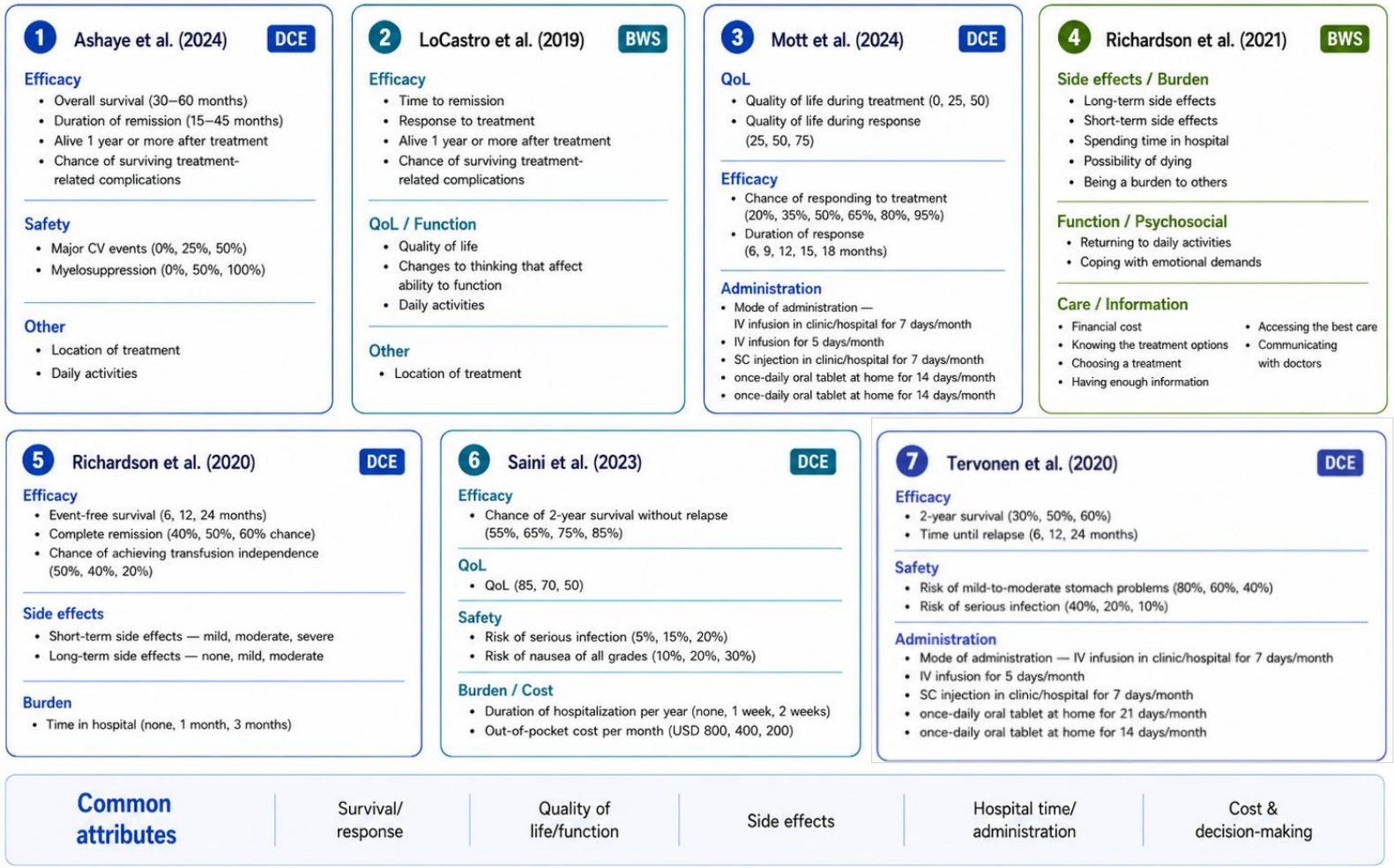
Searching Patient Preference Studies



Searching Patient Preference Studies



Summary of data extraction



Tervonen et al data extraction

Patient Preferences for Maintenance Treatment of Acute Myeloid Leukemia: Results of a Discrete Choice Experiment

Introduction: Standard treatments for older patients (pts) with acute myeloid leukemia (AML) can induce remission; however, responses are often short-lived and survival rates are poor upon relapse (Chen Y, et al. Medicine 2016;95:e4182). To date, no AML maintenance therapies have been approved by the US FDA, but a few have been used 'off-label'. Although some of these maintenance therapies, such as injectable azacitidine (AZA), improved disease-free survival in older pts with AML, overall survival (OS) benefits have been more difficult to achieve. Moreover, injectable therapies are associated with a higher administration burden and infusion reactions, and therefore may not be suitable for long-term maintenance therapy. The randomized, phase 3 QUAZAR AML-001 study (NCT01757535) of CC-486, a novel oral formulation of AZA, was the first maintenance study to demonstrate a significant and clinically meaningful improvement in OS (Wei AH, et al. Blood 2019;134:LBA-3). CC-486 was associated with a 9.9-month increase in OS (vs placebo) with a manageable safety profile and no compromise in health-related quality of life in pts with AML aged ≥ 55 years previously treated with intensive chemotherapy. Although previous research has examined patient preference for induction therapies, very little is known about preferences for AML maintenance therapies. We therefore used an online discrete choice experiment (DCE) survey to determine the relative importance that pts with AML place on key clinical benefits and risks, mode of administration, and out-of-pocket (OOP) costs. Methods: From November 2019 to April 2020, pts aged ≥ 55 years from the USA, Canada, Germany, and Italy who had undergone treatment for AML were invited to participate in an online DCE survey. Attributes in the DCE survey included clinical benefits (time until relapse [6, 12, 24 months] and 2-year survival rate [30%, 50%, 60%]), adverse events (risk of mild-to-moderate stomach problems [40%, 60%, 80%] and risk of serious infection [10%, 20%, 40%]), mode of administration (once-daily oral tablet for 14 or 21 consecutive days/month, subcutaneous [SC] injection in a clinic/hospital for 7 consecutive days/month, or intravenous [IV] infusion in a clinic/hospital for 5 or 7 consecutive days/month), and OOP costs (USD 200, 400, 800). Patient preferences for attribute levels were analyzed using a multinomial logit model and expressed as marginal utilities and maximum acceptable decrease in 2-year survival rate. Results: In total, 170 pts completed DCE surveys (USA, n = 104; Canada, n = 6; Germany, n = 30; and Italy, n = 30). Mean age was 63.0 years, and 54% of pts were male. In all, 73% of pts had achieved remission at any time, 74% had not received prior stem cell transplant, and 79% had been diagnosed with AML in the last 6 months. Based on the DCE survey results, pts valued a 30% increase in the chance of 2-year survival (marginal utility for 60% = 0.84; 95% confidence interval [CI] 0.70-0.99) more than changes in any other attribute. This was followed by a USD 600 decrease in OOP costs (marginal utility for USD 200 = 0.77; 95% CI 0.66-0.89), 18-month increase in time until relapse (marginal utility for 24 months = 0.61, 95% CI 0.48-0.74), an oral tablet for 14 days/month instead of IV infusion 7 days/month (marginal utility = 0.38, 95% CI 0.23-0.54), and a 30% decrease in the risk of serious infection due to injection (marginal utility for 10% = 0.20; 95% CI 0.09-0.32). Risk of mild-to-moderate stomach problems was not important to pts (Figure). Pts preferred an oral tablet taken either 14 or 21 days/month over SC injection 7 days/month (both P = 0.002). In addition, pts were willing to accept a 16% and 14% decrease in the chance of 2-year survival to switch from IV infusion in a clinic/hospital for 7 consecutive days/month to an oral tablet for 14 days/month (95% CI 9.95-22.08) or 21 days/month (95% CI 7.44-20.50), respectively. Conclusions: In this survey of pts with AML, the most important attribute during maintenance therapy was the probability of survival at 2 years. In addition, pts demonstrated a significant preference toward an oral tablet over IV infusions and SC injections in a clinic/hospital, and were willing to accept a significant decrease in treatment efficacy in favor of an oral mode of administration. This study provides valuable insights into patient preferences and may help inform decision-making for AML maintenance therapies. [Formula presented] Disclosures: Tervonen: Evidera: Current Employment, Current equity holder in publicly-traded company. Seo: Bristol Myers Squibb: Research Funding; Evidera: Current Employment. Nehme: Bristol Myers Squibb: Current Employment, Current equity holder in publicly-traded company. La Torre: Bristol Myers Squibb: Current Employment. Prawitz: Bristol Myers Squibb: Research Funding; Evidera: Current Employment. Chen: Bristol Myers Squibb: Current Employment, Current equity holder in publicly-traded company. Beach: Bristol Myers Squibb: Current Employment. Wang: Bristol Myers Squibb: Current Employment, Current equity holder in publicly-traded company.

