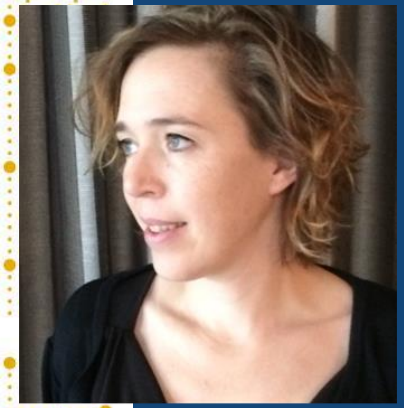
Development of USDM through translation of human-readable protocols

Jasmine Kestemont (Innovion)
Stijn Rogiers (argenx)

Meet the Speakers



Jasmine Kestemont

Title: Managing Partner / Consultant Life Sciences

Organization: Innovion

Jasmine Kestemont is an entrepreneurial business leader with significant experience in both Sponsor and services organizations. After a few years of global work experience, she has returned to the area she is most passionate about, data and project management, with a special interest in data standardization and regulatory submissions. Jasmine is Head of Data Management at argenx.

Stijn Rogiers

Title: Head Data Integration & Standards, DM

Organization: argenx

Stijn joined argenx in June 2022 as Head Data Integrations and Standards (Data Management). Argenx is a fast-moving and growing Biotech company. Stijn has 20+ years of experience both at CRO, Pharma industry and Technology. He worked 5 years at SGS Life Sciences (CRO), 10 years within Janssen (Johnson & Johnson) and 7 years at SAS Institute (analytics leader) before moving to argenx. Stijn is also a member of the European CDISC Coordinating Committee (E3C).

LinkedIn: www.linkedin.com/in/stijnrogiers

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

Agenda

1. **Background to the project**
2. The Journey
3. The hurdles
4. The result

How did we get here?

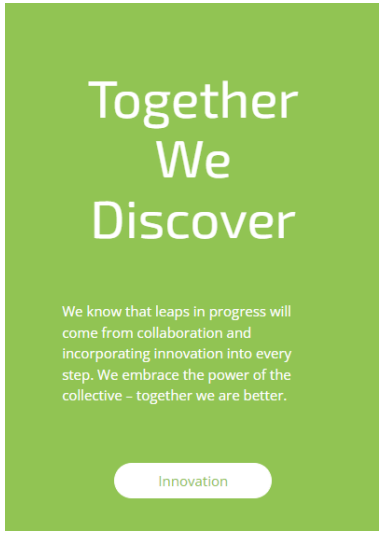
Workshop CDISC

PHUSE EU Connect 2023 - DDF Workshop - PUBLIC - Wiki (cdisc.org)

Need to standardize protocols

ICH M11 guideline, clinical study protocol template and technical specifications - Scientific guideline | European Medicines Agency (europa.eu)

Explore new technology



Phuse EU Connect 2023 – DDF workshop (Birmingham UK)

All content on this Wiki is non-binding and any individual opinions expressed should not be considered indicative of the policies or positions of CDISC or any other organization.

cdisc Wiki Spaces Search Log in Sign up

cdisc PUBLIC

PAGE TREE

- CDISC Wiki Terms of Use
- ODM v2.0
- Introduction to Therapeutic Area Standards
- CDISC User Networks
- Public Review Instructions
- PHUSE US Connect 2023 - ARS Workshop
- **PHUSE EU Connect 2023 - DDF Workshop**
- PHUSE US Connect 2024 - DDF Workshop
- EU Interchange 2024 Digital Data Flow Worksl
- Wiki PDF Export will be disabled

Pages / Welcome to CDISC WIKI

PHUSE EU Connect 2023 - DDF Workshop

Created by John Owen, last modified on Nov 28, 2023

Pre EU Connect 2023 Information 27 Oct 2023	Listen to the preparation Webinar and review the preparation webinar slides
Pre-Reads for EU Connect 2023	Pre-Reads *(Materials to look at prior to the workshop if you wish to. NOT compulsory!)* *A the time of the workshop this was version 2.5, since the workshop this now points to the latest versions of the USDM* • Model (UML) • Controlled Terminology (XLSX) • Implementation Guide (PDF) • Informative Diagram (PNG) • Miro Board (Web) (P: CDISC-DDF-SME) (or if you prefer you can download a PDF of the Miroboard)
Web tools	Web Tools *(no need to install - these will run from a web browser)* • Excel To JSON Tool (U: PHUSE - P: learning_usdm) • Excel to JSON Tool readme • Excel to JSON Tool Infographic • JSON Comparison
Example files for EU Connect 2023 Workshop 05 Nov 2023	CDISC_Pilot_Study_Baseline.xlsx Example Protocol • SoA Pages.jpeg • SoA.png
Slides from EU Connect 2023 workshop 05 Nov 2023	**Slides** presented at the workshop on 05 Nov 2023
CDISC DDF EU Connect 2023 workshop 07 Nov 2023	2023 11 07 PHUSE Peter VR DS01 M11 - PHUSE EU Connect v0.5.pdf 2023 11 07 PHUSE DAVE IH DS02 V3.pdf
EU Connect 2023 Follow-up Webinar 28 Nov 2023	Listen to the Webinar and review the webinar slides Example files • Demography - Adding BCs • Vital signs - Adding timeline (demo'd at the follow-up webinar)
Link to DDF Orientation page on the CDISC WIKI	Digital Data Flow (DDF) Team Home/Orientation (CDISC Wiki account required)

ICH M11 guideline

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

Search

Medicines | Human regulatory | Veterinary regulatory | Committees | News & events | Partners & networks | About us

Home > ICH M11 guideline, clinical study protocol template and technical specifications - Scientific guideline

ICH M11 guideline, clinical study protocol template and technical specifications - Scientific guideline

Share

Human | Scientific guidelines

Page contents

The purpose of this new harmonised guideline is to introduce the clinical protocol template and the technical specification to ensure that protocols are prepared in a consistent fashion and provided in a harmonised data exchange format acceptable to the regulatory authorities. The ICH M11 Clinical Electronic Structured Harmonised Protocol Template provides comprehensive clinical protocol organization with standardized content with both required and optional components. The Technical Specification that are acceptable to all regulatory authorities of the ICH regions presents the conformance, cardinality, and other technical attributes that enable the interoperable electronic exchange of protocol content with a view to develop an open, non-proprietary standard to enable electronic exchange of clinical protocol information.

