Johns Hopkins University

ENABLING AUTOMATION IN CLINICAL RESEARCH ADMINISTRATION THROUGH GOOD DATA MANAGEMENT

A Capstone Paper Submitted to the
Krieger School of Arts and Sciences
Advanced Academic Programs
in Partial Fulfillment of the Degree of
Master of Science in Research Administration

by

VICTOR POONAI

Baltimore, Maryland
December 2018

Abstract

The amount of clinical trials moving through the clinical trial lifecycle pipeline continues to increase every year. From 2000 to 2013, the number of registered clinical trials increased by two orders of magnitude. Also increasing are the amount of steps and processes both scientific and administrative required for a clinical trial to proceed through the pipeline. This trend contributes to increasing study complexity and burden of work on clinical research administrators. Traditional methods of facilitating and completing clinical research administrative tasks may not be effective in dealing with increasing numbers of simultaneous trials and increasing number of required processes to complete. There is a risk of further increasing the time of the clinical trial lifecycle due to overloaded clinical research administration. In this paper, concepts of clinical data management are introduced in the context of clinical research administration. Key data points common within clinical research administrative activities are presented, along with application of that data in improving clinical research administration efficiency through both non-programmatic and programmatic automation, reducing administrative work burden, costs, and ensuring current turnaround time for work processes will not be impacted by increasing workload.

Table of Content

Chapter 1. Introduction

1.1. Background

The number of clinical research trials being conducted continues to grow every year, along with clinical trial design and procedural complexity. The number of processes and procedures required to successfully execute a clinical trial have increased both for medical personnel and administrative personnel. The literature documents an increasing amount of trial procedures over time identified by in-depth analysis of institution-level processes and commonalities across institutions in executing a clinical trial. Also available is research into the impact of complexity on study outcomes, such as recruiting the required number of subjects for the study or ensuring appropriate timelines are met. Literature also exists discussing the impact of complexity on study workload and the increasing burden it presents on personnel at all stages of the clinical trial life cycle. In particular, steps in the process with the most significant impact on study workload were those deemed repetitive or that involved numerous parties.

Research into methods of assessing and alleviating study complexity have been conducted. Some researchers have presented scoring criteria to assess study design in order to better understand the impact a study's complexity may have on the workload of employees from an institution participating in a clinical trial as a study site. Other researchers have proposed a data science approach to assess study design and during concept development and workload impact after study approval. However, no research provides a practical and immediately implementable measure that equips institutions and clinical research administrators to handle increasing workload in clinical research administration.

1.2. Statement of the Problem

None of the literature addresses the need for clinical research administration to be equipped with new skill sets that prepare them for the trend of increasing study numbers and complexity, thus mitigating some of the negative impacts of large amounts of complex studies at a grass-roots level. Furthermore, none of the literature provides practical and immediately implementable steps that a clinical research administrator may take to alleviate the increased work burden of managing more studies every year that are increasing in complexity. Options that are available for clinical research administrators to handle increasing work burden are presented in this paper.

1.3. Research Questions

How can clinical research administrators and institutions benefit from applying data management principles? How can automation be applied to clinical research administration? What are some tasks in different areas within clinical research administration that can serve as examples of how to apply data management and automation principles? How can clinical research administrators accurately identify steps or tasks within their administrative processes that could be augmented by data management principles and automation?

1.4. Objectives

Methods to improve efficiency and prepare for larger clinical research administration workloads by using data management and automation will be explored. The concepts of data management, data management plans, data quality, ways to use data, and automation will be presented and reviewed. An approach will be developed on how to adapt the concepts of data management and automation to clinical research administration in a way to improve efficiency. A way to think about data management and automation from a clinical research

administration standpoint will be developed as well as how to assess internal processes for opportunities to apply data management and automation to improve work efficiency. Different approaches will be evaluated using practical scenarios and easily repurposed materials. Finally, code written specifically for this paper will be provided, along with detailed instructions, that demonstrate how data management principles can enable clinical research administrators to automate tasks and improve efficiency.

