

Detecting Logical Errors in Clinical Programming by Applying ChatGPT

ML16 Presenter
ILLIA SKLIAR
Senior Statistical
Programmer/Analyst

NOVEMBER 16-19, 2025

Agenda



Introduction



Methodology & Prompt Engineering



Case Study - BOR Derivation



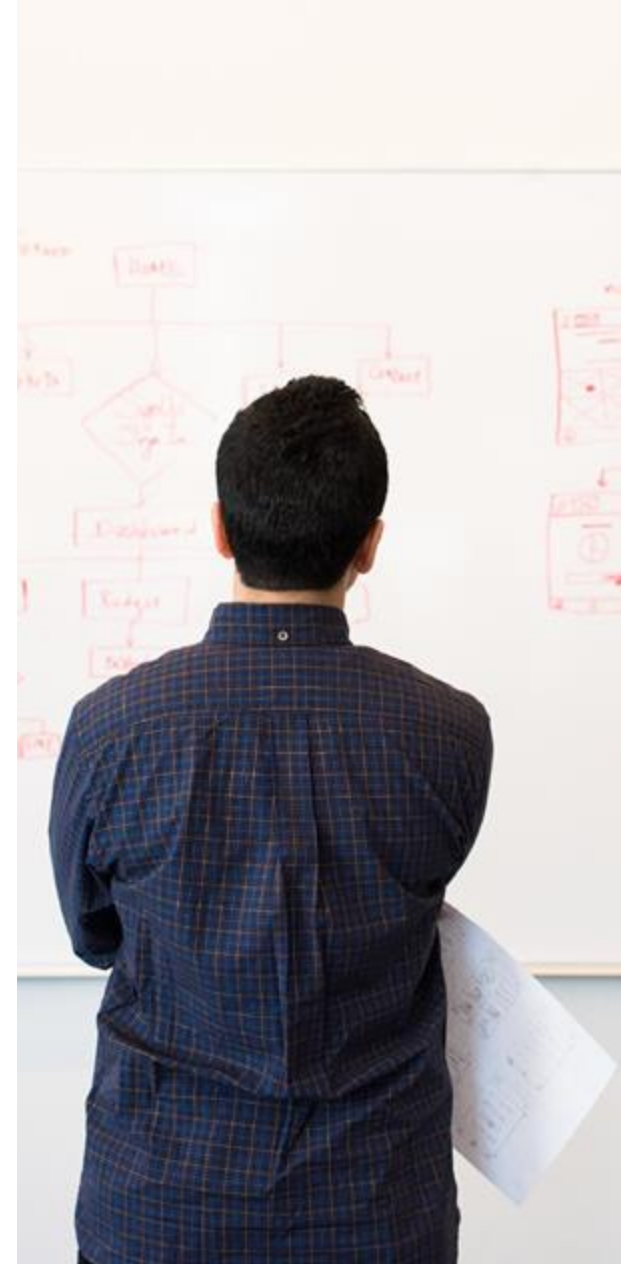
Complex Logic & Limitations



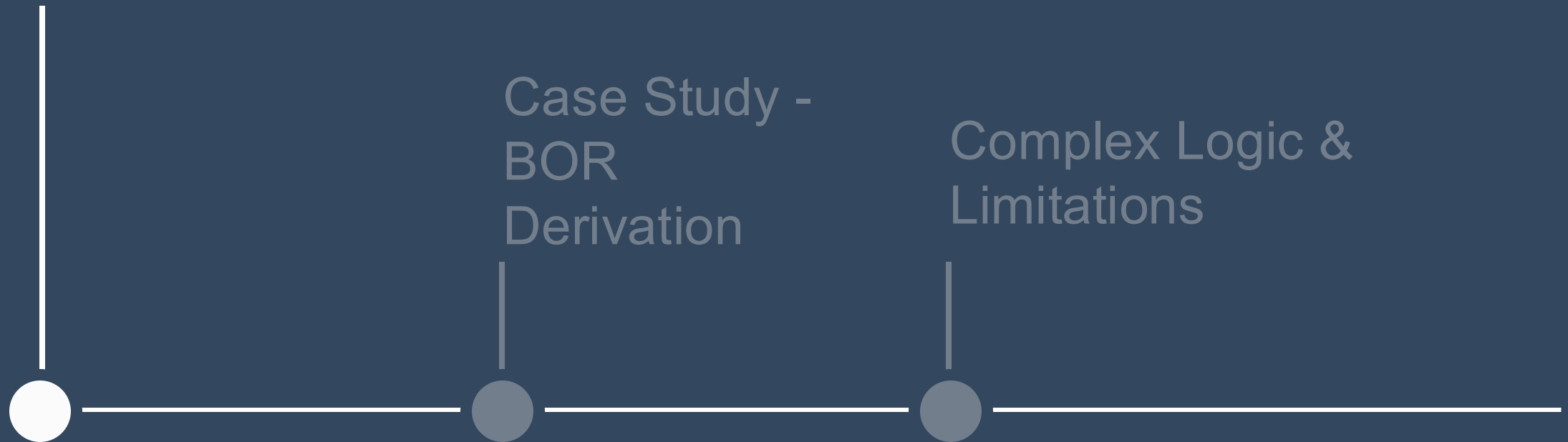
Regulatory Considerations

Introduction

- Logical errors produce **plausible** but **incorrect** results when SAS code misapplies SAP rules, making detection challenging
- Traditional QC via double programming can't guarantee **semantic correctness** on early stages of development
- Existing QC **focuses on output** differences rather than alignment with the **intended analysis logic**



Methodology & Prompt Engineering



Integration of ChatGPT in SAS Environment

```
1 %macro RunGPTquery(query =);
2
3   %let API_KEY= sk-5cZRL3N4W...;
4   %let QUESTION = %str("%&query.%");
5
6   /* Body of the POST request */
7   filename in temp;
8   data _null_;
9   file in;
10  put;
11  put "{";
12  put '"model": "gpt-5", "messages": [{"role": "user", "content": '"&question "}"]';
13  put "}";
14  run;
15
16  /* reference that file as IN= parm in PROC HTTP POST */
17  filename resp "%sysfunc(getoption(WORK))/echo.json";
18  proc http
19    method="POST"
20    url="https://api.openai.com/v1/chat/completions"
21    ct="application/json"
22    in=in
23    out=resp;
24    headers "Authorization" = "Bearer &api_key.";
25  run;
26
27  /* Tell SAS to parse the JSON response */
28  libname response JSON fileref=resp;
29
30  proc report data=response.choices_message;
31    column content;