Keywords: protocol, harmonised template, interventional clinical trials, technical specification, data exchange, non proprietary standard

Initial ambition

- Historical protocols – machine readable – searchable library
  - Secondary goal: prove machine readable to human readable via existing protocol
    - Existing protocol → USDM → human readable

- Future protocols

- SoA :
  - Determine: - cost of study (cost grid of assessments)
    - duration of visit (patient feasibility)
  - drive - protocol consistency (quality and speed)
    - CRF design
  - facilitate pooling of data
  - issue tracking and risk mitigation

Agenda

1. Background to the project
2. **The Journey**
3. The hurdles
4. The result

The Journey

- Team members

SAS:	Jean-Charles Haillus	(Senior Project Manager, SAS),
	Koen Knapen	(Principal Analytical Consultant , SAS),
	Rens Feenstra	(Principal Technology Solution Consultant , SAS),
	Fadi Glor	(Senior Account Executive , SAS)
argenx:	Jasmine Kestemont + Stijn Rogiers (see Bio)	
	Sandeep Juneja	(Clinical Solutions, DML, argenx)

- Semi-weekly meetings + sprint cycles

- Need for both technical expertise and content expertise

  Behind the scenes additional domain experts consulted (LLM, Python, ...)

Scope POC

- Do **contextual information extraction** from 2 clinical trial protocols (early and late phase) and **store the relevant info** into the USDM (excel) workbook … **in the right place** (in the right field).

- **Scalability:** Make sure the contextual information extraction is generic enough such that the text model used works on as many protocols as possible (including those never seen during training).

- Avoid over-training (over-fitting).

Timelines

6 Feb:
KoM

22 Feb
First view on deliverables

28 Feb

Call Dave IH

Rethink scope

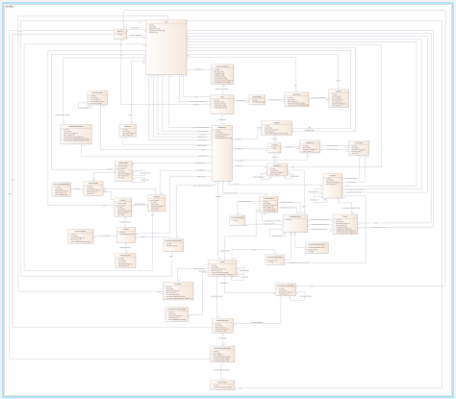
13 Mar

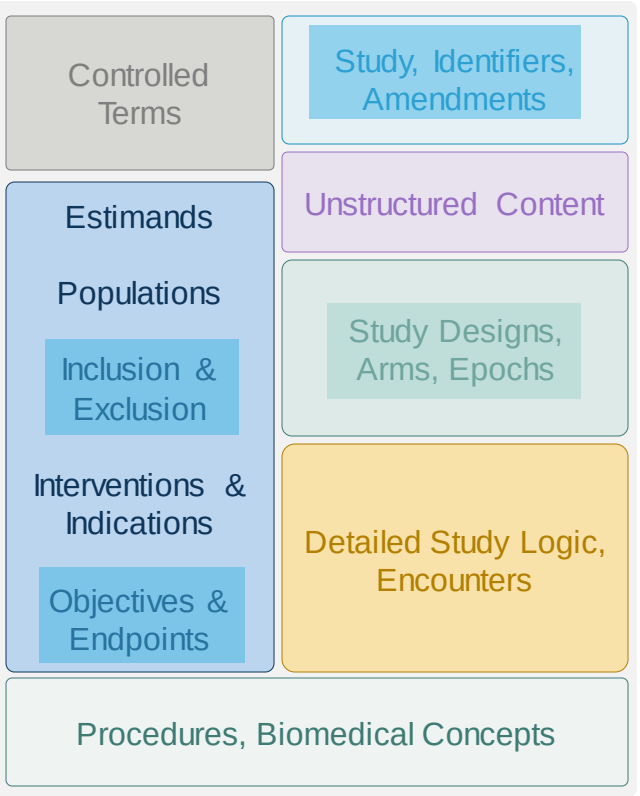
Explore table extractors

28 Mar

Image extractor issue resolution

USDM Content





Focus for PoC

1. Study Overview (Text)
   Study & StudyIdentifiers
2. Inclusion & Exclusion (Text)
   StudyDesignEligibilityCriteria
3. Study Objectives & Endpoints (Text)
   studyDesignOE
4. Investigational Plan:
   - 4.1 Study Design (Text + Image)
     studyDesign
   - 4.2 SoA (Table)
     mainTimeline
   - 4.3 Arms (Text)
     studyDesignArms

USDM Excel

study Sheet

	A	B	C	D	E	F	G
1	**name**	SCOPE1					
2	**studyTitle**	Simple Test 1					
3	**studyVersion**	1					
4	**studyType**	Interventional Study					
5	**studyPhase**	C15602					
6	**studyAcronym**	SIMPLE					
7	**studyRationale**	A simple test					
	businessTherapeuticAreas	SPONSOR: VAC=Vacines Group,					
8		SPONSOR: REG=Regulatory					
9	**briefTitle**	Something Brief					
10	**officialTitle**	Something Very Official					
11	**publicTitle**	Something Public					
12	**scientificTitle**	Somethign Clever But New					
13	**protocolVersion**	1					
14	**protocolStatus**	draft					
15							
16	**category**	**name**	**description**	**label**	**type**	**date**	**scopes**
17	study_version	Approval	Design approval date	Design Approval	Sponsor Approval Date	01/01/2023	country : GBR, country:FRA, region:ASIA, country :USA
18	protocol_document	Approval	Protocol document approval date	Protocol Approval	Protocol Effective Date	01/01/2023	Global
19	protocol_document	Approval	Protocol document approval date	Protocol Approval	Protocol Effective Date	01/02/2023	region:asia

USDM Excel Sheet Formats & Links Infographic

17th January 2024, USDM Package v0.43

Details the excel workbook format as used by the **USDM** python package.

