



Investigator's Brochure content automation

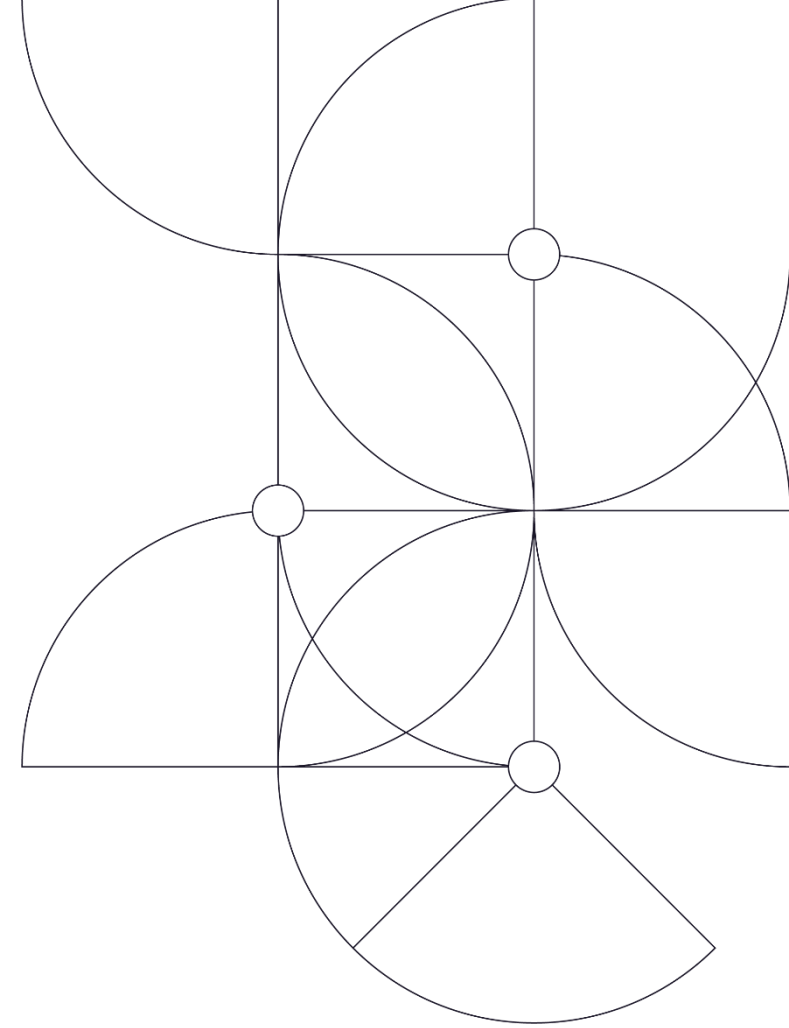
Prepared for PHUSE US Connect 2023

March 6, 2023

Debiprasad Roy, Exec. Dir & Head of R&D Technology, Data Science and
Digital Strategy. Allogene Therapeutics

Kumail Fazal, Associate Principal, ZS Associates India Pvt. Ltd.

Impact where it matters.



Disclaimer

This presentation is a collaboration project between Allogene and ZS Associates to address medical writing automation challenges.

Allogene does not promote services and technology provided by ZS Associates.

For a medical writer, authoring a new document requires substantial amount of manual effort and decision making

1

Pick the correct template

- **Larger biotech:** Which template should we use?
- **Emerging biotech:** We have no templates. Where should we begin?

4

Create new content

- What summaries do I need to create?
- Are there any sections that I need to write from scratch (de-novo content)?