1.5. Significance

The significance of this project is two-fold. First, this project introduces clinical research administrators to technological approaches that have been trending in other industries for years. These industries have used this technology for operationally beneficial applications, such as forecasting project workload, predicting project success, and automating a plethora of tasks, thus freeing up resources for more mission-critical work. As trends indicate that the number of clinical trials will not only continue to increase in number, but in complexity as well, it is vital that approaches to clinical research administration tasks adapt to a higher volume of work at a fundamental level. Specifically, clinical research administrators must look at the processes and procedures they directly execute on a daily basis and identify opportunities to improve efficiency in these processes and procedures through readily available technologies. This approach frees up time to either handle other tasks critical to their institution's mission or to further prepare their institution's clinical research administration processes for the future.

Second, this project is significant in its potential to further drive changes to how clinical research is conducted. Numerous organizations both public and private agree that the way clinical research is designed, developed, and conducted is inefficient and takes too much

time. By adopting new technologies and schools of thought from other professional fields, clinical research administrators could drive transformation in clinical research execution from a grass-roots level by harnessing data and automation. This grass-roots approach could enable institutions to more quickly adopt new technologies and deploy them on a broad scale, as their own personnel could demonstrate the technologies benefits.

1.6. Limitations – The Nature of Clinical Research Administration

The scope of work for clinical research administration professionals can vary greatly. Whether responsible for safety committee organization, study accrual monitoring, case report form (CRF) development, or protocol amendment analysis, clinical research administrators can be found at almost any part of the clinical trial lifecycle. While as many examples as possible will be covered, this paper is unable to cover every possible process or task that a clinical research administrator may face. The goal is to provide readers with an introduction to the concepts of data management and automation in the context of example tasks clinical research administrators in different institutions and organizations may face.

Chapter 2. Literature Review

2.1. Overview of Literature Review

The National Institutes of Health (NIH) tracks the number of registered clinical trials over time at the *ClinicalTrials.gov* website, starting from the year 2000 up to 2018[1]. Records indicate that from 2000 to 2013 the number of registered clinical trials increased by two orders of magnitude, from 1,255 registered trials to 137,532. As of October 2018, there are 262,431 registered clinical trials. Calculating the average percentage increase of registered trials per year from 2000 to 2018 yields an average increase in registered clinical trials per

[1] National Institutes of Health. (2018). *ClinicalTrials.gov Trends, Charts, and Maps*. From https://clinicaltrials.gov/ct2/resources/trends

year of 1,156%. The number of registered clinical trials compounded by the amount of studies involving multiple hospitals as study sites, all with differing processes and procedures, paints a more specific image of what the clinical research administration landscape may look like in the future should these trends continue.

Getz and Campo monitored and assessed the state of clinical trial complexity over the course of 14 years. In their publication titled *New Benchmarks Characterizing Growth in Protocol Design Complexity*, they found that of the sites monitored, the mean number of total procedures to execute for a clinical trial increased for all phases across all therapeutic areas[2]. These results reflect an increase in workload needed to execute all required procedures both medical and administrative in a clinical trial. The average increase of total procedures for a clinical trial was 202 for Phase I studies, 177 for Phase II studies, and 150 for Phase III. With all Phases experiencing increases in total procedures, traditional approaches to completing clinical research administration tasks may not be sufficient enough to keep up with the increasing administrative workload.

A study by Dilts et al. sought to map all processes and procedures for Phase III cooperative group clinical trials from first inception as a concept at an institution to study opening and participant registration. By analyzing electronic records, trial initiation data, and process documentation, they found that 769 steps were required to open the Phase III study[3]. In their analysis, they identified two types of processes required to open the study. Work steps were activities that needed to be performed before a study could open. Decision points determined routes where the products of work steps would go if they met a certain criterion.