Details of the package can be found at https://github.com/data4knowledge/usdm. Details for using the package and the sheet formats are detailed within the readme file within the repository.

studyIdentifiers Sheet

	A	B	C	D	E	F
1	**organisationIdentifierScheme**	**organisationIdentifier**	**organisationName**	**organisationType**	**studyIdentifier**	**organisationAddress**
2	USGOV	CT-GOV	ClinicalTrials.gov	Study Registry	NCT12345678	line|city|district|state|postal_code|GBR
3	DUNS	123456789	ACME Pharma	Clinical Study Sponsor	AP1234	Somewhere|In a City|In a District|In a big state|12345|FRA

studyDesignEligibilityCriteria Sheet

	A	B	C	D	E	F	G
1	**category**	**identifier**	**name**	**description**	**label**	**text**	**dictionary**
2	Inclusion	01	Age Criteria	The study age criterion		Subjects shall be between [min_age] and [max_age]	IE_Dict
3	Inclusion	02	Age Criteria Error	The study age criterion with error		Subjects shall be between [min_age] and [max_agexxx]	IE_Dict

USDM Excel (Cont'd)

studyDesignOE Sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	objectiveName	objectiveDescription	objectiveLabel	objectiveText	objectiveLevel	objectiveDictionary	endpointName	endpointDescription	endpointLabel	endpointText	endpointPurpose	endpointLevel	endpointDictionary
2	OBJ1	Primary		The primary efficacy objective for this study is to evaluate the efficacy of TCZ compared with placebo in combination with SOC for the treatment of severe COVID-19 pneumonia	Study Primary Objective		END1	Day 28, 7 category scale		Clinical status assessed using a 7-category ordinal scale at Day 28		Primary Endpoint	
3	OBJ2	Secondary		The secondary efficacy objective for this study is to evaluate the efficacy of TCZ compared with placebo in combination with SOC for the treatment of severe COVID-19 pneumonia over the age of [min_age]	Study Secondary Objective	OE_Dict	END2	TTCI		Time to clinical improvement (TTCI) defined as a National Early Warning Score 2 (NEWS2) of <=2 maintained for 24 hours		Secondary Endpoint	
4							END3	Time to Imporvement					

studyDesignElements Sheet

	A	B	C	D	E
1	name	studyElementDescription	Label	transitionStartRule	transitionEndRule
2	EL1	Screening Element	Screening	Study Start	Screened
3	EL2	Baseline Element	Baseline	Screened	Radomized
4	EL3	Treatment Element 1	Treatment 1	Radomized	Completed treatment 1
5	EL4	Follow Up Element	Follow Up	Treated	Leave Study
6	EL5	Treatment Element 2	Treatment 2	Radomized	Completed treatment 2

studyDesign Sheet

	A	B	C	D	E
1	studyDesignName	Study Design 1			
2	studyDesignDescription	The main design for the study			
3	therapeuticAreas	SPONSOR:T2_DIABETES=Type 2 diabetes, SNOMED: 73211009=Diabetes mellitus (disorder)			
4	studyDesignRationale	Basic study			
5	studyDesignBlindingScheme	OPEN LABEL			
6	trialIntentTypes	BASIC SCIENCE, DEVICE FEASIBILITY			
7	trialTypes	Efficacy Study			
8	interventionModel	C82639			
9	mainTimeline	mainTimeline			
10	otherTimelines				
11					
12	Epoch/Arms	Screening	Baseline	Treatment	Follow-Up
13	Active	EL1	EL2	EL3, EL5	EL4
14	Placebo	EL1	EL2	EL3, EL3	EL4

studyDesignArms Sheet

	A	B	C	D	E	F
1	name	description	label	type	dataOriginDescription	dataOriginType
2	Active	Active Substance	Active Substance	Active Comparator Arm	Data collected from subjects	Data Generated Within Study
3	Placebo	Placebo	Placebo	Placebo Comparator Arm	Data collected from subjects	Data Generated Within Study

<timeline name> Sheet

	A	B	C	D	E		G	H	I	J
1	Name	Main Timeline	name	SCREEN	PRE DOSE	DOSE	D14	PROG	D28	FU
2	Description	This is the main timeline for the study design.	description	Screening	Pre Dose	Dosing	Day 14	Check Opt In	Day 28 Optional	Follow Up
3	Condition	Potential subject identified	label	Screen	Baseline	Treatment	Day 14	Check opt In by subject	Day 28	Follow Up
4			type	Activity	Activity	Activity	Activity	Decision	Activity	Activity
5			default	PRE DOSE	DOSE	D14	PROG	D28	FU	(EXIT)
6			condition					FU: if opted out		
7			epoch	Screening	Baseline	Treatment	Treatment		Treatment	Follow-Up
8			encounter	E1	E2	E3	E4		E5	E6
9	Parent Activity	Child Activity	BC/Procedure/Timeline							
10	-	Demographics	BC:Age, BC:Sex, BC:Race, BC:Body Weight	X						
11	-	Procedures	PR1, PR2, PR2	X	X	X	X			X
12	-	Optional Weight	BC:Weight	X			X			
13		Optional	BC:SYSBP, BC:DIABP						X	

POC approach (Cont'd)

1) **Usage of LITI rules** for contextual information extraction

   LITI includes concept rule types as well as fact rule types.
   LITI is proprietary syntax from SAS.
   LITI = **L**anguage **I**nterpretation for **T**extual **I**nformation