▼ **Classification**

Is PPS	Yes
Confidence	99.39%

Level 1 data extraction

Level 1 Summary

Disease Area	Oncology
Disease	Acute myeloid leukemia (AML)
Design	Patient Preference Study (PPS)
Stage	Full scale
Study Type	Quantitative
Elicitation Method	Discrete choice experiment (DCE)
Exploration Method	Non applicable
Statistical Model	Multinomial logit
Participants	170
Countries	USA, Canada, Germany, Italy
Population	Patients
Recruitment	Other
Funding	Private
Attributes	Time until relapse, 2-year survival rate, risk of mild-to-moderate stomach problems, risk of serious infection, mode of administration, out-of-pocket costs
# Attributes	6
Attribute Types	Effectiveness of intervention, Cost of intervention, Safety of intervention, Mode of intervention
Alternatives	—
# Alternatives	—

Level 2 data extraction

multinomial logit DCE main results

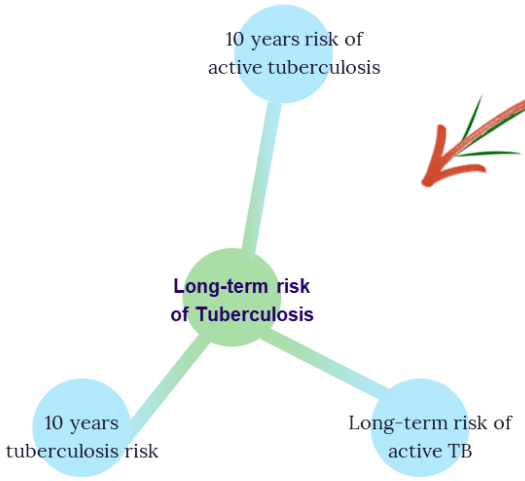
Main discrete choice experiment analysis using a multinomial logit model reporting marginal utilities (preference weights) for attribute levels, significance comparisons between modes of administration, and trade-offs expressed as maximum acceptable decreases in 2-year survival.

ATTRIBUTE	CLASS	LEVEL	WEIGHT	SE	P-VALUE	CI LOW	CI HIGH
2-year survival							
	Class 1						
		30% (reference)	0	—	—	—	—
		50%	0.65	—	—	0.55	0.75
		60%	0.8	—	—	0.68	0.92
Time until relapse							
	Class 1						
		6 months (reference)	0	—	—	—	—
		12 months	0.25	—	—	0.15	0.35
		24 months	0.55	—	—	0.42	0.68
Risk of mild-to-moderate stomach problems							
	Class 1						
		80% (reference)	0	—	—	—	—
		60%	0.05	—	—	-0.03	0.13
		40%	0.1	—	—	0.03	0.17
Risk of serious infection							
	Class 1						
		40% (reference)	0	—	—	—	—
		20%	0.18	—	—	0.1	0.26
		10%	0.22	—	—	0.14	0.3
Mode of administration							
	Class 1						
		IV infusion in a clinic/hospital for 7 consecutive days/month (reference)	0	—	—	—	—
		IV infusion in a clinic/hospital for 5 consecutive days/month	0.04	—	—	-0.04	0.12
		SC injection in a clinic/hospital for 7 consecutive days/month	0.06	—	—	-0.02	0.14
		Once-daily oral tablet at home for 21 consecutive days/month	0.58	—	—	0.45	0.71
		Once-daily oral tablet at home for 14 consecutive days/month	0.5	—	—	0.38	0.62
Out-of-pocket cost per month							
	Class 1						
		USD 800 (reference)	0	—	—	—	—
		USD 400	0.45	—	—	0.33	0.57
		USD 200	0.75	—	—	0.6	0.9