2

Locate the tables, listings and figures

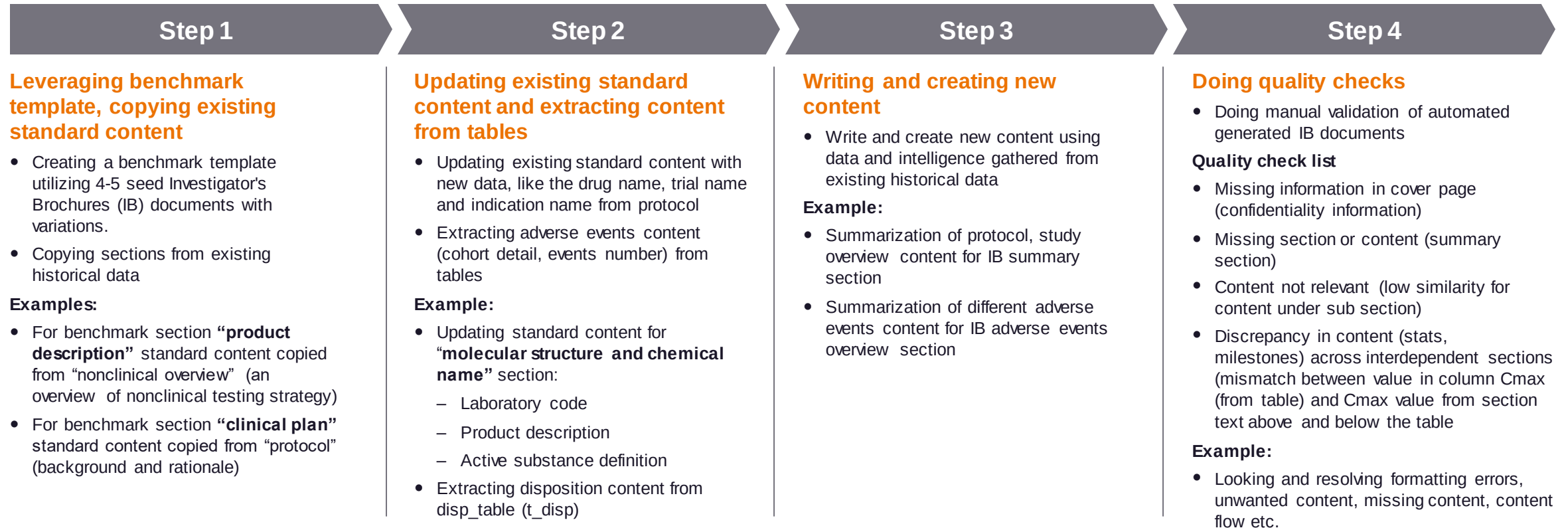
- Where can I find the data?
- Which tables, listings and figures are relevant?
- Which data sets will require processing?

3

Reuse and adjust information from already created content

- Is there any preexisting content that can be leveraged directly?
- Which sections are generic and can be reused?

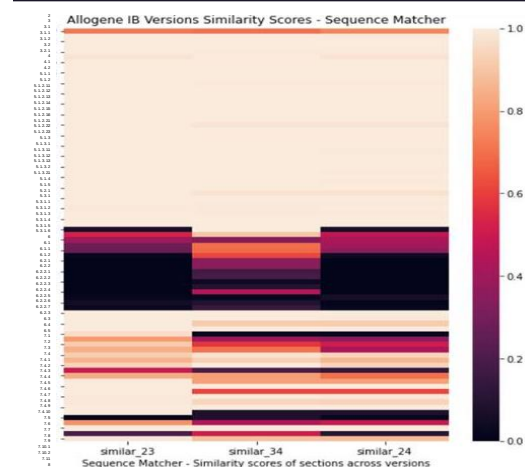
For any pharma organization, the multi-step end-to-end process of authoring IB document manually takes 3-4 weeks



ZS collaborated with Allogene to automate these tasks and gain efficiencies in authoring an IB, given it is a yearly process and is highly manual and time consuming.

Advanced AI and ML techniques in combination with content mapping help lay a solid foundation for content automation

Sections



Our Methodology to assess overlap:

- Algorithm Method – Preprocessed content in all sections and compared each section across different versions to calculate similarity scores

*Darker color - Lesser overlap
Lighter color - More overlap

Legend:

- similar_23 (comparison of version 2 & version 3)
- similar_34 (comparison of version 3 & version 4)
- similar_24 (comparison of version 2 & version 4)

Sections with high content overlap (similarity scores are higher than 90%):