32    define content / display "" style(column)=[cellwidth=6in fontsize=12pt asis=ON];
33  run;
34
35 %mend RunGPTquery;
```

Prompt Engineering Strategy

Role Instruction

“You are a senior statistical programmer or an FDA auditor reviewing SAS code for compliance with the SAP.”

User Prompt

SAP Rule: Event that start before treatment start date should not be classified as TEAEs

SAS Code: if TRTSDT <= ASTDT & ASTDT <= min(TRTEDT + 30, FUACDT);

Validation Instruction

“Identify any discrepancies between the SAP rule and the SAS code logic. Focus on logic only and ignore coding style or syntax.”

Case Study - BOR Derivation



Case Study: BOR Derivation



SAP rule: CR/PR must be confirmed at least **4 weeks** later without intervening PD;
unconfirmed responses are excluded

SAS code chooses the best response (CR > PR > SD > PD) but **omits** the confirmation check

Expecting **ChatGPT** to spot this mismatch and highlighted the risk of overstated efficacy

Code Input

```
1
2 %let CODE_PART= %nrstr(
3 proc sort data=responses out=resp_sorted;
4     by subject response_date;
5 run;
6
7 data bor_derivation;
8     set resp_sorted;
9     by subject;
10    retain best_resp 'PD';           /* initialise BOR as worst (PD) */
11    if first.subject then best_resp = 'PD';
12    /* update best_resp based on hierarchy CR > PR > SD > PD */
13    if response = 'CR' then best_resp = 'CR';
14    else if response = 'PR' and best_resp not in ('CR') then best_resp = 'PR';
15    else if response = 'SD' and best_resp not in ('CR','PR') then best_resp = 'SD';
16    if last.subject then do;
17        bor_resp = best_resp;        /* assign derived BOR result */
18        output;
19    end;
20 run;
21 );
```