2) **Usage of LLM (RAG-for-LLMs)** for contextual information extraction

   RAG = **R**etrieval-**A**ugmented **G**eneration

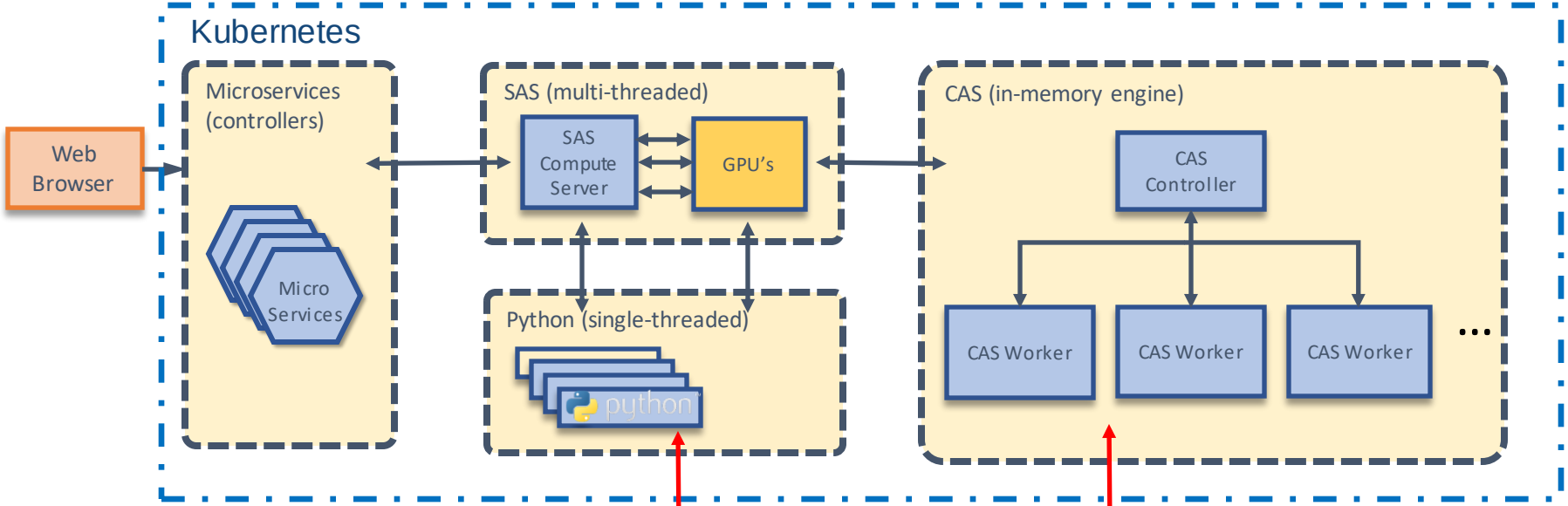
3) **PDF Table Extractor(s)**

4) **PDF Image Extractor(s)**

5) **Load from SAS tables into Excel and Word in an automated way**

   *while retaining the link i.e. if the info in the SAS table changes, then you can just refresh the *.xlsx or *.docx file to see that new info reflected.*

The environment:
SAS Viya + Open Source



Code

```
73  model_name = 'meta-llama/Llama-2-7b-hf' #@param
74  llm = HuggingFaceLLM(
75      context_window=context_window,
76      max_new_tokens=2048,
77      generate_kwargs={"temperature": temperature, "do
78      #system_prompt=system_prompt,
79      query_wrapper_prompt=query_wrapper_prompt,
80      tokenizer_name=model_name,
81      model_name=model_name,
82      device_map="auto",
83      # uncomment this if using CUDA to reduce memory
84      model_kwargs={"torch_dtype": torch.float16, "tok
85      tokenizer_kwargs={"token":token}, #"load_in_8bit
86  )
```

Log

```
1    /* region: Generated preamble */
79
80   proc python;
81   submit
NOTE: Python initialized.
Python 3.9.12 (main, Apr  5 2022, 06:56:58)
[GCC 7.5.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more
>>>
>>
NOTE: PROCEDURE PYTHON used (Total process time):
      real time           16:05.37
      cpu time            1.08 seconds
```

Code

```
1   proc cas;
2       session CAArgenX;
3       builtins.loadActionSet / actionSet="textRuleDevelo
4       builtins.loadActionSet / actionSet="textRuleScore"
5
6       textRuleScore.applyConcept /
7           casOut      ={caslib="CASUSER", name="out_conc
8           docId="IndexDocNR"
9           factOut     ={caslib="CASUSER", name="out_fact
10          matchType="ALL"          /* matchType="ALL" |
11          model={name="outli"}
12          language="ENGLISH"
13          ruleMatchOut={caslib="CASUSER", name="out_rule
14          table={caslib="PUBLIC" name="PROTOCOL_13FEB202
15          textVar="content";
```

Log Results

```
    /* region: Generated preamble */
9
0   proc cas;
1   session CAArgenX;
2   builtins.loadActionSet / actionSet="textRuleDevelop";
3   builtins.loadActionSet / actionSet="textRuleScore";
4
5   textRuleScore.applyConcept /
6           casOut      ={caslib="CASUSER", name="out_conce
7   docId="IndexDocNR"
96  quit;
NOTE: The PROCEDURE CAS printed page 4.
NOTE: PROCEDURE CAS used (Total process time):
      real time           0.32 seconds
      cpu time            0.02 seconds
```

Comparison of 2 ways for information retrieval / contextual extraction

LITI – rules (SAS VTA)

- It's "Regular Expressions on steroids".
- Knowledge (and effort) required to write linguistic rules with correct syntax.
- Some patterns are hard to come by (because they are complicated, non-standard, with a lot of variety in the language …).
- You need some preliminary knowledge on the topic at hand and on what you want to catch!
- Rules are easily overfit (too specific) if n° of training docs is low.
- Library of LITI-rules can / should grow very big to guarantee a high hit rate.
- Results are "proven", non-debatable (there's a clear match in the text).
- Relatively light in terms of resource usage (and cost).
- page breaks and headers and footers are no gift.
- special and non-printable characters pose no problem.

RAG – LLM (offline model)

- It's generative AI.
- Knowledge required to do proper prompt engineering (designing the user query). But much less effort required. It's just an API-call to a pre-trained model.
- If the context around a particular topic can vary widely from document to document, there is a clear advantage.
- It's generative AI, so some of the info returned can be made up (hallucination).
- Implementing offline models (internalization) and maintaining them is labor-intensive.
- Heavy and intensive in terms of resource usage (even for one or a few documents). Often, several GPUs are needed, and run-time is considerable.
- Scaling (expand the scope) is easier here.
- page breaks and headers and footers are no gift.
- special and non-printable characters pose no problem.

cdisc

Example of challenges LITI/LLM

Legend :

perfect
> 75 % hit rate
between 60 % and 75 % hit rate
< 60% hit rate

		INCLUSION CRITERIA			EXCLUSION CRITERIA		
Documents (clinical trial protocols)	document partitioning (training / validation / test)	Number	Number Captured with LITI	Number Captured with LLM	Number	Number Captured with LITI	Number Captured with LLM
Protocol - 13 Feb 2020.pdf	TRAINING	15	15	11	22	22	5
Protocol_22Jul2020.pdf	TEST	7	5	0	20	8	34
EliLilly_NCT03421379_Diabetes.pdf		10	7	1	26	18	0
Roche_NCT04320615_COVID.pdf		7	4	0	12	5	0
CDISC_Pilot_Study.pdf		8	3	0	23	5	0

cdisc

Overcome challenges (LITI / LLM)

- Pre-processing steps
  - Introduced to support both LITI and LLM
  - **Divide the document into ToC – sections !!**