---

[2] Getz, K., Campo, R. (2017) *New Benchmarks Characterizing Growth in Protocol Design Complexity.* doi: 10.1177/2168479017713039

[3] Dilts et al. (2010). *Phase III Clinical Trial Development: A Process of Chutes and Ladders.* doi: 10.1158/1078-0432.CCR-10-1273

Of the 769 steps identified, 84% were work steps, the majority of which were clinical research administrative tasks ranging from form development to communication and coordination. The authors noted that there is a risk that some work steps could be slowed down or overwhelmed should higher volumes of more complex clinical trials begin to flow through these process routes. Of more concern is the additional slow-down that could occur should certain steps require repeating. The authors compared the process to the child's game Chutes and Ladders, illustrating how many steps act as chutes, requiring that they be completed again (up to six times in some cases). The authors also noted concerns about how much overwhelmed administrative processes would further increase the time taken to complete the clinical trial lifecycle, and thus increase the time that possible life saving therapeutic innovations take to reach patients in need.

Getz further elucidates the nature of clinical "chutes" and their impact on clinical research administrator work burden in an article focused on a familiar topic to many clinical research administrators; protocol amendments. In *Measuring the Incidence, Causes, and Repercussions of Protocol Amendments*, Getz presents his results from monitoring changes to 3,410 clinical protocols over a two to four-year period[4]. Of these protocols, 6,855 protocol amendment changes were categorized. 39% were modifications to the eligibility criteria description, adjustments to the number and types of safety assessments, and alterations of general protocol information (e.g. contact information, site information, etc.). The median time taken to implement changes was 65 days, with almost half of that time used on finding the correct way to rewrite the protocol to accurately reflect the amendment. This process can involve a significant amount of back and forth communication between clinical research

[4] Getz et al. (2011). *Measuring the Incidence, Causes, and Repercussions of Protocol Amendments*. doi: 10.1177/009286151104500307

administrators, medical practitioners, and others before a revised protocol is generated that all parties find agreeable. More than half of amendments made in Phase I studies occurred before first patient enrollment, with one-third of all amendments being classified as avoidable (i.e. caused by human error). This outcome is interesting when considering Getz findings in *Trends in clinical trial design complexity* from 2017, where he reported on a positively correlating trend between study complexity and protocol amendment amount continuing with no sign of slowing down[5].

Dilts and Sandler dedicated a project to assess the clinical research administration barriers to opening clinical trials, where they identify redundancies and deficiencies present between administrative departments within their target institution and their interactions with external entities[6]. Note that in the process map presented in Figure 1 of *Invisible Barriers to clinical Trials*, a single work step submitting some forms comes before a large series of reviews and approvals. Before that step are two steps related to form preparation. From an administrative standpoint, this step is crucial to ensure research activities can proceed in a timely manner. However, consider this subset of the greater clinical trial development process while accounting for the trend of increasing clinical trial numbers and complexity. There is a risk of a bottleneck emerging should a clinical research administrator not be well equipped to handle a significant number of forms for complex studies.

Concerns over operational efficiency in clinical research administration have not gone unvoiced by professionals in the field. Additionally, possible solutions and approaches that enable clinical research administrators to improve operational efficiency have been written on. In 2017, Regan wrote on the ways technology could improve clinical research execution.

[5] Getz, K., Campo, R. (2017). *Trial Watch, Trends in clinical trial design complexity.* doi: 10.1038/nrd.2017.65.
[6] Dilts, D., Sandler, A.B. (2006). *Invisible Barriers to Clinical Trials*. doi: 10.1200/JCO.2005.05.0104

Specifically, he points to automation as a means to mitigate risk inherent in manually completing clinical research administration tasks while improving task execution efficiency[7]. Study status reports are an example of a regularly required task that could be automated to avoid reporting errors and to save report preparation time. Furthermore, automated study status reports based on high quality data can enable sponsors to act on new information quickly and with confidence while allowing clinical research administrators time to handle other vital tasks.

Malikova discussed the problems faced in clinical trial execution due to increasing protocol complexity and how the industry is adapting by creating innovative methods of optimization both in study design and execution[8]. She noted the importance of data from previous studies in enabling optimization throughout a study's life, especially during the early development phase. Clinical research administration professionals are uniquely positioned to enable this time saving early development phase technique by acting as stewards of high quality administrative data. Such information can be effectively used to categorize studies into specific clusters for further analysis that can help researchers and administrators alike determine the best course of action to take that will lead to an accepted and efficient study. Smuck et al. developed a thorough protocol complexity scoring system which has been adopted by numerous institutions as an effective tool to gauge study complexity and work burden. The tool, presented in the *Ontario Protocol Assessment Level: Clinical Trial Complexity,* considers common clinical research-related tasks under categorized groups, a