- IB Sections: Summary, Introduction, Formulation and Nonclinical Studies

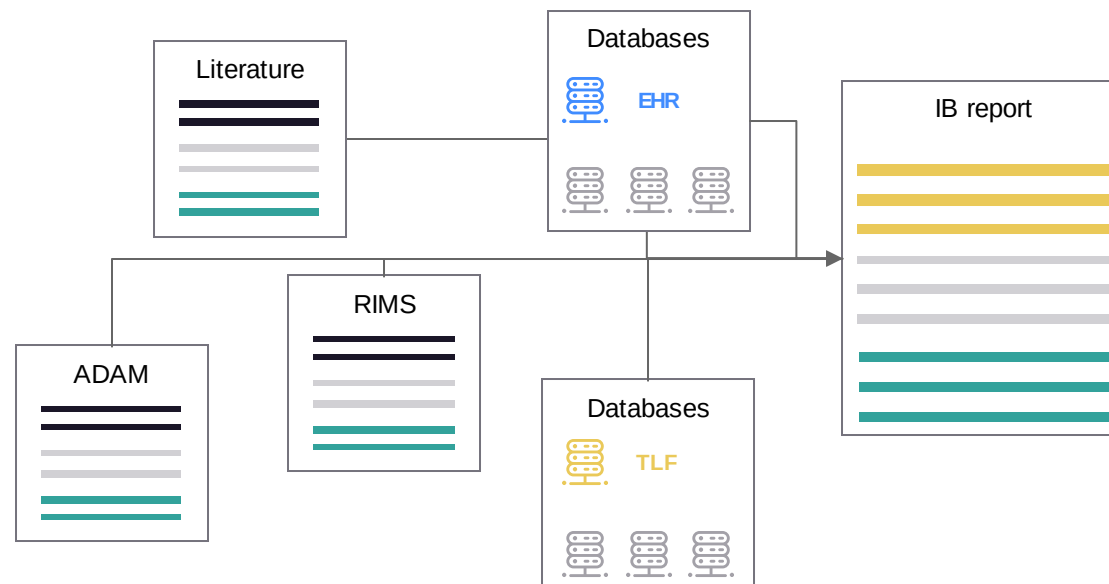
Sections with moderate content overlap (similarity scores between 60%-90%):

- IB sections: Exposure, Adverse Events

Sections with low content overlap (similarity scores are lower than 60%):

- IB Sections: Effects in Humans (AEs, Deaths, Lab Evaluations, Exposure, etc.), Summary of Data and Guidance for Investigator

SCA + NLP + NLU technologies to lay the foundation



In harmonization, standardization and NLP analytics and data, pipelines become an essential step mapping of content.

Leveraging content overlap, a benchmark template was created, and data ingestion pipelines were built to generate the document automatically, which significantly reduced the manual content authoring effort.

The next steps will be about the level of automation required

Level - 1



Level 1: Copying content from existing document

- **Matching and extraction of section name and its content from parsed data**

Example:

- For the benchmark section “**Product description,**” standard content was copied from “Nonclinical Overview” [overview of nonclinical testing strategy]
- For the benchmark section “**Clinical Plan,**” standard content was copied from “Protocol” [background and rationale]

Level - 2



Level 2a: Updating existing standard content with new data

Level 2b: Extracting content from existing tables

- Keyword extraction using **biomedical NER models**
- Plug-in/Fill blanks with extracted keywords in the standard content
- Read table, use pattern recognition for defining input/output
- Fetched preferred terms along with values from defined input/output columns
- Generated new content by combining standard text and fetched table value

Example:

- Updating standard content for “**Molecular Structure and Chemical Name**” section:
 - Laboratory Code
 - Product Description
 - Active Substance Definition
- Extracting disposition content from disp_table (t_disp)

Level - 3



Level 3: Write and create new content using data and intelligence gathered from existing historical data

- Text summarization using **GPT-2 model** using 60% ratio of summarized text to original text
- 1. **Create content from other documents (“Protocol” and “Nonclinical”):** Extractive summarization to fetch relevant, highly ranked sentences, with an aim to reduce the content length fetched from the data source.
- 2. **Create content from IB sections:** Extractive summarization to fetch relevant, highly ranked sentences, with an aim to reduce the content length fetched by joining the relevant sections content from IB.

Example:

- Summarization of “Protocol,” “Study Overview” content for IB summary section
- Summarization of different adverse events content for IB adverse events overview section



NLP / NLU/ NLG-augmented techniques, SCA-based document generation

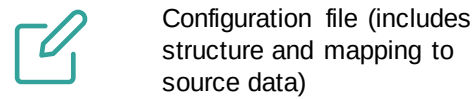
Our approach to automate the IB document



Data ingestion



Source data + other source documents



Configuration file (includes structure and mapping to source data)



Parse documents to extract relevant information / section localization



Preprocessing



Text cleaning (lower casing, removing unwanted characteristics, etc.)