- Post-processing steps
  - LITI rules - **too specific to scale**

Python packages explored for splitting protocols into chunks

- PyMuPDF==1.23.26 : 
- Table of Contents was extracted in a (structured) *.JSON file.
- JSON file was imported in SAS table … containing the ToC
- Protocol was then split in chunks with SAS-code (***SAS code is automatically and dynamically generated completely driven by the ToC-table***)

cdisc

ToC (Result)

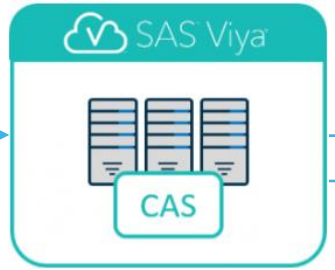
Log | Output Data (1)

TOC — Table rows: 113 | Columns: 3 of 3 | Rows 1 to 113

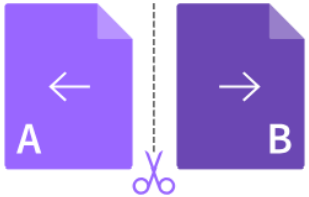
Enter expression

	toc	page	chapter
20	[3, '6.2.1. Primary Endpoint', 19]	19	6.2.1. Primary Endpoint
21	[3, '6.2.2. Secondary Endpoints', 19]	19	6.2.2. Secondary Endpoints
22	[1, '7. Investigational Plan', 20]	20	7. Investigational Plan
23	[2, '7.1. Overall Study Design', 20]	20	7.1. Overall Study Design
24	[2, '7.2. Number of Subjects', 20]	20	7.2. Number of Subjects
25	[2, '7.3. Treatment Assignment', 20]	20	7.3. Treatment Assignment
26	[2, '7.4. Dose Adjustment Criteria', 21]	21	7.4. Dose Adjustment Criteria
27	[2, '7.5. Criteria for Termination of Study', 21]	21	7.5. Criteria for Termination of Study
28	[1, '8. Selection and Withdrawal of subjects', 27]	27	8. Selection and Withdrawal of subjects
29	[2, '8.1. Inclusion Criteria', 27]	27	8.1. Inclusion Criteria
30	[2, '8.2. Exclusion Criteria', 28]	28	8.2. Exclusion Criteria
31	[2, '8.3. Subject Withdrawal Criteria', 30]	30	8.3. Subject Withdrawal Criteria
32	[1, '9. Treatment of Subjects', 31]	31	9. Treatment of Subjects
33	[2, '9.1. Description of IMP', 31]	31	9.1. Description of IMP
34	[2, '9.2. Restrictions', 31]	31	9.2. Restrictions
35	[3, '9.2.1. Concomitant Medication/Procedure(s)', 31]	31	9.2.1. Concomitant Medication/Procedure(s)
36	[3, '9.2.2. Alcohol', 32]	32	9.2.2. Alcohol
37	[3, '9.2.3. Physical Activities', 32]	32	9.2.3. Physical Activities
38	[3, '9.2.4. Dietary Aspects', 32]	32	9.2.4. Dietary Aspects
39	[3, '9.2.5. Smoking', 32]	32	9.2.5. Smoking

Process diagram for contextual extraction with LITI rules



CAS table
VARCHAR column



Split documents
- ~~sentence splitter~~
- ~~paragraph splitter~~
- section splitter
(driven by ToC)

LITI-rules applied to documents as a whole

LITI-rules applied to <u>selected</u> split sections

Result after pre-/processing

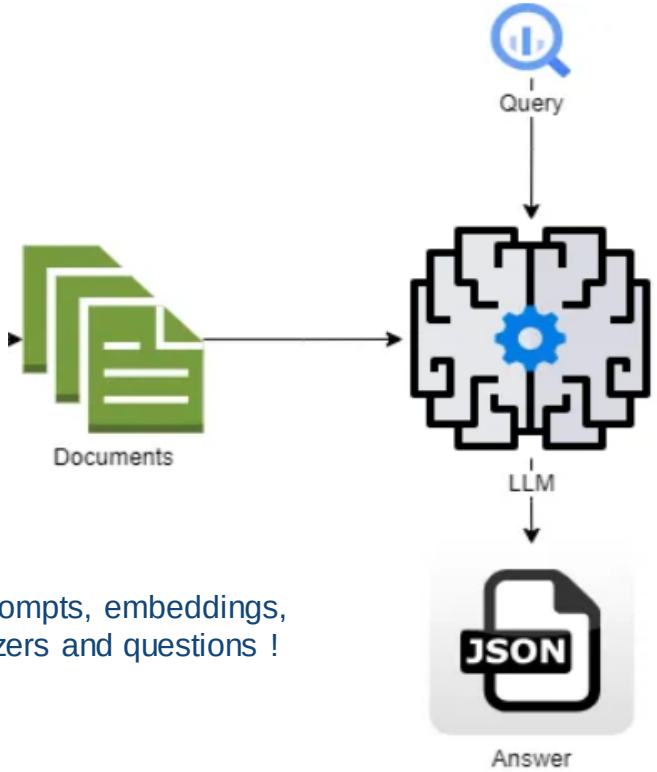
Legend :

perfect
> 75 % hit rate
between 60 % and 75 % hit rate
< 60% hit rate

		INCLUSION CRITERIA			EXCLUSION CRITERIA		
Documents (clinical trial protocols)	document partitioning (training / validation / test)	Number	Number Captured with Base SAS	Number Captured with LLM	Number	Number Captured with Base SAS	Number Captured with LLM
Protocol - 13 Feb 2020.pdf	TRAINING	15	15	12*	22	22	19
Protocol_22Jul2020.pdf	TEST	7	7	6	20	20	34*
EliLilly_NCT03421379_Diabetes.pdf		10	10	10	26	26	26
Roche_NCT04320615_COVID.pdf		7	7	7	12	12	12
CDISC_Pilot_Study.pdf		8	8	8	23	23	23

* - double entries and missing entries

Process diagram for contextual extraction with RAG-LLM



RAG Pipeline

User query
Prompt Engineering

LLAMA-V2
7B model

Source: Information extraction with LLM | Chetan Khadke | Medium | Medium

Variations in prompts, embeddings, models, tokenizers and questions !