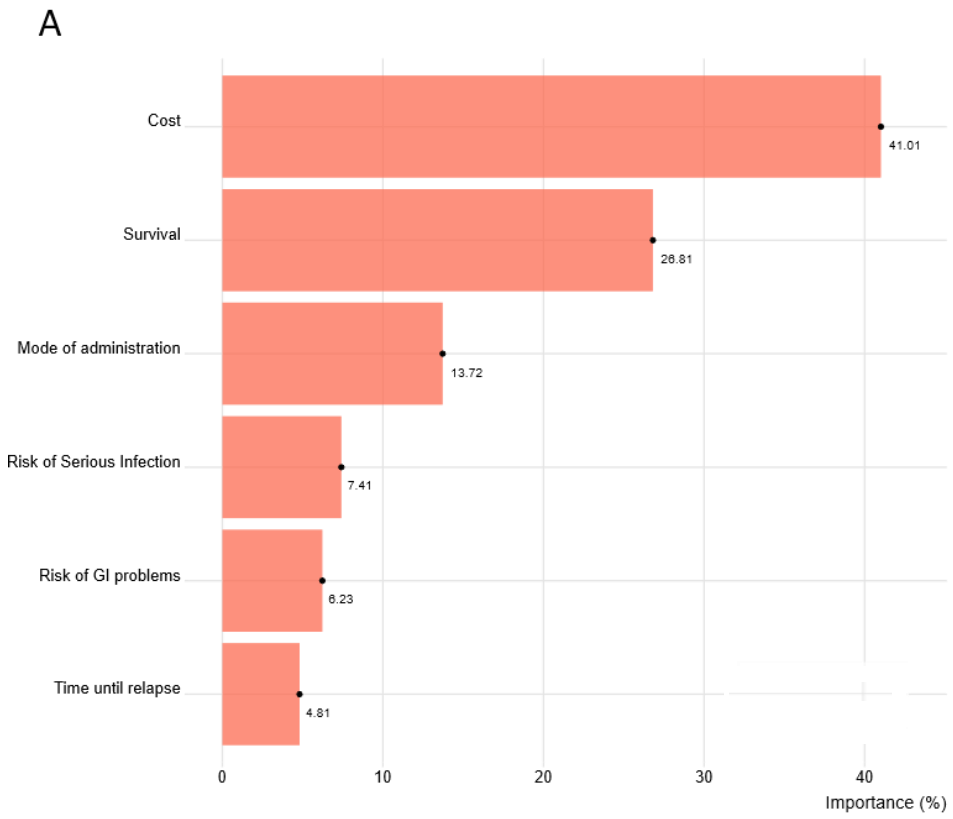
Searchable library of DCE's attributes

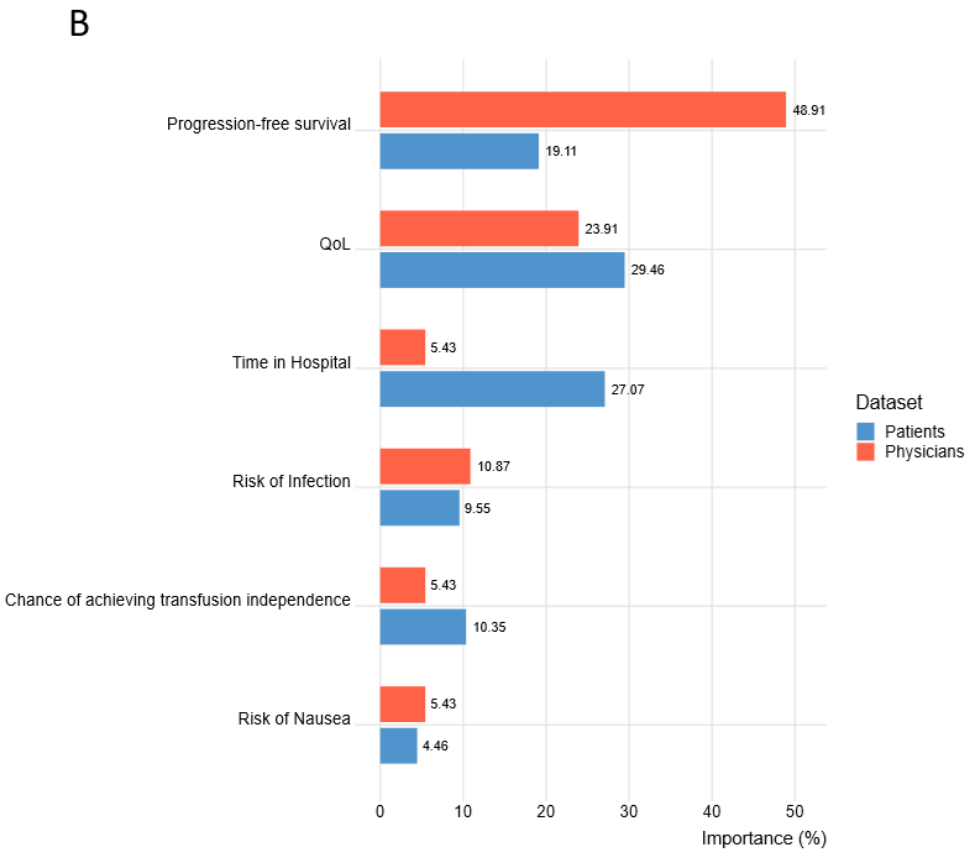
DCEs Attribute and Level Ontology

- Searchable Library
  - Facilitate new studies designs
- Harmonization for Meta-analysis