SAP Input & RunGPT call

```
23 %let SAP_INSTRUCTIONS = %nrstr(  
24 1. BOR is defined as the best tumour response from treatment start until progression  
25 responses are prioritised as CR > PR > SD > PD.  
26 2. A Complete Response (CR) or Partial Response (PR) must be confirmed  
27 by a subsequent tumour assessment of CR or PR, at least four weeks later,  
28 with no intervening Progressive Disease (PD).  
29 3. Unconfirmed CR/PR should not be counted as confirmed responses.  
30 Stable Disease (SD) and PD follow their standard definitions.  
31 );  
32  
34 %RunGPTquery(query=SYSTEM MESSAGE: You are a clinical data analysis assistant tasked with  
35 verifying that a SAS program correctly implements an SAP rule.  
36 USER MESSAGE:  
37 <<SAP RULE>> &SAP_INSTRUCTIONS.  
38 <<SAS CODE>> &CODE_PART.  
39 <<TASK>>  
40 Identify any logical inconsistencies between the SAP's BOR criteria and the SAS code.  
41 Explain any discrepancies);  
42
```

[CODE](#)[LOG](#)[RESULTS](#)[▶ Table of Contents](#)

Key inconsistencies between the SAP and the SAS code

No confirmation logic for CR/PR

- SAP: CR and PR must be confirmed by a subsequent CR or PR at least 4 weeks later with no intervening PD.

Unconfirmed CR/PR should not be counted as confirmed responses.

- Code: Sets best_resp to CR or PR at the first occurrence and never checks for a confirming assessment 4-week timing or intervening PD.

- Impact: Subjects with a single unconfirmed CR/PR will be incorrectly classified as CR/PR for BOR.

Complex Logic & Limitations

Methodology
& Prompt
Engineering

Case Study -
BOR
Derivation



SAP Input

Table 1.1: RECIST v1.1 Best Overall Response when confirmation of CR and PR required

Overall Response first timepoint	Overall Response subsequent timepoint	Best Overall Response
CR	CR	CR
CR	PR	PR
CR	SD	SD provided minimum criteria for SD duration met, otherwise, PD
CR	PD	SD provided minimum criteria for SD duration met, otherwise, PD
CR	NE	SD provided minimum criteria for SD duration met, otherwise, NE
PR	CR	PR
PR	PR	PR
PR	SD	SD provided minimum criteria for SD duration met, otherwise, NE
PR	PD	SD provided minimum criteria for SD duration met, otherwise, PD
PR	NE	SD provided minimum criteria for SD duration met, otherwise, NE
SD	SD	SD provided minimum criteria for SD duration met, otherwise, NE
SD	PD	SD provided minimum criteria for SD duration met, otherwise, PD
SD	NE	SD provided minimum criteria for SD duration met, otherwise, NE
NE	NE	NE
No post-baseline tumor assessments		Not Done

Note: PR or CR assessments separated by one NE assessment (e.g., PR-NE-PR or CR-NE-CR) will be considered confirmed PR or confirmed CR, respectively.

Code Input

```

188 /*=====
189 * 5) Decide BOR with precedence and timing:
190 *   1) Confirmed CR before PD → CR
191 *   2) Else confirmed PR before PD → PR
192 *   3) Else SD eligible before PD → SD
193 *   4) Else if PD exists → PD
194 *   5) Else if any post-baseline assessment → NE
195 *   6) Else → Not Done
196 *=====*/
197 proc sql;
198   /* Any post-baseline assessment? */
199   create table any_post as
200   select USUBJID, (min(ADT) is not null) as HAS_POST length=3
201   from resp
202   group by USUBJID;
203
204   create table bor_stage as
205   select a.USUBJID,
206          d.PD_DT,
207          c.CR_CONF_DT, c.FIRST_CR_DT,
208          p.PR_CONF_DT, p.FIRST_PR_DT,
209          s.SD_ELIG_DT,
210          ap.HAS_POST
211   from adsl_srt as a
212   left join first_pd as d on a.USUBJID=d.USUBJID
213   left join conf_cr as c on a.USUBJID=c.USUBJID
214   left join conf_pr as p on a.USUBJID=p.USUBJID
215   left join sd_hits as s on a.USUBJID=s.USUBJID
216   left join any_post as ap on a.USUBJID=ap.USUBJID;
217 quit;