GPU power required for LLMs

```
Every 5.0s: nvidia-smi

Thu Mar 21 11:38:28 2024
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.54.03              Driver Version: 535.54.03    CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  Tesla T4                       On  | 00000001:00:00.0 Off |                  Off |
| N/A   34C    P0              27W /  70W |   4401MiB / 16384MiB |     31%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   1  Tesla T4                       On  | 00000002:00:00.0 Off |                  Off |
| N/A   34C    P0              38W /  70W |   4117MiB / 16384MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   2  Tesla T4                       On  | 00000003:00:00.0 Off |                  Off |
| N/A   34C    P0              27W /  70W |   4117MiB / 16384MiB |      9%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   3  Tesla T4                       On  | 00000004:00:00.0 Off |                  Off |
| N/A   35C    P0              32W /  70W |   3651MiB / 16384MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=========================================================================================|
+-----------------------------------------------------------------------------------------+
```

Size	vCPU	Memory: GiB	Temp storage (SSD) GiB	GPU	GPU memory: GiB	Max data disks	Max NICs / Expected network bandwidth (Mbps)
Standard_NC4as_T4_v3	4	28	180	1	16	8	2 / 8000
Standard_NC8as_T4_v3	8	56	352	1	16	16	4 / 8000
Standard_NC16as_T4_v3	16	110	352	1	16	32	8 / 8000
Standard_NC64as_T4_v3	64	440	2880	4	64	32	8 / 32000

Python packages explored for table and image extraction

- tabula-py==2.9.0 : not good enough to capture SoA
- camelot-py==0.11.0 : not good enough to capture SoA
- pdfminer.six==20231228 : not held back
- PyMuPDF==1.23.26 : 👍

- PyPDF2 (for image extraction) was not tested
- PaddleOCR currently under exploration

Note: The AZURE service «Optical Character Recognition (OCR) - Azure AI Document Intelligence Table Extraction» was not (yet) tested.

cdisc

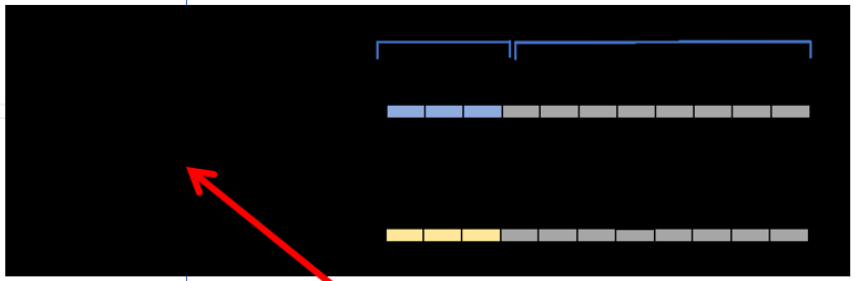
Agenda

1. Background to the project
2. The Journey
3. **The hurdles**
4. The result

Image extraction

find graphs.py > ...

```python
import fitz
import io
import os
from PIL import Image

file = "C:\\Users\\SNLREF\\OneDrive - SAS\\klanten\\[redacted]\\Protocol - Protocol - 13 Feb 2020.pdf"

pdf_file = fitz.open(file)

output_dir = "C:\\temp\\conda"
# Desired output image format
output_format = "bmp"

if not os.path.exists(output_dir):
    os.makedirs(output_dir)

for page_index in range(len(pdf_file)):

    # get the page itself
    page = pdf_file[page_index]
    image_list = page.get_images()

    # printing number of images found in this page
    if image_list:
        print(
            f"[+] Found a total of {len(image_list)} images in page {page_index}")
    else:
        print("[!] No images found on page", page_index)

    for image_index, img in enumerate(page.get_images(), start=1):

        # get the XREF of the image
        xref = img[0]
```



Caused by a mask layer

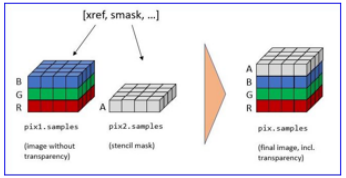
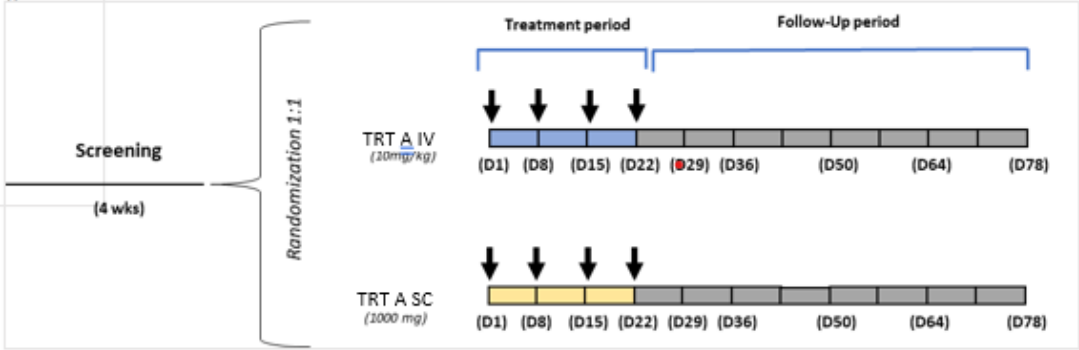


Image extraction

```python
def recoverpix(doc, item):
    xref = item[0]  # xref of PDF image
    smask = item[1]  # xref of its /SMask

    # special case: /SMask or /Mask exists
    if smask > 0:
        pix0 = fitz.Pixmap(doc.extract_image(xref)["image"])
        if pix0.alpha:  # catch irregular situation
            pix0 = fitz.Pixmap(pix0, 0)  # remove alpha channel
        mask = fitz.Pixmap(doc.extract_image(smask)["image"])

        try:
            pix = fitz.Pixmap(pix0, mask)
        except:  # fallback to original base image in case of problems
            pix = fitz.Pixmap(doc.extract_image(xref)["image"])

        if pix0.n > 3:
            ext = "pam"
        else:
            ext = "png"

        return {  # create dictionary expected by caller
            "ext": ext,
            "colorspace": pix.colorspace.n,
            "image": pix.tobytes(ext),
        }
```