[7] Regan,B., (2017). *Can Technology Foster Better Collaboration? Applied Clinical Trials*. from: http://www.appliedclinicaltrialsonline.com/can-technology-foster-better-collaboration

[8] Malikova, M.A. (2016). *Optimization of Protocol Design: A Path to Efficient, Lower Cost Clinical Trial Execution*. doi: 10.4155/fso.15.89

significant number of which are clinical research administration tasks[9]. The fact that administrative tasks were considered enough of a complicating factor to be included in the OPAL complexity scoring system speaks to the importance of enabling clinical research administrators to utilize tools that improve their efficiency without sacrificing quality. Data Monitoring Committees (DMCs) or Data Safety Monitoring Boards (DSMBs) are independent safety oversight bodies required for large, randomized, multisite clinical trials[10]. Organizing these committees requires a significant amount of time and coordination. Furthermore, studies may not begin until the first meeting of the committee, which in addition to the committee typically involves the principal investigator, sponsor, regulatory affairs personnel, data collection entity, and independent safety monitors.

Clinical research administration professionals may find themselves tasked with coordinating committee organization. Coordination activities can involve identifying committee member candidates with specific expertise, organize their conflict of interest (COI) analysis, and coordinating data reviews with them[11]. This process was previously time consuming and cumbersome, as work was exclusively conducted manually. To enable more efficient committee organization while managing an ever-increasing number of studies, some NIH divisions have organized the creation new systems dedicated to committee and safety data management that mitigate human error and save time.

In summary, the number of clinical trials entering the the clinical trial pipeline is increasing every year. Additionally, the overall process for designing, obtaining approval, and

---

[9] Smuck, B. et al. (2010). *Ontario Protocol Assessment Level, Clinical Trial Complexity Rating Tool*. doi: 10.1200/JOP.2010.000051.

[10] Food & Drug Administration. (2006). *Guidance for Clinical Trial Sponsors*. From: https://www.fda.gov/downloads/regulatoryinformation/guidances/ucm127073.pdf

[11] Division of Microbiology and Infectious Diseases. (2017). *DMID Safety Information Sheet*. From: https://www.dmidcroms.com/Shared%20Documents/Safety%20Oversight%20and%20SAE%20Reporting%20Info%20Sheet.pdf

completing a clinical trial are also increasing. As additional steps are added to the process, everyone involved in facilitating the clinical trial lifecycle will experience increasing burden in workload. Concerns over trial design complexity and amount are significant enough that institutions have developed ways to assess workload burden based on a trial's design.

Chapter 3. Need(s) Assessment

As demonstrated in the literature, the amount of clinical trials being submitted and initiated is trending upwards. Simultaneously, the means by which these trials are assessed and conducted are becoming increasingly complex with each year. Numerous research projects have elucidated the nature of trial complexity and the means by which that complexity may be alleviated. However, none of these research projects have provided methods specifically for clinical research administration related tasks to be completed more efficiently, despite many articles pointing towards clinical research administration as a source of study complexity and long development times. Furthermore, should the amount of clinical trials and their complexity continue at increase at the current rate, existing processes and procedures utilized by clinical research administration professionals may be insufficient to maintain current task completion rates and may in fact become slower. While needed, proposing a mass implementation of broad and sweeping changes across the clinical research administration field is neither practical nor reasonable. Instead, clinical research administration professionals must be equipped with tools that enable them to drive transformation in clinical research operational and administrative conduct. Using this grass-roots approach, the pipeline through which protocols start at concept and end in study close-out and final clinical study report will be able to handle the increasing flow of ever more complicated studies without slowing down.