Tokenization



Structured data with cleaned text



Content generation



Keywords extraction using NER models



Text summarization



Level 1: Content from existing document



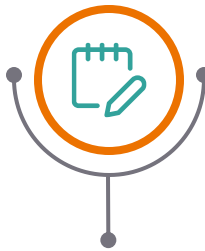
Level 2: Content from existing table



Level 3: Generate content

NLP / NLU Techniques

NLG-augmented, SCA-based document generation



Post processing



Size formatting for headers and content



Table of Content generation



Validation based on Benchmark document

Frontend



Impact generated



50%-60%

Reduction in
manual effort



~75%

Sections automated of
benchmark document



~99%

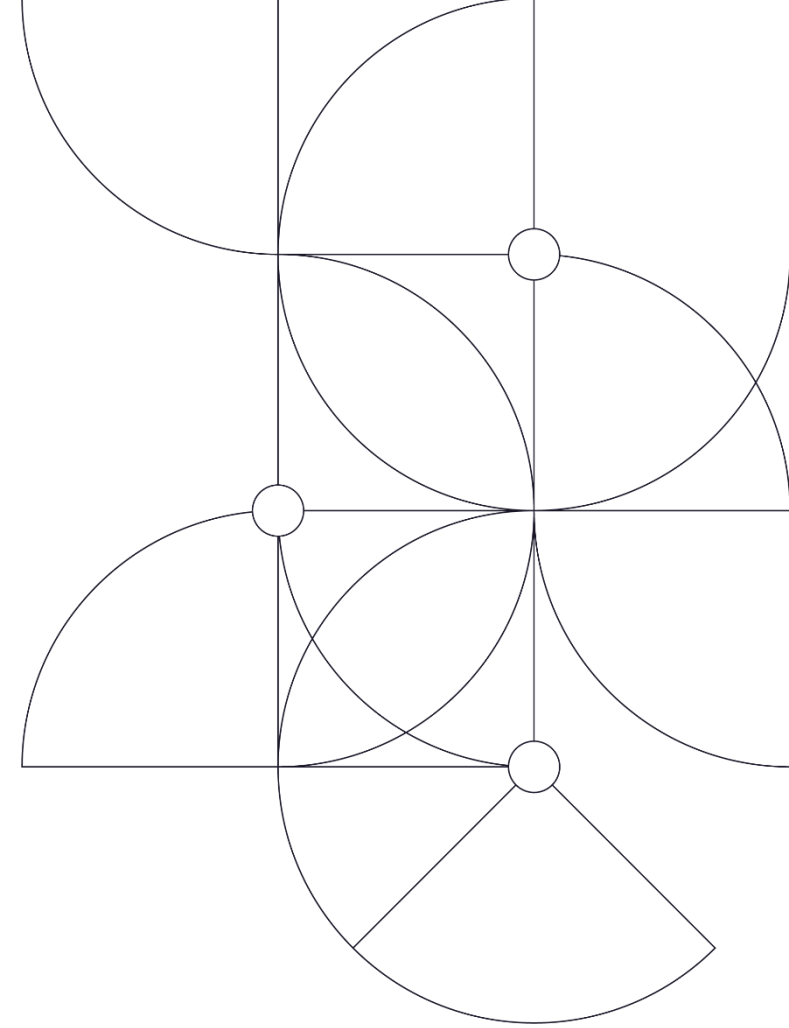
Accuracy achieved in
automating the sections

We don't want to stop here. We look to automate CSRs and protocols in the next phase.



Thank you

Impact where it matters.



IB document automated generation: Demo

Content Automation Tool

Enter study parameters

Molecule Name: SPNS-123 ▼ Study Name: Type Name Of The Study

Phase: I ▼

Indication: R&R Multiple myeloma ▼

Version: 1 ▼

Language: EN ▼

Country: United States ▼

Format: DOC ▼

Config File: Enter Config File Path

📁 Upload Config (0)

Document(s) to be created

☐ Investigator's Brochure

☐ Clinical Study Report

☐ eCTD modules

Data sources

☐ Protocol

☐ Non-clinical

☐ CMC report

☐ TLFs

Generate IB Document

Show Document

Download Document



Level 1 examples

Data source: Protocol (Summary of Study Design)