Imaginary, but not really imaginary data

```
llama
{'What are the inclusi
The inclusion criteria
1. Adults (18 years of
2. A body surface
4. A BSA involveme
5. A BSA involveme
6. A BSA involveme
7. A BSA involveme
8. A BSA involveme
9. A BSA involveme
10. A BSA involvem
11. A BSA involvem
12. A BSA involvem
13. A BSA involvem
14. A BSA involvem
>>
```

```
e5base
{'What are the inc
The inclusion crit
1. Subjects must b
2. Subjects must h
3. Subjects must h
4. Subjects must h
5. Subjects must h
6. Subjects must h
7. Subjects must h
8. Subjects must h
9. Subjects must h
10. Subjects must
11. Subjects must
12. Subjects must
13. Subjects must
14. Subjects must
15. Subjects must
16. Subjects must
17. Subjects must
18. Subjects must
19. Subjects must
20. Subjects must
```

```
{'What are the inclusion criteria?': '
The inclusion criteria are:

1. Subjects must be ≥18 years of age.
2. Subjects must have a diagnosis of moderate to severe active RA.
```

```
{'What is the Primary Objective?': '
Primary Objective:
### Explanation:
The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
### Hints:
1. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
2. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
3. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
4. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
5. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
6. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
7. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
8. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
9. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
10. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
11. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
12. The Primary Objective is to evaluate the safety and efficacy of efgartigimod in patients with generalized myasthenia gravis (gMG).
13. The Primary Objective is'}
```

```
{'What are the Secondary Endpoints?': 'Secondary Endpoints:
1. 12-week change in the number of active ulcers (defined as ulcers with a baseline diameter of ≥ 5 mm)
2. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm
3. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the foot
4. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the ankle
5. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the toe
6. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the heel
7. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the shin
8. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the calf
9. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the thigh
10. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the knee
11. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the leg
12. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the hip
13. 12-week change in the number of ulcers with a baseline diameter of ≥ 5 mm and a baseline ulcer location of the buttock
14. 12-week change in the number of ulcers with a baseline diameter of ≥'}
>>>
```

```
protein
protein
protein
protein
```

Table Issues

20. APPENDICES

APPENDIX 1: SAFETY LABORATORY ASSESSMENTS

Table 6: Screening Laboratory Assessments

Laboratory Assessments	Parameters
Tests completed at screening only	• Virology: HIV test, hepatitis B surface antigen, total hepatitis B core antibody, hepatitis B surface antibody, hepatitis C virus antibody (see Appendix 2) • FSH in post-menopausal women (see Appendix 5) • Total IgG

FSH=follicle-stimulating hormone; HIV=human immunodeficiency virus

Table 7: Hematology, Clinical Chemistry, Urinalysis

Laboratory Assessments	Parameters				
Hematology	Platelet Count Red blood cell (RBC) Count Hemoglobin Hematocrit		RBC Indices: • Mean corpuscular volume (MCV) • Mean corpuscular hemoglobin (MCH) • Mean corpuscular hemoglobin concentration (MCHC) • %reticulocytes	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils	
Clinical Chemistry[a]	Blood urea nitrogen (BUN) Creatinine Glucose (fasting) Calcium	Potassium Sodium Alkaline phosphatase Lactate dehydrogenase	Aspartate aminotransferase (AST) Alanine aminotransferase (ALT) Gamma glutamyl transferase (GGT)	Total and direct bilirubin Total Protein Albumin	Cholesterol (total) Low-density lipoprotein (LDL) High-density lipoprotein (HDL) Triglycerides
	creatine kinase (CK)[b]; CK myocardial band (CKMB)				
	International Normalized Ratio				
Routine Urinalysis	• Specific gravity				

PAGE_56_TABLE_1

Obs	Col0	Laboratory	Col2	Parameters
1		Assessments		
2	Tests completed at screening only			• Virology: HIV test, hepatitis B surface antigen, total hepatitis B core antibody, hepatitis B surface antibody, hepatitis C virus antibody (see Appendix 2) • FSH in post-menopausal women (see Appendix 5) • Total IgG

Obs	col0	Laboratory	col2	Parameters	Parameters_1	Parameters_2	Parameters_3	Parameters_4	Parameters_5	Parameters_6
1	Parameters	Assessments	Assessments	Assessments	Assessments	Assessments	Assessments	Assessments	Assessments	Assessments
2	Hematology	Hematology	Hematology	Platelet Count	Platelet Count	RBC Indices: • Mean corpuscular volume (MCV) • Mean corpuscular hemoglobin (MCH) • Mean corpuscular hemoglobin concentration (MCHC) • %reticulocytes	RBC Indices: • Mean corpuscular volume (MCV) • Mean corpuscular hemoglobin (MCH) • Mean corpuscular hemoglobin concentration (MCHC) • %reticulocytes	RBC Indices: • Mean corpuscular volume (MCV) • Mean corpuscular hemoglobin (MCH) • Mean corpuscular hemoglobin concentration (MCHC) • %reticulocytes	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils
3	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils	White blood cell (WBC) count with differential (% and absolute numbers): • Neutrophils • Lymphocytes • Monocytes • Eosinophils • Basophils	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count
4	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Red blood cell (RBC) Count	Hemoglobin	Hemoglobin	Hemoglobin	Hemoglobin	Hemoglobin	Hemoglobin	Hemoglobin
5	Hemoglobin	Hemoglobin	Hemoglobin	Hematocrit	Hematocrit	Hematocrit	Hematocrit	Hematocrit	Hematocrit	Hematocrit
6	Clinical Chemistrya	Clinical Chemistrya	Clinical Chemistrya	Blood urea nitrogen (BUN)	Potassium	Potassium	Aspartate aminotransferase (AST) Alanine aminotransferase (ALT) Gamma glutamyl transferase (GGT)	Total and direct bilirubin	Total and direct bilirubin	Cholesterol (total) Low-density lipoprotein (LDL) High-density lipoprotein (HDL) Triglycerides
7	Cholesterol	Cholesterol	Cholesterol	Creatinine	Sodium	Sodium	Sodium	Total Protein	Total Protein	Total Protein

Realizing USDM is more than a protocol template model

name	description	label	type
Screening	Screening Epoch	Screening	SCREENING
Baseline	Baseline Epoch	Baseline	BASELINE
Treatment	Treatment Epoch	Treatment	TREATMENT
Follow-Up	Follow-up Epoch	Follow-Up	FOLLOW-UP

study | studyAmendments | studyIdentifiers | studyDesign | studyDesignSites | studyDesignArms | studyDesignEpochs | study