Chapter 4: Project Description

4.1 Description

As evidenced in the literature, the burden of clinical research administration workload remains a concern among policy makers and leaders in the United States health research enterprise, as there is a continued trend towards higher workloads that may be difficult to handle utilizing standard approaches. This project seeks to provide practical knowledge and practical solutions from the fields of data management and data science that clinical research administrators can utilize in their day-to-day work activities. This goal would be achieved by introducing clinical research administrators to core concepts from data-centric professional fields and providing clinical research administration focused tools to assess administrative tasks for opportunities to utilize data-centric approaches. Engaging in this approach can benefit clinical research administration professionals now by expanding their skill sets and knowledge to augment their institutions administrative efficiency, and in the future by helping clinical research administrators prepare for the above-mentioned trends of increasing workload.

Chapter 5. Methodology

5.1.1 Project Methodology Overview

This project will utilize a practical approach for educating clinical research administrators by identifying relatable tasks within the clinical research lifecycle. These tasks will then be framed in a way that best practices and principles for data management can be applied. Best practices will be taken from the Good Clinical Data Management (GCDMP) guidelines. This approach will also allow for tools to be introduced to readers in a logical and natural manner, where each step in the education process leads to the creation of a unique tool

which can be adjusted by the reader to suit their specific needs in day-to-day clinical research administration activities.

It is important to note that this project is not intended to encompass all possible clinical research administration scenarios. Instead, this project is meant to provide clinical research administration professionals with a general overview of how best practices and trends from data-centric professional fields, such as data management practices and automation implementation, might be applied in the course of their normal work-related tasks. This project may be expanded upon by future researchers to cover more specific topics and further provide tools and educational materials that prepare clinical research administrators for the future, both in terms of managing increasing workloads and acting as a grass-roots force for transformation in the clinical research process.

The project will culminate in a set of tools and resources which clinical research administration professionals can refer to as examples and adjust as needed to suit the tasks they must complete in their work. Example tools that will be provided include a checklist of questions to determine if a task is conducive to data management principle application, a summary of clinical data management practices and principles in the context of clinical research administration, and basic and advanced example scenarios where data-centric approaches to clinical research administration tasks enable improved work efficiency through automation using standard office programs and scripts developed for this project using the Python programming language. To that end, it is necessary for this project to narrow its scope of work in two steps.

5.1.2 Establishing Project Scope

First, this project will define what a clinical research administrator is and the type of work they are engaged in. This definition will be backed by documentation from academic literature and professional organizations. Second, this project will limit common tasks in the clinical trial process to two task types contributing to study initiation (e.g. protocol development stages) and two task types executed during the course of the study (e.g. study recruitment monitoring). The project will proceed to identify the types of tasks in the clinical research lifecycle where clinical research administrators are likely to be involved.

5.1.3 Clinical Component- Methodology and Rationale

To identify common clinical research administration tasks, this project will analyze the Good Clinical Practice (GCP) guidelines created by the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). This document provides standards for the "design, conduct, performance, monitoring, auditing, recording, analyses, and reporting of clinical trials" designed to protect a study's credibility and the rights of recruited study subjects[12]. Specifically, an overview of ICH GCP E6 (R2) will be used, as this addendum to the original GCP guidelines provides additional guidance and best practices that account for the increased use of electronic health care records and IT technologies that improve the efficiency and credibility of clinical trials. This document will also be used later as an additional source of examples of where data-centric approaches to clinical research administration tasks.

Initially, this project considered utilizing Dilts' extensive mapping of the clinical trial process because the literature review identified him as the foremost researcher of the clinical

[12] https://ichgcp.net/

trial life cycle and its increasing complexity. A significant amount of the literature related to this topic came from Dilts. However, mapping created from Dilts' research was extremely narrow in terms of study Phase and therapeutic area. This project seeks to provide an overview of generalized clinical research lifecycle steps and data-centric approaches to these steps. Therefore, the ICH GCP guidelines were chosen over Dilts' research.

The number of steps and location of these steps within the clinical trial lifecycle will be limited as explained above. These steps will be identified by (1) their presence in ICH GCP documentation in the context of tasks within the purview of the clinical research administrator definition as established earlier in the project and (2) their relevance to the scope of work as per this paper's definition of a clinical research administrator.

For example, ICH GCP guidelines discuss establishing an Independent Data Monitoring Committee (IDMC). These Committees can be known by multiple names, such as Data and Safety Monitoring Board or Safety Monitoring Committee. Clinical research administrators may be required to coordinate or organize the creation of such Committees, which involves a significant number of steps. Such a task would be a suitable candidate for use in this project as an educational example.

