Service Definition Document: Clinical Trial Risk Tool (CTRT)

Attribute	Detail
Service Owner	Fast Data Science (FDS)
Service Name	Clinical Trial Risk Tool (CTRT)
Version	2.0
Review Date	January 2026
Author	Thomas Wood, Director, thomas@fastdatascience.com
Tool accessible at	https://clinicaltrialrisk.org/
Fast Data Science	https://fastdatascience.com/

1. Service Overview

The Clinical Trial Risk Tool (CTRT) is an Artificial Intelligence (AI) and Natural Language Processing (NLP) powered application designed to automate and standardise the risk assessment of clinical trial protocols. Its primary function is to analyse large, unstructured PDF protocol documents (up to 200 pages) and create itemised clinical trial site budgets (per patient budgets), predict costs, and quantify the risk of a trial concluding without delivering statistically significant or informative results (i.e., the risk of uninformativeness). The service augments human review processes by providing rapid, objective, and data-driven insights into critical design flaws and potential cost drivers.

2. Service Objectives

The CTRT service aims to achieve the following core objectives:

1. **Clinical Trial Budgeting:** Create a clinical trial budget in Excel format from the schedule of events.
2. **Risk Quantification:** Provide an objective, numerical risk score (0-100) and a categorical risk rating (Low, Medium, High) for trial uninformativeness.
3. **Cost Prediction:** Accurately predict the likely total cost of a clinical trial in dollars based on extracted trial parameters and historical trial data.

3. Key Features and Functionality

The CTRT operates through the automated extraction and analysis of specific features from the uploaded protocol document:

3.1. Protocol Ingestion & Feature Extraction (Input)

- **Document Upload:** Accepts clinical trial protocols in PDF format.
- **NLP Engine:** Uses an ensemble of Machine Learning (ML) and rule-based models (Python/spaCy/Scikit-learn) with a PDF parsing library (Apache Tika) to scan the text.
- **Key Feature Identification:** Automatically extracts the following critical trial parameters:
  - **Pathology:** The disease or condition under investigation (e.g., HIV, TB, Oncology).
  - **Trial Phase:** The stage of the trial (Phase 1, 1/2, 2, 3, 4).
  - **Statistical Analysis Plan (SAP) Presence:** Binary classification (Yes/No).
  - **Effect Estimate:** Identification of disclosed effect size or confidence intervals.
  - **Sample Size (Number of Subjects):** Extracted patient count, factoring in "per arm" or "per cohort" modifiers.
  - **Number of Arms/Groups.**
  - **Countries of Investigation.**
  - **Simulation Use:** Determines if simulation was used for sample size determination.

3.2. Risk Modelling and Cost Modelling

- **Risk Scoring:** The extracted features are fed into a linear risk model/scoring formula to generate a risk score and corresponding High/Medium/Low rating.
- **Cost Prediction:** Leverages predictive analytics models trained on past trial data to estimate the total cost based on phase, pathology, sample size, and country count.

3.3. Reporting and Insights (Output)

- **Automated Reporting:** Generates a comprehensive output report in PDF or Excel format.
- **Design Issue Flags:** Highlights potential design problems (e.g., underpowered study, missing SAP).
- **Recommendations:** Provides targeted suggestions for protocol improvement (e.g., increase sample size).
- **Itemised Budget:** Provision for detailed, per-patient budget spreadsheets calculated from the Schedule of Events (SoE).
- **Cost Benchmarking**: The Clinical Trial Risk Tool calculates a benchmark cost for your trial by comparing your trial's conditions, interventions, and other properties to a dataset of past trials where total cost information is in the public domain.

4. Target Users and Use Cases

User Group	Use Case	Benefits

Sponsors	**Rapid Triage:** Quickly assess risk for incoming funding proposals to determine which protocols require deeper review.	Consistent, objective, and rapid assessment for resource allocation.
CROs	**Budgeting**: Create an itemised per-subject budget	Quickly iterate trial protocol amendments and new amended budgets
Investigators/Academics	**Pre-Submission Check:** Perform a preliminary risk assessment on their own trial designs before submitting for funding or regulatory approval.	Improve protocol quality and increase funding/approval success rates.
Regulators/Ethics Committees	**Quality Check:** Ensure core elements, like the SAP and adequate powering, are explicitly addressed in the protocol.	Enhanced oversight and patient protection through better trial integrity.

5. Technology and Architecture

Component	Description	Technology Stack
Front-End/Application Layer	Interactive web application for protocol upload and results display.	Javascript
Data Extraction Layer	Handles document parsing and text preparation.	Java (Apache Tika), Python
Analytics/ML Layer	Machine learning and rule-based models for feature extraction and scoring.	Python (Scikit-learn, spaCy, NLTK, optional use of OpenAI)
Risk/Cost Model	The core linear scoring formula and predictive	Proprietary Algorithm, Python

	regression models.	

6. Service Availability and Support

6.1. Availability

The public-facing web application of the CTRT is targeted for **99.9% uptime** (excluding scheduled maintenance) as a non-critical assessment tool according to the Microsoft Azure VM SLA.

6.2. Service Support (Tiered)

Tier	Type	Scope
Tier 1	**Community Support**	Users of the open-source version are directed to the public repository for bug reporting and community-driven Q&A.
Tier 2	**Direct Commercial Support**	Available for enterprise clients with commercial licensing. Includes SLA-backed bug fixes, feature requests, and dedicated consultancy.

7. Service Metrics (KPIs)

Key performance indicators for the CTRT service measure its accuracy, efficiency, and adoption:

- **Protocol Processing Speed:** Average time from PDF upload to report generation (Target: < 30 seconds).
- **Risk Model Accuracy:** Reported here: https://clinicaltrialrisk.org/accuracy/