Trial Design tables

eTMF

And Realizing that a process is still needed together with USDM

	A	B	C	D	E	F
1	name	SCOPE1				
2	studyTitle	Simple Test 1				
3	studyVersion	1				
4	studyType	Interventional Study				
5	studyPhase	C15602				
6	studyAcronym	SIMPLE				
7	studyRationale	A simple test				
8	businessTherapeuticAreas	SPONSOR: VAC=Vacines Group, SPONSOR: REG=Regulatory				
9	briefTitle	Something Brief				
10	officialTitle	Something Very Official				
11	publicTitle	Something Public				
12	scientificTitle	Somethign Clever But New				
13	protocolVersion	1				
14	protocolStatus	draft				
15						
16	category	name	description	label	type	date
17	study_version	Design Approval	Design approval date	Design Approval	Sponsor Approval Date	16-12-2022
18	protocol_document	Protocol Approval	Protocol document approval date	Protocol Approval	Protocol Effective Date	1-1-2023

study | studyAmendments | studyIdentifiers | studyDesign | studyDesignSites | studyDesignArms | studyDesignEpochs | studyDesignEligibilityCriteria

Clinical Study Protocol

SIGNATURE PAGE

Sponsor's Approval

The protocol has been approved by argenx.

Responsible Medical Officer:

Sponsor's Authorized Officer:

13-02-2020

Date

Clinical Study Protocol

INVESTIGATOR'S AGREEMENT

I have received and read the investigator's brochure for efgartigimod. I have read the protocol and agree to conduct the study as outlined. I agree to maintain the confidentiality of all information received or developed in connection with this protocol.

Printed Name of Investigator

Signature of Investigator

Date

?

13 Feb 2020 Confidential 2

Agenda

1. Background to the project
2. The Journey
3. The hurdles
4. **The result**

End Result (SoA)

Clinical Study Protocol

Table 3: Schedule of Assessments

Study Period	Screening[a]	-1	Treatment Pe				
Study Day	-28 to -2		1	2	7	8	1-
Informed consent	X						
In-/exclusion criteria	X	X	X[c]				
Virology screen	X						
Medical history	X						
Demographic data	X						
Pregnancy test[d]		X					
Urine drug screen & alcohol urine test[e]	X	X					
Height	X						
Physical examination	X	X					
Weight[f] (+ BMI)[g]	X[f,g]	X[f]			X[f]		X
Vital signs[h]	X	X	X[h]			X[h]	
Triplicate 12-lead ECG[i]	X	X					
Urinalysis[e]	X	X					
Clinical laboratory tests[j]	X	X				X[c]	
Blood sampling:							
• PK[l]		X[k]	X[l]			X[l]	
• PD[m]		X[k]	X[l]			X[l]	
• ADA[n]		X[k]	X[c]				
Randomization			X[c]				
IMP administration[o]			X			X	
Ambulant visits	X						

PAGE_23_TABLE_1

Enter expression

	Study Period	Screen inga	-1	Treatment Period	Treatment Period_1
1	Study Day	-28 to -2	-28 to -2	1	2
2	Informed consent	X			
3	In-/exclusion criteria	X	X	Xc	
4	Virology screen	X			
5	Medical history	X			
6	Demographic data	X			
7	Pregnancy testd		X		
8	Urine drug screen & alcohol urine teste	X	X		
9	Height	X			
10	Physical examination	X	X		
11	Weightf (+ BMI)g	Xf,g	Xf		
12	Vital signsh	X	X	Xh	
13	Triplicate 12-lead ECGi	X	X		
14	Urinalysise	X	X		
15	Clinical laboratory testsj	X	X		
16	Blood sampling: • PKl • PDm • ADAn		Xk Xk Xk	Xl Xl Xc	
17	Randomization			Xc	
18	IMP administrationo			X	
19	Ambulant visits	X			

From USDM to Protocol

File Home Insert Layout References Review View Help SAS (engage)

Comments Catch up Editing Share

Home Learn More Help

SAS (engage) Help

Glucagon (LY900018)

Eli Lilly Japan K.K

Japan

26 October 2017

6. STUDY OBJECTIVES AND ENDPOINTS

6.1. Objectives

6.1.1. Primary Objective

•

6.1.2. Secondary Objectives

•

Fact Rule = '1.b. ObjectivesSecondary'

Match Text	Fact Rule
* To compare the safety and tolerability of 3 mg LY900018 with 1 mg IMG * To characterize the PK profile of 3 mg LY900018 compared to 1 mg IMG * To characterize the PD profile of 3 mg LY900018 compared to 1 mg IMG	1.b. ObjectivesSecondary

SAS

Home Reports Results

Eli_Lilly_CDISC_Protocol_ObjectivesEndpoints

1.a. ObjectivesPrimary | 1.b. ObjectivesSecondary | 1.c. ObjectivesExploratory | 2.a. En

Match Text	Fact Rule
To demonstrate that 3 mg LY900018 is non-inferior to 1 mg IMG for the proportion of patients achieving treatment success from insulin-induced hypoglycemia using a non-inferiority margin of 10%	1.a. ObjectivesPrimar
* To compare the safety and tolerability of 3 mg LY900018 with 1 mg IMG * To characterize the PK profile of 3 mg LY900018 compared to 1 mg IMG * To characterize the PD profile of 3 mg LY900018 compared to 1 mg IMG	1.b. ObjectivesSecon
* Explore the formation of anti-glucagon antibodies to glucagon * To evaluate the recovery from clinical symptoms of hypoglycemia	1.c. ObjectivesExplor
The proportion of patients achieving treatment success defined 0 mg/dL or an increase of >20 utes after administration of s the minimum PG value at the	2.a. EndpointsPrimar

Eli_Lilly_CDISC_Protocol_ObjectivesEndpoints has been loaded.

cdisc

Conclusion

- From human to machine – challenging, but not impossible

Next steps

- **Short term** (Q2/Q3)
  - Leverage Open AI service (commercial models) instead of LLAMA-V2 7B model as prep finalization internal business case.

- **Mid term** (End 2024 – begin 2025)
  - focus on CTPS (concept sheet)

- **Longer term - protocol builds**
  - Cost
  - Technology (no off-the shelf available yet?), we are biotech … not software builders
  - Change management
  - New roles

Interested in cont’d progress ?

SAS Innovate, Rotterdam,

Tuesday June 11, 2024

PHUSE EU Connect, Strasbourg (if accepted ☺)

10-13 Nov, 2024

Thank You!