Attributes ranking



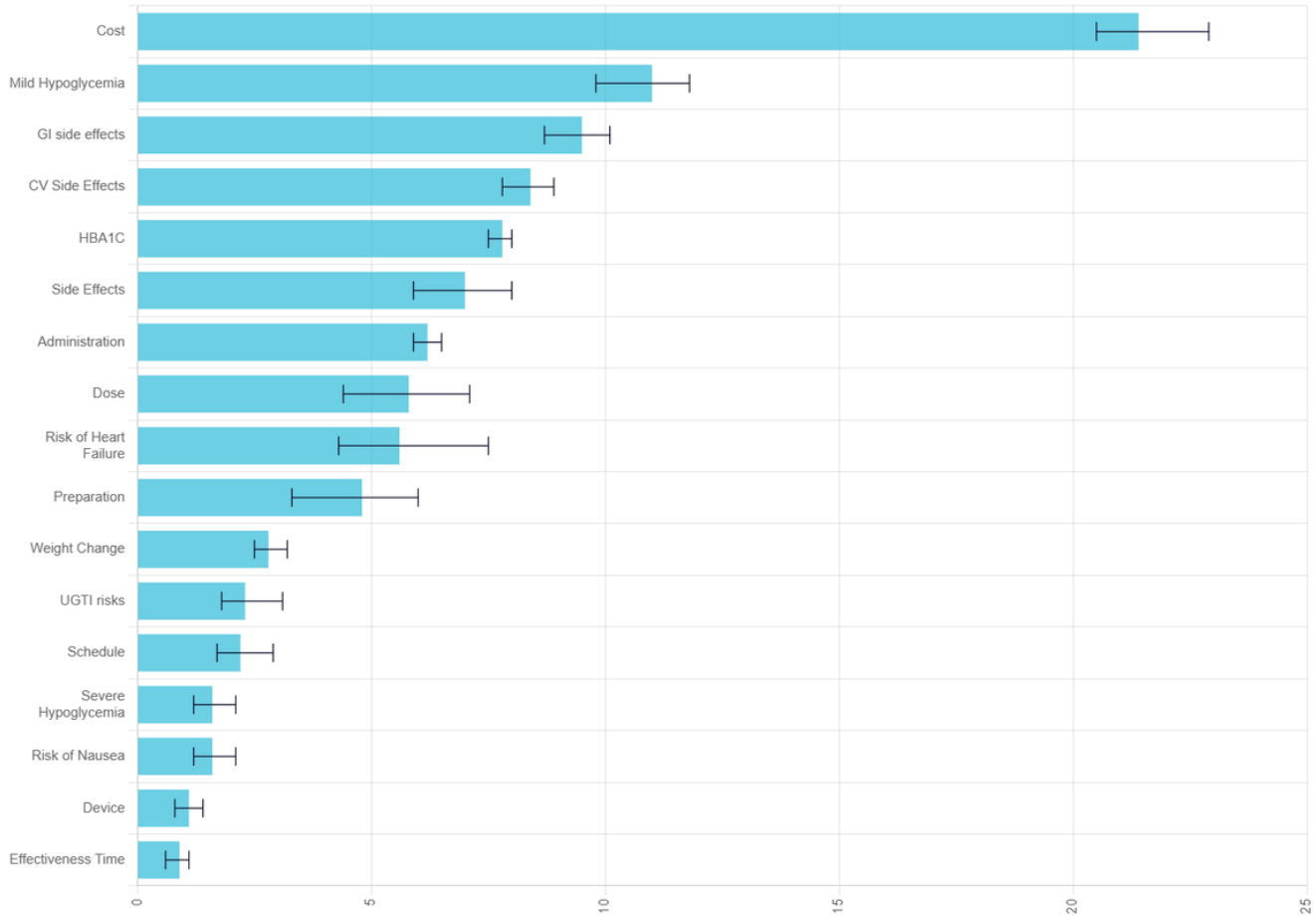


UNIVERSITÀ DI TORINO
Department
Clinical and
Biological Sciences

RESULTS

Ranking attributes appearing in at least − 1 + papers

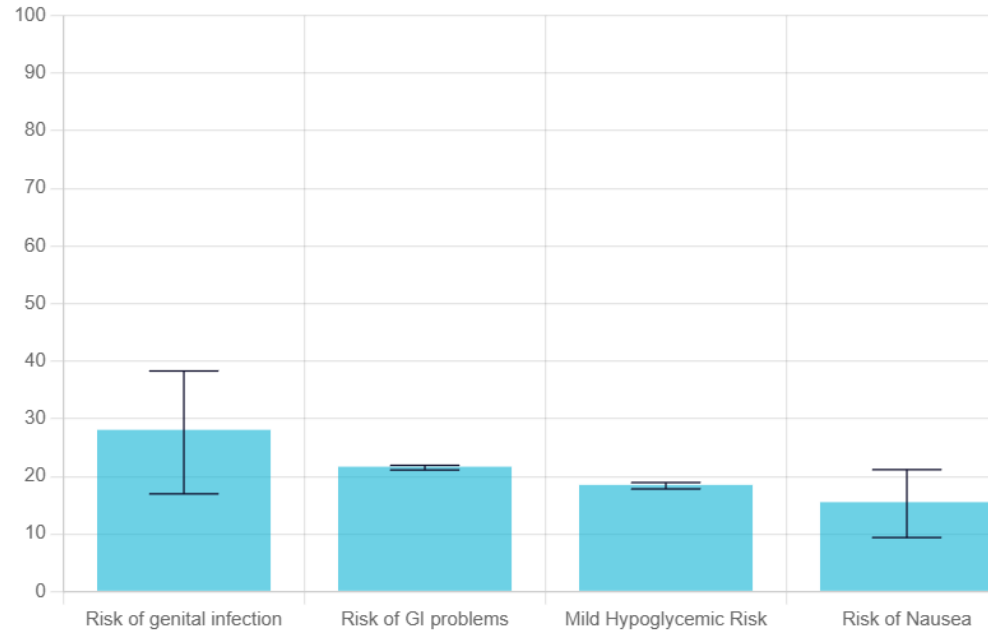
21.4 Cost | 11.0 Mild Hypoglycemia | 9.5 GI side effects | 8.4 CV Side Effects



RESULTS

MARs for HBA1C Reduction (1 %)

28.1 Risk of genital infection | 21.7 Risk of GI problems | 18.5 Mild Hypoglycemic Risk | 15.5 Risk of Na



RESULTS

MARs for

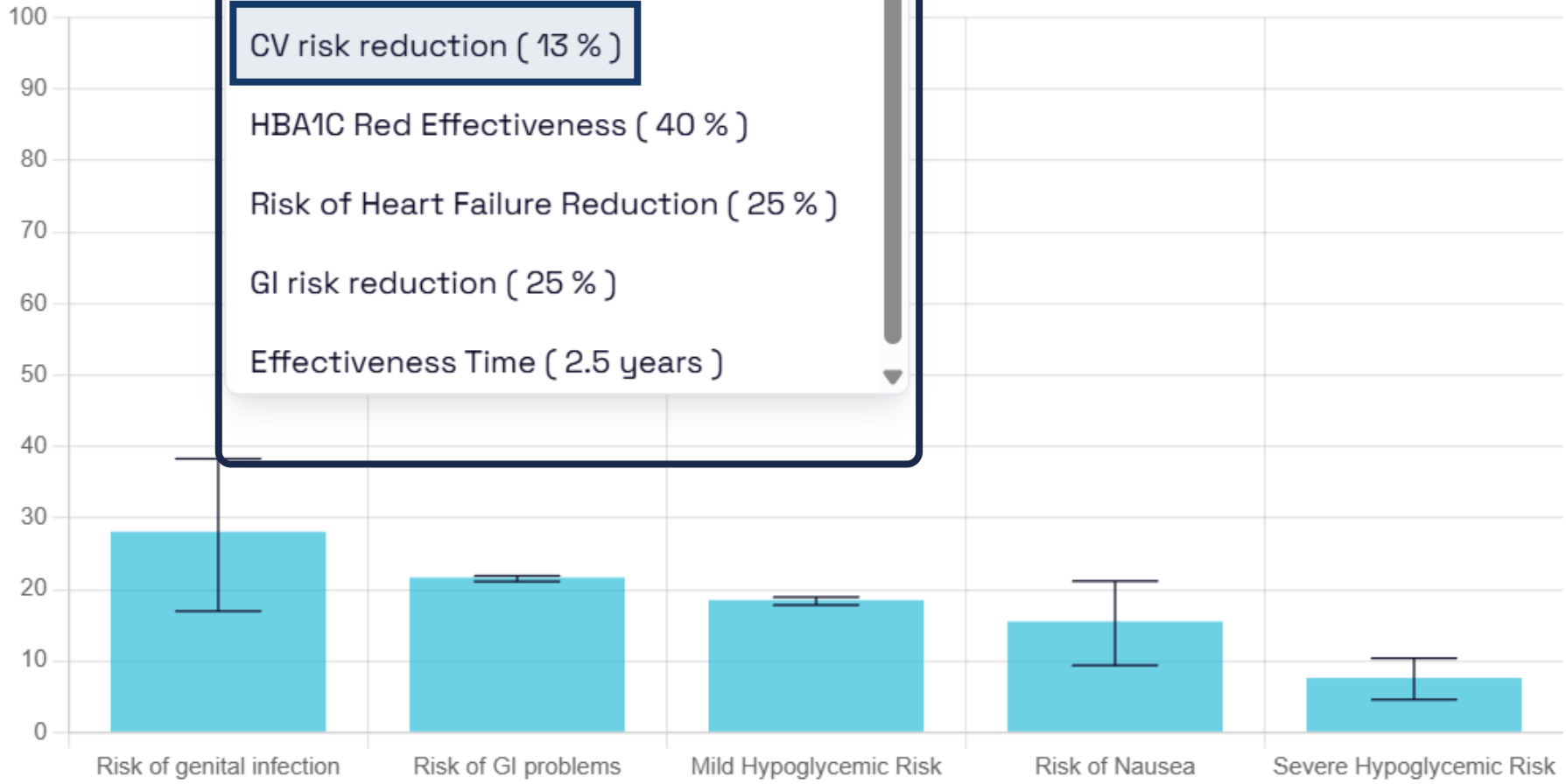
HBA1C Reduction (1 %)

28.1 Risk of genita

emic Risk

15.5 Risk of Nausea

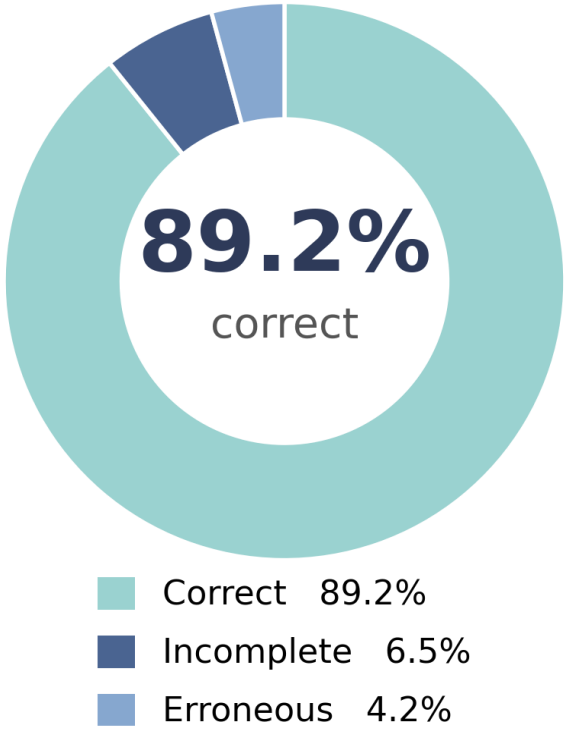
- HBA1C Reduction (1 %) ✓
- CV risk reduction (13 %)
- HBA1C Red Effectiveness (40 %)
- Risk of Heart Failure Reduction (25 %)
- GI risk reduction (25 %)
- Effectiveness Time (2.5 years)