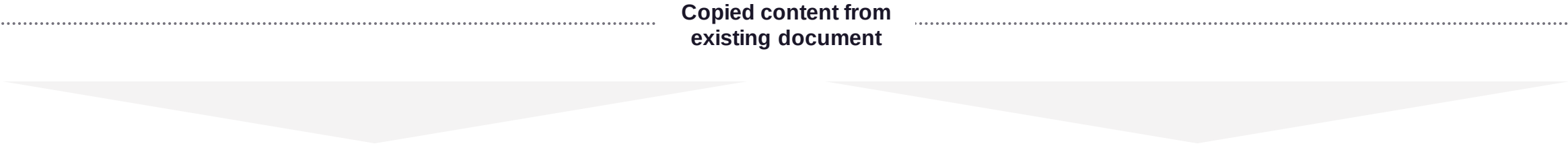
Study 123-PQRS (PQRS) is a multicenter, randomized, double-blind, Phase 3 trial for women with HR+, HER2- locally advanced or metastatic breast cancer. Figure PQRS.8.1 illustrates the study design. Approximately 630 EP patients will be randomized 2:1 between the 2 arms:

Experimental Arm A: DRUG XX 150 mg orally Q12H on Days 1 to 28 of a 28-day cycle plus DRUG XY 500 mg intramuscularly on Days 1 and 15 of Cycle 1, then on Day 1 of Cycle 2 and beyond Control (Placebo) Arm B: Placebo orally Q12H on Days 1 to 28 of a 28-day cycle plus DRUG XY 500 mg intramuscularly on Days 1 and 15 of Cycle 1, then on Day 1 of Cycle 2 and beyond.

IB content: Summary section

Study 123-PQRS (PQRS) is a multicenter, randomized, double-blind, Phase 3 trial for women with HR+, HER2- locally advanced or metastatic breast cancer. Figure PQRS.8.1 illustrates the study design. Approximately 630 EP patients will be randomized 2:1 between the 2 arms:

Experimental Arm A: DRUG XX 150 mg orally Q12H on Days 1 to 28 of a 28-day cycle plus DRUG YY 500 mg intramuscularly on Days 1 and 15 of Cycle 1, then on Day 1 of Cycle 2 and beyond Control (Placebo) Arm B: Placebo orally Q12H on Days 1 to 28 of a 28-day cycle plus DRUG XY 500 mg intramuscularly on Days 1 and 15 of Cycle 1, then on Day 1 of Cycle 2 and beyond.



Data source: Protocol (Introduction)

DRUG XX is a potent and selective small molecule inhibitor of CDK4 and CDK6. In the Phase 1 Study 123-PQRS (PQRS), DRUG XX has shown acceptable safety across five tumor-specific cohorts, with the most common treatment-emergent adverse events (TEAEs) possibly related to study drug including diarrhea, nausea, fatigue, vomiting, and neutropenia.

IB content: Product description

DRUG XX is a potent and selective small molecule inhibitor of CDK4 and CDK6. In the Phase 1 Study 123-PQRS (PQRS), DRUG XX has shown acceptable safety across five tumor-specific cohorts, with the most common treatment-emergent adverse events (TEAEs) possibly related to study drug including diarrhea, nausea, fatigue, vomiting, and neutropenia.

Level 2 examples

Example 1: Keyword extraction for updating existing standard content with new data

Extracted keyword	Examples	Data source section and subsection names
Indication name	Advanced or Metastatic clear cell renal cell carcinoma	Protocol Cover page (Trial heading)
Trial name	Protocol no.- SPNS-123-101	Protocol Cover page (Trial heading)
Product description	CD70CAR+_TCR αβ-_CD52+/-	M3 document Structure
Active substance	CAR+, TCRαβ-viable cells	M3 document Structure

Example 2: Extraction of serious adverse events content from table

	Number (%) of patients			
Preferred term	10mg ABC (N=5)	20mg PQR (N=10)	30mg XYZ (N=5)	All patients (N=20)
No of patients with any serious TEAE	4 (80%)	8 (80%)	3 (60%)	15 (75%)
Cystitis	0 (0.0%)	2 (20%)	1 (20%)	5 (25%)
Pemphigoid	1 (20%)	1 (10%)	0 (0.0%)	3 (15%)
Aplastic anaemia	0 (0.0%)	3 (30%)	0 (0.0%)	2 (10%)
Hypokalemia	2 (40%)	1 (10%)	1 (20%)	1 (15%)
Neutropenia	1 (20%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Hypertension	0 (0.0%)	1 (10%)	1 (20%)	4 (20%)

Extraction is done using
biomedical NER model

Overall, Out of 20 subjects, 15 (75%) subjects have experienced a serious adverse event. Serious adverse events reported as related to SPNS-123 by either the investigator or the sponsor included, Cystitis with 5 (25%) patients, Pemphigoid with 3 (15%) patients, Aplastic anaemia with 2 (10%) patients, Hypokalemia with 1 (15%) patients, Neutropenia with 0 (0.0%) patients, , Hypertension with 4 (20%) patients . No deaths have been reported.