Once four tasks are identified (that is two tasks contributing to study initiation and two tasks during the study), a general overview of the necessary work to complete these steps will be provided. With the clinical scope established, the project will proceed to the data-focused component.

5.1.4 Data Management Component- Methodology and Rationale

An overview of Good Clinical Data Management Practices (also known as GCDMP) will be introduced to the reader. GCDMP are a set of guidelines and methods of best practices

for clinical data and electronic records management set forth by the Society for Clinical Data Management (SCDM)[13]. While a general overview of the document will be established, concepts and practices relevant to the tasks selected in the narrowed clinical research administration component will be reviewed in detail. Specifically, guidelines will be compared to the steps needed to complete the tasks identified using through the method described above in the clinical component section. The objective is to guide readers into thinking about their tasks from a data-centric perspective and to consider GCDMP in the context of their work.

5.1.4 Planned Output – Practical Application of Practical Knowledge

The above approach will lead to the introduction of a checklist readers can use as a starting point to assess GCDMP applications for their professional tasks, adjusting as required for their unique situation. Additionally, the utilized guidance from GCDMP will be presented in a way that fits the context of clinical research administrator activities. This approach is used to further encourage clinical research administration professionals to think about GCDMP and their day-to-day tasks and to demonstrate how guidelines can be modified to suit unique needs. By seeing how GCDMP can be adapted to suit administrative work, other clinical research administration professionals may continue to "translate" other best practices found in GCDMP.

With GCDMP and its application to clinical research administration introduced, the project will guide readers from practical education to practical execution of what has been presented thus far. Using the four previously identified tasks in the clinical research lifecycle and the steps required to complete them, scenarios which demonstrate how these tasks can be

[13] https://www.scdm.org/publications/gcdmp/

completed using a data-centric approach. Emphasis will be placed on demonstrating how utilizing GCDMP can enable automation in clinical research administration tasks with common office programs. These scenarios will culminate in a step-by-step instructional guide to complete the scenario task. The guide will be created in a modular way, allowing readers to adjust the tool for their unique needs.

In addition to simple automation methods using office programs, advanced automation methods will be presented. These advanced methods will further encourage clinical research administration professionals to visualize their work flow and how data structures within it can move and be adjusted to create their final product. Because the advanced automation methods utilize a programmatic approach, step-by-step instructions will not be practical. Instead, a walkthrough of the program's concept will be provided to assist readers in understanding the approach and utilizing it in their professional tasks. The final code for each program will be included in the Appendix of this project. Advanced methods will be presented to show readers how programmatic approaches to clinical research administration are feasible and executable by anyone. The advanced methods are also included to demonstrate the practicality, readability, and easily grasped syntax of the Python programming language when used for clinical research administration.

Chapter 6. Project Results and Discussion

6.1. A Definition of Clinical Research Administration

There is no single definition limiting what clinical research administration is. Numerous educational programs and offices focused on clinical research administration exist in various institutions. While no single definition exists, overlap in many clinical research administration duties is found when reviewing mission statements and offered services from

the aforementioned offices and other sources, such as professional organizations and academic publications.

Rush University's Office of Clinical Research Affairs (RU OCRA) defines their scope of work by providing a definition of clinical research administration. The office defines clinical research administration as the facilitation of administrative and financial aspects of clinical research[14]. Among their goals, the office seeks to ensure that PIs and research teams are audit ready, studies start in an efficient manner, and clinical research related activities are effectively executed. RU OCRA further expands upon the ways that they support their investigators in clinical trials, listing activities such as clinicaltrials.gov application support, synchronization of protocol related documents prior to contract execution, and research coordinator support.

The Society of Research Administrators International (SRAI) offers training sessions for a variety of research administration related areas, including clinical trials research administration[15]. The training session covers elements crucial to clinical research administration, including protocol review, recruitment, compliance, and risk management and analysis.