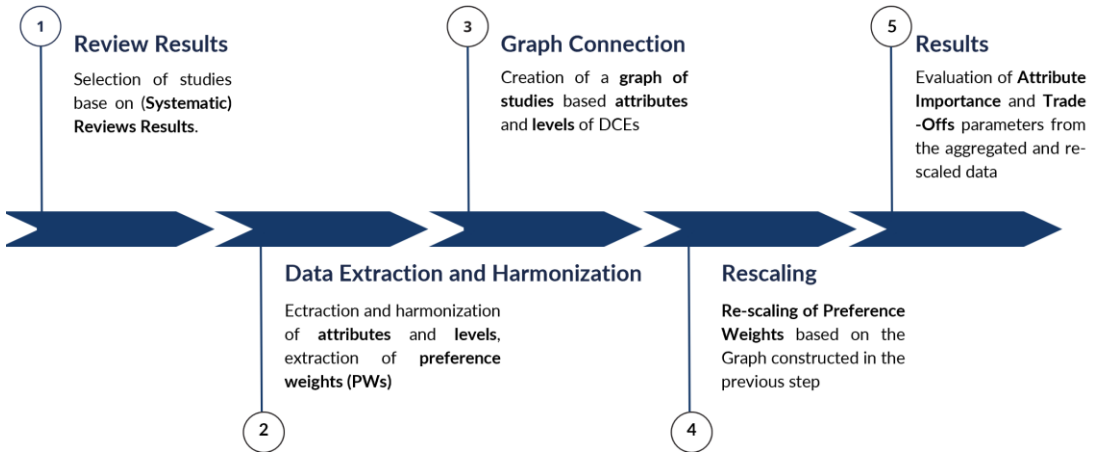
Validation: how reliable, and where it fails

C	**Correct** matches the source
H	**Hallucinated** no basis in the source (fabricated)
I	**Incomplete** partially correct; detail missing
M	**Missing** should have been extracted, but omitted
E	**Erroneous** extracted, but wrong value

Each extracted item scored against its original source. Reported separately by extraction level (classification, Level 1, Level 2), because reliability falls with depth.



AI-enabled meta-analysis is not simply a faster traditional meta-analysis



- It changes the workflow from a one-off manual process into a structured, reusable evidence infrastructure
- In HPR this is relevant because preference evidence requires interpretation at several levels:
  - extraction of study characteristics and preference estimates
  - harmonisation of attributes and levels across studies
  - rescaling of preference weights
  - model specification across heterogeneous designs
  - validation of extracted and synthesised results

Human-in-the-loop review remains central

- AI support was limited when tasks required contextual or methodological interpretation, for example:
  - deciding whether two attributes are conceptually equivalent
  - interpreting ambiguous outcome definitions
  - checking whether extracted values correspond to the correct model or subgroup
  - resolving inconsistencies between text, tables, figures, and supplementary materials
  - judging whether a harmonised synthesis remains clinically and methodologically meaningful
- AI platform should be understood as a hybrid system: automated where possible, expert-supervised where necessary

AI makes preference evidence synthesis more scalable *but only if it is auditable*

- The value of AI in HPR is not only efficiency
- Its real promise is to support a more transparent, reproducible, and cumulative model of preference evidence synthesis
- This requires that every AI-assisted step remains linked to the source evidence, open to expert review, and incorporated into the uncertainty assessment
- AI should therefore be seen as methodological infrastructure for evidence synthesis
  - not as a substitute for methodological expertise.

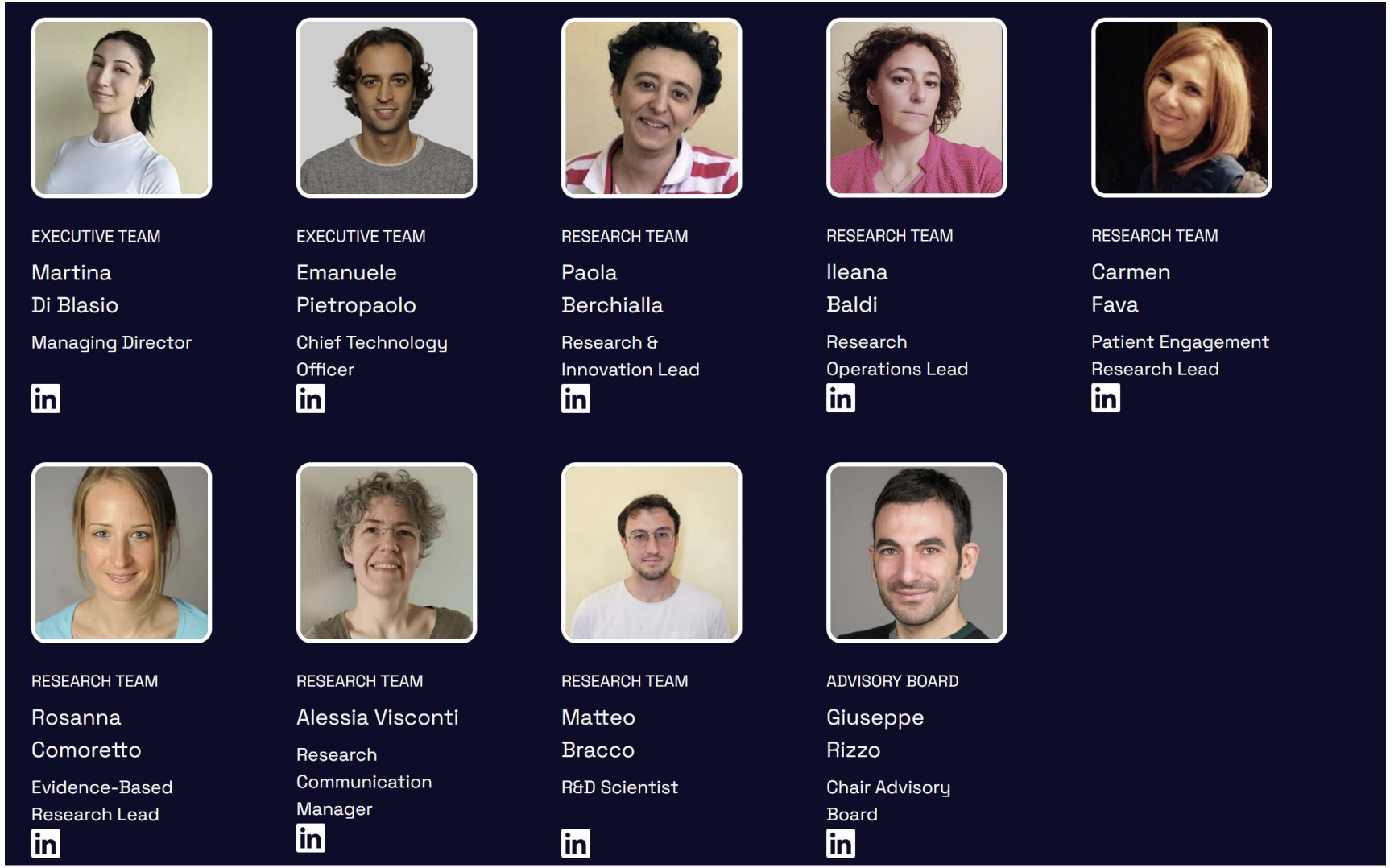
From automation to accountability

- AI-enabled meta-analysis can help address a major bottleneck in health preference research
  - the difficulty of synthesising heterogeneous preference studies in a timely and reproducible way
- The appropriate standard is a validated hybrid workflow
  - not *human-out-of-the-loop*
- **AI-assisted extraction, human-in-the-loop validation, reproducible modelling, and explicit uncertainty propagation**
  - it makes preference evidence more reusable, transferable, and trustworthy for decision making