Summarization of text



1- Create content from other documents (“Protocol” and “Nonclinical”)

Raw Study Design text from Protocol

Study XXX is a multicenter, randomized, double-blind, Phase 3 trial for women with YYY locally advanced or metastatic breast cancer. Figure JPBL.8.1 illustrates the study design.

Enrollment will close when approximately 450 EP patients have been randomized at a starting dose of N mg Q12H. Patients will be randomized using the following stratification factors: nature of disease (visceral metastases versus bone only metastases versus other) and sensitivity to endocrine therapy (primary resistance versus secondary resistance). The presence of visceral metastases refers to lung, liver, pleural, or peritoneal involvement at the time of randomization. Primary clinical resistance to endocrine therapy is defined as follows: 1) for endocrine therapy in the adjuvant setting, recurrence within the first 2 years of adjuvant endocrine therapy while on endocrine therapy or 2) for endocrine therapy in the locally advanced or metastatic setting, progression within first 6 months of initiating first-line endocrine therapy while on endocrine therapy. Patients receiving prior endocrine therapy who do not meet the definition of primary clinical resistance will be considered to have secondary clinical resistance.

Summarization is done using GPT-2 model using bert-extractive-summarizer with 60% threshold

Summarized text (going into IB document)

Study XXX is a multicenter, randomized, double-blind, Phase 3 trial for women with YYY locally advanced or metastatic breast cancer. The presence of visceral metastases refers to lung, liver, pleural, or peritoneal involvement at the time of randomization. Primary clinical resistance to endocrine therapy is defined as follows: 1) for endocrine therapy in the adjuvant setting, recurrence within the first 2 years of adjuvant endocrine therapy while on endocrine therapy or 2) for endocrine therapy in the locally advanced or metastatic setting, progression within first 6 months of initiating first-line endocrine therapy while on endocrine therapy. Patients receiving prior endocrine therapy who do not meet the definition of primary clinical resistance will be considered to have secondary clinical resistance.



1- Create content from other documents (“Protocol” and “Nonclinical”)

Different adverse events subsection content from IB

Most Common TEAEs

Overall, the most commonly reported TEAEs following initial treatment were General disorders and administration site conditions with 9 (100.0%) subjects, Fatigue with 8 (88.9%) subjects, Pyrexia with 4 (44.4%) subjects, Investigations with 8 (88.9%) subjects, C-reactive protein increased with 3 (33.3%) subjects, Neutrophil count decreased with 3 (33.3%) subjects, White blood cell count decreased with 3 (33.3%) subjects, Blood and lymphatic system disorders with 5 (55.6%) subjects, Anaemia with 5 (55.6%) subjects, Neutropenia with 3 (33.3%) subjects.

TEAEs by Severity (Grade ≥ 3)

Overall, 7/9 patients have experienced at least 1 Grade ≥ 3 TEAE. The most commonly reported Grade ≥ 3 TEAEs were blood and lymphatic system disorders (4/9 patients), neutropenia (3/9 patients), anaemia (2/9 patients), leukopenia (1/9 patients) and lymphopenia (1/9 patients).

TEAEs Related to DRUG NAME

The most commonly reported TEAEs related to DRUG NAME (all grades) were: Neutropenia with 2 (50.0%) subjects, Anaemia with 2 (50.0%) subjects, Fatigue with 2 (50.0%) subjects, Lymphopenia with 1 (25.0%) subjects, Platelet count decreased with 1 (25.0%) subjects.

Summarized text going into adverse event overview section in IB

The most commonly reported TEAEs related to DRUG NAME (all grades) were: Neutropenia with 2 (50.0%) subjects, Anaemia with 2 (50.0%) subjects, Fatigue with 2 (50.0%) subjects, Lymphopenia with 1 (25.0%) subjects, Platelet count decreased with 1 (25.0%) subjects. Overall, Out of 9 subjects, 1 (11.1%) subjects have experienced a serious adverse event. Serious adverse events reported as related to DRUG NAME by either the investigator or the sponsor included, General disorders and administration site conditions with 9 (100.0%) patients, Pyrexia with 4 (44.4%) patients, Infections and infestations with 4 (44.4%) patients. Serious TEAS reported were Cytomegalovirus infection with 1 (11.1%) patients. No deaths have been reported as a serious TEAS.