```

```

219 data bor_final;
220   set bor_stage;
221   length BOR $2 BORDT LABEL $60;
222   format BORDT date9.;
223
224   %macro before_pd(dt);
225     (not missing(&dt) and (missing(PD_DT) or &dt <= PD_DT))
226   %mend;
227
228   /* Precedence and timing */
229   if %before_pd(CR_CONF_DT) then do; BOR='CR'; BORDT=CR_CONF_DT; end;
230   else if %before_pd(PR_CONF_DT) then do; BOR='PR'; BORDT=PR_CONF_DT; end;
231   else if %before_pd(SD_ELIG_DT) then do; BOR='SD'; BORDT=SD_ELIG_DT; end;
232   else if not missing(PD_DT) then do; BOR='PD'; BORDT=PD_DT; end;
233   else if HAS_POST then do; BOR='NE'; BORDT=.; end;
234   else do; BOR=''; BORDT=.; end;
235
236 run;

```

There was no clear GPT-5 response

Missed mismatches:
multi-step logic and
minimum SD duration
often not recognized

False positives:
intermediate datasets
and macros
misinterpreted as errors

Limited context window
leads to contradictory
suggestions and loss of
earlier context

Reliability drops sharply
for complex,
macro-driven programs



Regulatory Considerations

Protect participant confidentiality. Send only de-identified specifications and code

Validate systems per Part 11. Maintain audit trails, record retention and version control

Follow AI ethics: human oversight, transparency and accountability. Treat AI output as advisory

Mask identifiers, version prompts and models, and combine AI checks with traditional QC

Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products

Discussion Paper and Request for Feedback

“

...These applications [AI/ML] can potentially increase data accuracy and improve the speed at which data are prepared for analyses...

”

TITLE 21--FOOD AND DRUGS
CHAPTER I--FOOD AND DRUG ADMINISTRATION
DEPARTMENT OF HEALTH AND HUMAN SERVICES
SUBCHAPTER A - GENERAL

PART 50 PROTECTION OF HUMAN SUBJECTS

TITLE 21--FOOD AND DRUGS
CHAPTER I--FOOD AND DRUG ADMINISTRATION
DEPARTMENT OF HEALTH AND HUMAN SERVICES
SUBCHAPTER A - GENERAL

PART 11 ELECTRONIC RECORDS; ELECTRONIC SIGNATURES

TITLE 21--FOOD AND DRUGS
CHAPTER I--FOOD AND DRUG ADMINISTRATION
DEPARTMENT OF HEALTH AND HUMAN SERVICES
SUBCHAPTER A - GENERAL

PART 56 INSTITUTIONAL REVIEW BOARDS

TITLE 21--FOOD AND DRUGS
CHAPTER I--FOOD AND DRUG ADMINISTRATION
DEPARTMENT OF HEALTH AND HUMAN SERVICES
SUBCHAPTER A - GENERAL

PART 54 FINANCIAL DISCLOSURE BY CLINICAL INVESTIGATORS

AI
NLP
GPT

Conclusion & Future Directions



- ChatGPT is a helpful semantic reviewer for simple, linear SAS programs



- Reliability declines on complex logic, macros and long context. Outputs can miss mismatches or hallucinate



- Use AI output as advisory. Combine with human expertise, decomposition of tasks and iterative prompting



- Refine prompt strategies, extend context windows, and explore improved models

Thank you for watching

Contact

Illia.Skliar@intego-group.com

www.intego-group.com