Department
Clinical and
Biological Sciences

UNIVERSITÀ
DI TORINO

This is our team



paola.berchialla@unito.it

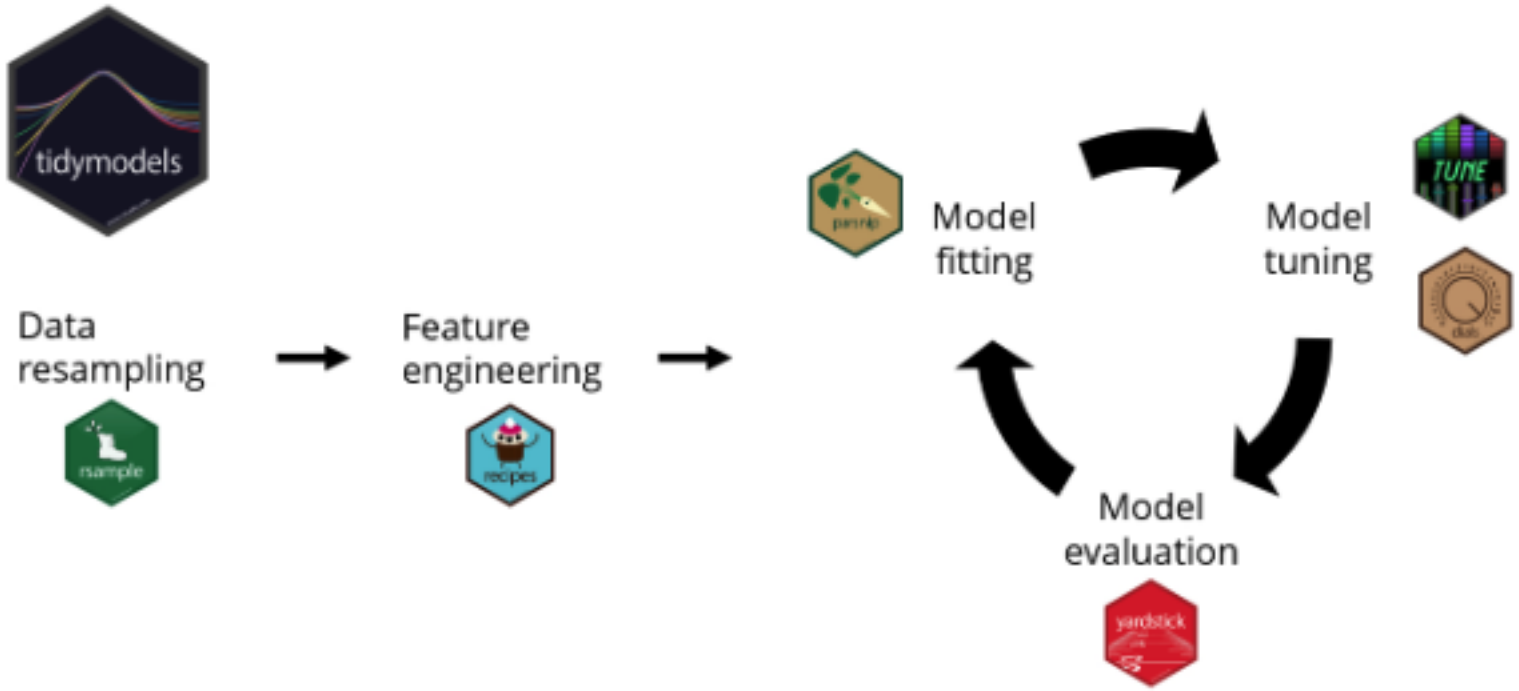
A Tutorial on tidymodels: Building Reproducible Modelling Workflows in R

Jason Nicholas, Harry Parr

GSK

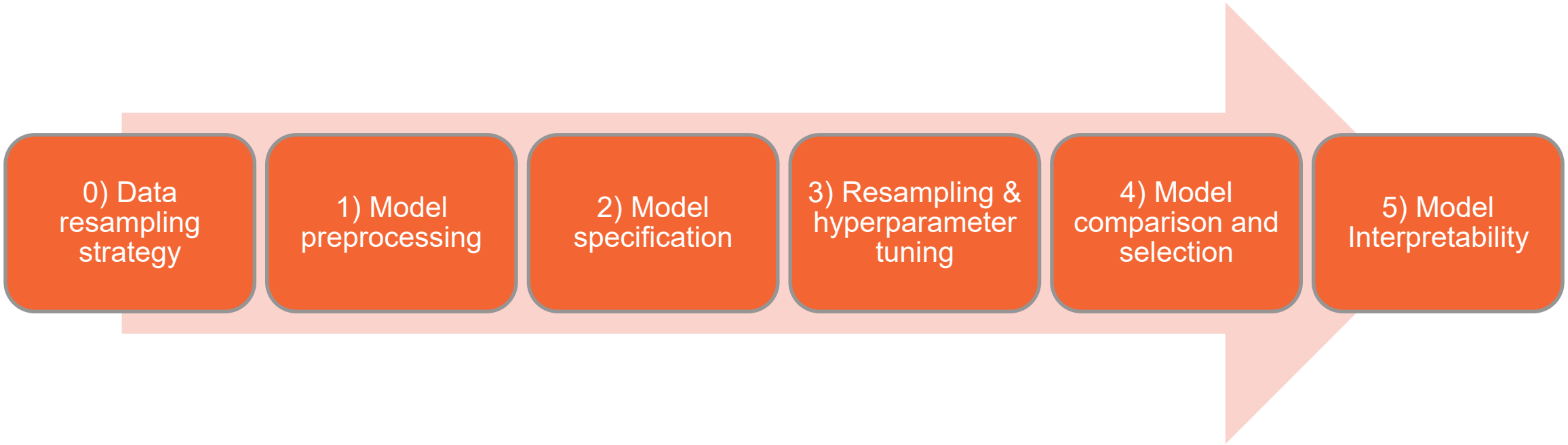
Tidymodels

- Tidymodels is a collection of R packages designed for statistical modelling and machine learning.
- It provides a structured and consistent framework for separating preprocessing, model specification, resampling, and evaluation within reproducible workflows



GSK

Tidymodels Workflow



GSK

Data Splitting and resampling strategy

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

- *initial_split()* is used to create separate training and test datasets.
  - The training set is used for model development, resampling, and hyperparameter tuning.
  - The test set is held out until the final model evaluation.

Data Splitting and resampling strategy

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

- *initial_split()* is used to create separate training and test datasets.
  - The training set is used for model development, resampling, and hyperparameter tuning.
  - The test set is held out until the final model evaluation.
- Repeated k-fold cross-validation (e.g. 5-fold CV x10) is created within the training data.
  - This provides an internal resampling framework for comparing candidate models and tuning hyperparameters.

Data Splitting and resampling strategy

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

- *initial_split()* is used to create separate training and test datasets.
  - The training set is used for model development, resampling, and hyperparameter tuning.
  - The test set is held out until the final model evaluation.
- Repeated k-fold cross-validation (e.g. 5-fold CV x10) is created within the training data.
  - This provides an internal resampling framework for comparing candidate models and tuning hyperparameters.
- Alternatively, internal cross-validation can be performed on the full dataset to estimate model performance if there is external data available for testing.

Data Splitting and resampling strategy

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

- *initial_split()* is used to create separate training and test datasets.
  - The training set is used for model development, resampling, and hyperparameter tuning.
  - The test set is held out until the final model evaluation.
- Repeated k-fold cross-validation (e.g. 5-fold CV x10) is created within the training data.
  - This provides an internal resampling framework for comparing candidate models and tuning hyperparameters.
- Alternatively, internal cross-validation can be performed on the full dataset to estimate model performance if there is external data available for testing.
- Stratification can be used for binary outcomes, particularly when classes are imbalanced, this helps to preserve the outcome distribution across resamples.

Data Splitting and resampling strategy

```
splits <- initial_split(df, prop = .8)
training_data <- training(splits)
testing_data <- testing(splits)
cv_folds <- vfold_cv(training_data , strata = OUTCOME, v=5, repeats = 10)
cv_folds_INTERNAL <- vfold_cv(df, strata = OUTCOME, v=5, repeats = 10)
```

- *initial_split()* is used to create separate training and test datasets.
  - The training set is used for model development, resampling, and hyperparameter tuning.
  - The test set is held out until the final model evaluation.
- Repeated k-fold cross-validation (e.g. 5-fold CV x10) is created within the training data.
  - This provides an internal resampling framework for comparing candidate models and tuning hyperparameters.
- Alternatively, internal cross-validation can be performed on the full dataset to estimate model performance if there is external data available for testing.
- Stratification can be used for binary outcomes, particularly when classes are imbalanced, this helps to preserve the outcome distribution across resamples.
- Other resampling approaches, such as bootstrapping, can also be used depending on the modelling objective.
  - Bootstrapping repeatedly samples the training data with replacement – particularly useful when the dataset is small.

Preprocessing via Recipes

- The Recipes package automates data transformations like normalisation, encoding and imputation.
- Ensures training and test datasets are processed consistently, preventing errors or data leakage.
- The Recipe is a reusable object which makes it easy to apply the same steps to multiple models

```
recipe <-
  recipe(OUTCOME ~ ., data = df) %>%
  update_role(STUDYID, SUBJID, new_role = "IDs") %>%
  step_zv(all_predictors()) %>%
  step_rm(REGIMEN) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_impute_median(all_numeric_predictors()) %>%
  step_normalize(all_numeric_predictors()) %>%
  step_naomit(OUTCOME)
```

GSK

Model Specification

- Defines what type of model you want to use
- Allows you to separate model definition from data, making workflows more modular
- Hyperparameters can be marked with *tune()* to allow systematic optimisation during training
- Tidymodels separates the model from the underlying engine, making it easy to switch implementations.
- This promotes reproducible and consistent model building across datasets.

```
mod_rf <-
  rand_forest(
    mtry = tune(),
    min_n = tune(),
    trees = 1000
  ) %>%
  set_engine("ranger") %>%
  set_mode("classification")
```

```
mod_logistic <-
  logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
```

```
mod_xgb <-
  boost_tree(
    trees          = 1000,
    tree_depth     = tune(),
    learn_rate     = tune(),
    min_n          = tune(),
    loss_reduction = 0,
    sample_size    = 0.8,
    mtry           = 0.8
  ) %>%
  set_engine("xgboost") %>%
  set_mode("classification")
```

```
mod_glmnet <- logistic_reg(
  penalty = tune(),
  mixture = tune()
) %>%
  set_engine("glmnet") %>%
  set_mode("classification")
```

GSK

Workflow Creation

- Workflows combine the preprocessing (recipe) and model specification into a single pipeline.
- This ensures all steps are applied consistently during training, tuning and testing
- Adding the recipe to *workflow_set()* automatically pairs it with every model, ensuring consistent preprocessing across all workflows.
- This also makes the workflow modular and reusable, so you can easily swap models or recipes.

```
wf_set <- workflow_set(
  preprocessing = list(recipe),
  models = list(
    random_forest = mod_rf,
    logistic = mod_logistic,
    xgboost = mod_xgb,
    glmnet = mod_glmnet
  )
)
```

GSK

Hyperparameter Tuning

- In the model specification stage, we identified which hyperparameters we plan to tune.
- In this stage we can automatically extract all tuneable hyperparameters using the *dials* package.
- We can then create a grid which generates 200 hyperparameter combinations and add this to the workflow.
- The grid size scales with the number of hyperparameters, simpler models require fewer parameter combinations, whereas models like XGBoost need a larger grid to explore the space properly.

```
rf_param <-
  mod_rf %>%
  extract_parameter_set_dials() %>%
  update(mtry = mtry(c(2, 30)))

rf_grid <- grid_latin_hypercube(
  rf_param,
  size = 200
)

wf_set <- wf_set %>%
  option_add(id = "recipe_random_forest", grid = rf_grid)
```

Source: https://www.youtube.com/watch?app=desktop&v=Evua529dAgc

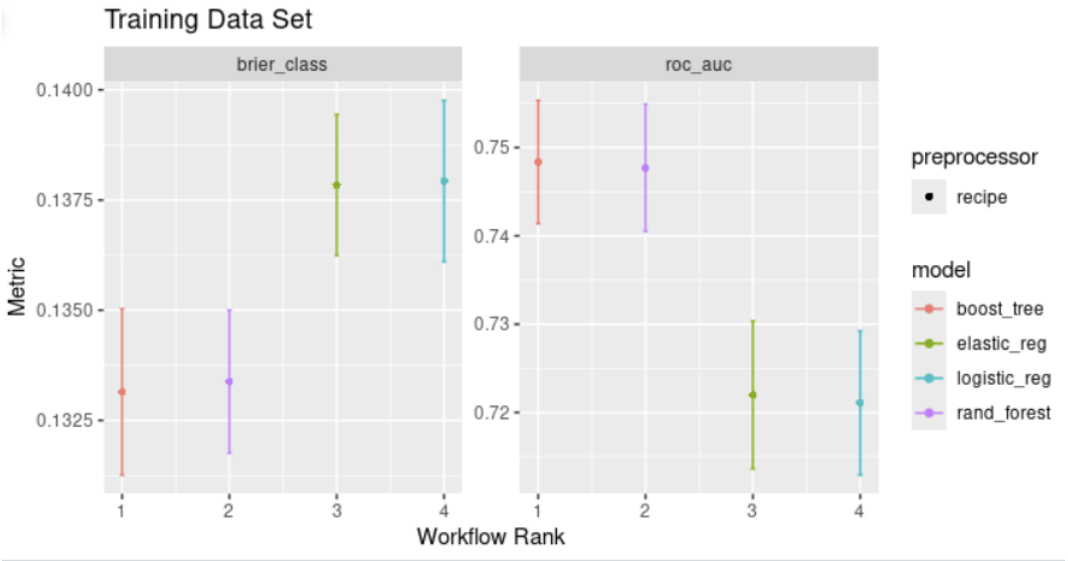
GSK

Cross-validation procedures to find optimal hyperparameters

- For each workflow, each hyperparameter combination is trained and then performance is evaluated across all CV folds.
- We then select the hyperparameter combination with the best cross-validated performance.

```
cv_folds <- vfold_cv(training_data, strata = OUTCOME,
                     v = 5, repeats = 10)

wf_set_tuned <- wf_set %>%
  workflow_map(
    "tune_grid",
    resamples = cv_folds,
    metrics = metric_set(roc_auc, accuracy)
  )

best_parameters <- wf_set_tuned %>%
  select_best(metric = "roc_auc")

autoplot(wf_set_tuned, select_best=TRUE)
```

```
wf_set_raced <- wf_set %>%
  workflow_map(
    "tune_race_anova",
    resamples = cv_folds,
    grid = rf_grid,
    metrics = metric_set(roc_auc, brier_class),
    control = control_race(verbose_elim=TRUE)
  )
```



GSK

Fitting the Final Model

- The finalised workflow combines the best hyperparameters with the preprocessing recipe and model specification into a single workflow ready for fitting.
- Fits the fully specified model using the training or complete dataset.

```
rf_wf_updated <- workflow() %>%
  add_model(rf_spec) %>%
  add_recipe(recipe)

rf_wf_final <- finalize_workflow(rf_wf_updated, best_parameters)
rf_fit <- fit(rf_wf_final, data = training_data)
```

GSK

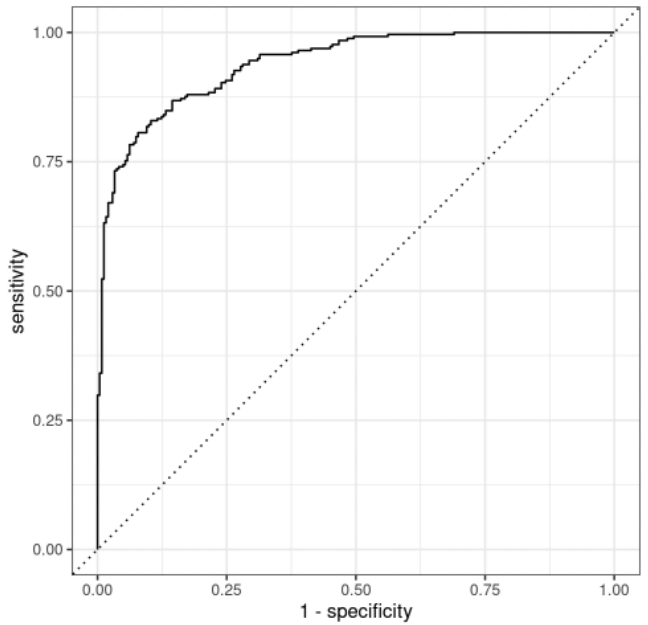
Model Evaluation

- Tidymodels uses **yardstick** to compute performance metrics for the predictive models.
- ***yardstick*** provides a consistent and standardized interface for model evaluation.
- It supports a wide range of metrics (e.g. ROC AUC, accuracy, confusion matrices).

```
preds <- predict(rf_fit, new_data=test_data, type = "prob") %>%
  bind_cols(predict(rf_fit, test_data, type = "class"))

roc_auc(preds, truth = OUTCOME, .pred_0)

conf_mat(preds, truth = OUTCOME, .pred_class)

roc_auc_curve <- roc_curve(preds, OUTCOME, .pred_1)
```

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.740
```

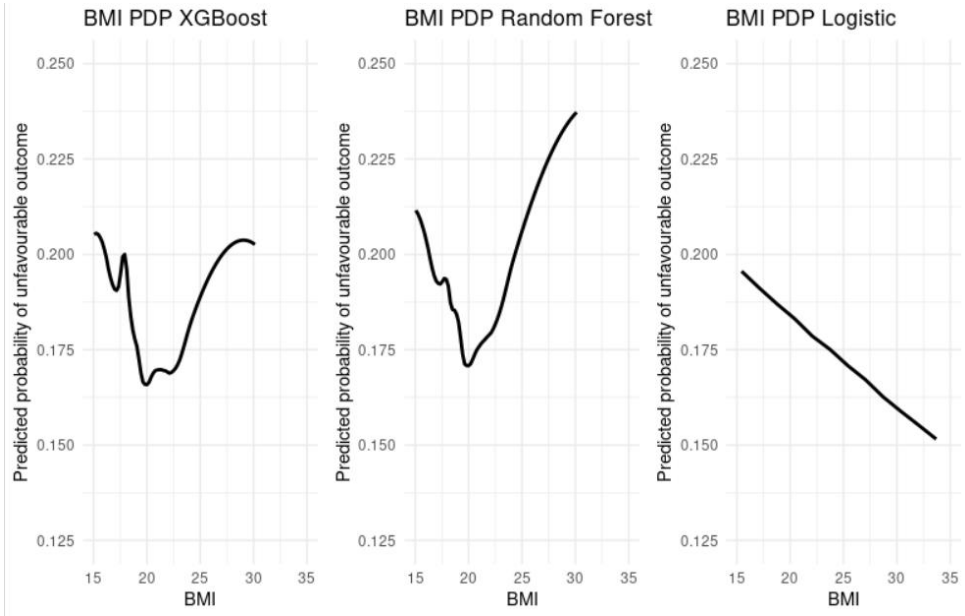
```
            Truth
Prediction   0   1
         0  56  18
         1  86 184
```

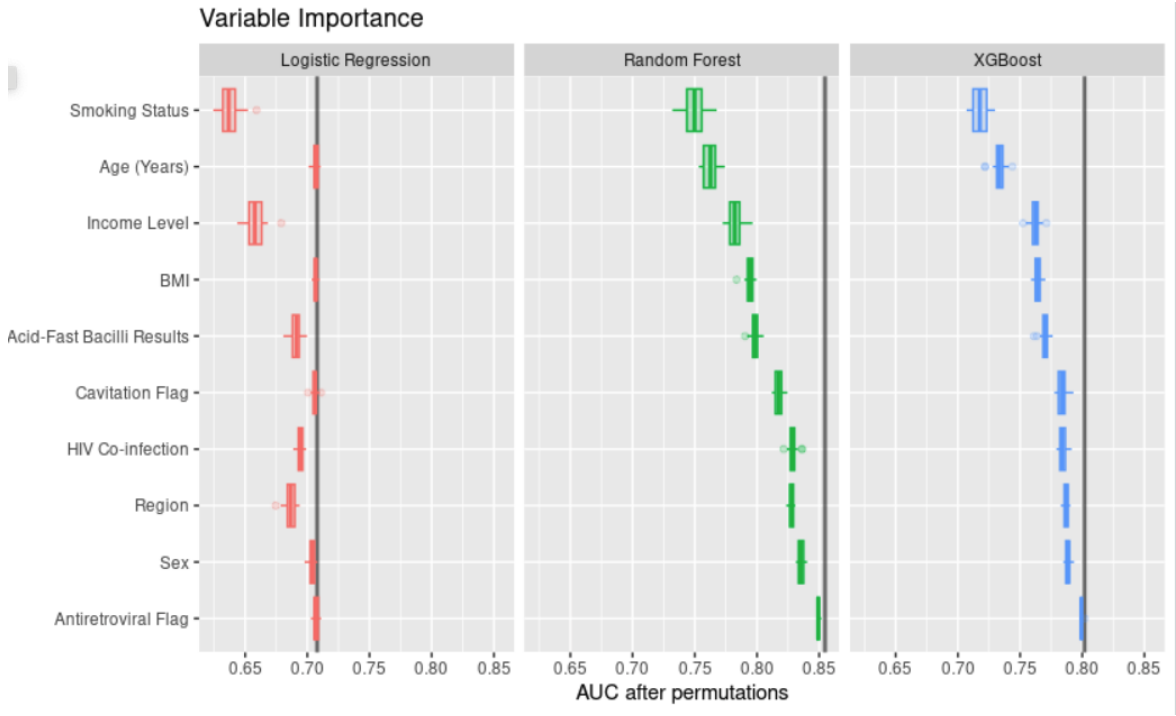


GSK

Model interpretation

- Performance metrics describe how well a model predicts
- Interpretability tools help explain how the model is using the predictors.
- Variable importance plots (VIPs) can identify which predictors contribute most to model performance.
- Partial dependence plots (PDPs) show the average relationship between a predictor and the modelled outcome.





Closing Remarks

Key technical takeaways

- **Tidymodels enforces good practice by design** — encourages reproducibility, leakage prevention, and consistent evaluation
- **Everything is modular** — you can swap models, recipes, or metrics without rebuilding from scratch
- **The workflow is end-to-end** — from splitting data all the way to interpreting results, it's one coherent pipeline rather than a patchwork of separate steps.

Here we presented a concise primer on Tidymodels, which provides a cohesive ecosystem for building, tuning, and validating predictive models in R, with consistent workflows that strengthen reproducibility and evaluation rigor.

GSK