The Association of Academic Health Centers (AAHC) released a report in 2009 discussing the increasing importance of clinical research administrator activities before, during, and after a clinical trial[16]. In an attempt to quantify the extent of clinical research administration's importance in clinical research execution, the report sent a survey to eight institutions asking them to identify clinical research administration tasks common to all

[14] https://www.rushu.rush.edu/research/office-research-affairs/clinical-research-administration
[15] https://www.srainternational.org/learn-0/certificate-programs-0/ctra101-cert
[16] http://www.aahcdc.org

entities. The survey's results identified fourteen tasks that all surveyed institutions shared in common. These tasks were broad in their subject matter. Among the fourteen tasks identified in the report are protocol development, defining standard of care, patient recruitment and scheduling, and compliance.

Moffitt Cancer Center's Clinical Trial Office (MCC CTO) provides a list of clinical trial administration tasks which the office executes on behalf of investigators[17]. These tasks include providing information management support for clinical trials, ensuring study compliance, submitting regulatory documents, adverse event reporting, and data collection.

The University of Arkansas for Medical Science Translational Research Institute (TRI) developed a system dedicated to facilitating fundamental clinical research administration related tasks[18]. The system, named CLARA (CLinicAl Research Administrator), has specialized modules for interfacing with various oversight committees, mapping adverse event reporting workflows, and facilitating general clinical trials management.

The Society of Clinical Research Associates (SOCRA) provides what may be the most comprehensive explanation of what clinical research administration entails[19]. As demonstrated thus far, clinical research administration encompasses a broad range of tasks. SOCRA recognizes and takes into consideration the diversity of work a clinical research administrator may be required to complete. First, SOCRA states that clinical research administration professionals may fall under different work titles, including coordinator, data manager, and regulatory affairs manager. Second, the duties listed are as varied as the titles a

[17] https://moffitt.org/clinical-trials-research/clinical-trials/clinical-trials-administration/
[18] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4173184/
[19] https://www.socra.org/

clinical research administrator may work under. These duties include but are not limited to data collection, data monitoring, committee coordination, participant recruitment, preparation of various study progress reports, and preparation of adverse event reports.

Another duty clinical research administrators may handle relates to independent data monitoring committees (IDMC). While 21 CFR does not explicitly require the use of IDMCs or data and safety monitoring boards (DSMB) outside of emergency settings (21 CFR 50.24(a)(7)(iv)), the NIH requires that each institute and center under its purview maintain a system to manage and implement safety oversight for patients in clinical trials. Some federal offices still require IDMCs for some or all studies regardless of 21 CFR. For example, the National Institute of Allergy and Infectious Diseases (NIAID), an institute under the NIH, requires IDMCs for "all randomized clinical trials of any phase that involve both investigator-masked interventions and enrollment of greater than 100 subjects"[20]. NIAID's Division of Microbiology and Infectious Diseases (DMID) utilizes a support contractor called the Safety Oversight Committee Support (SOCS) group to facilitate the organization, coordination, and management of these committees. In addition to committee management, the group's activities encompass a myriad of clinical research administration tasks including serious adverse event processing, study progress reporting, and clinical document management.

6.2. Common Research Administration Duties

As shown above, clinical research administration duties can greatly differ from professional to professional. However, there are certain duties that are continually mentioned across all information sources common to clinical research administration. Furthermore, these common duties are not relegated to one step within the clinical research lifecycle. For

[20] https://www.niaid.nih.gov/research/safety-oversight-clinical-research

example, a critical duty that clinical research administrators may be tasked with mentioned in multiple information sources was protocol development. A Protocol is a comprehensive plan by which a trial is conducted, explaining how each component is carried out, elucidating eligibility criteria, and timelining the length of the study[21]. Development occurs before a study is approved to begin and recruit its first participant. Another critical duty a clinical research administrator may be tasked with, regulatory compliance, occurs throughout a study's active period and beyond. All tasks are related to tasks outside the medical science of a clinical trial, such as the treatment of patients, execution of medical therapies, and design of patient care procedures. This leaves tasks related to operations, management, and administration not necessarily relegated to one type of duty (e.g. budgeting).

Therefore, this paper will define clinical research administration as the management and execution of any and all administrative and/or operational duties related to clinical research development and execution. This definition can encompass all above mentioned duties and reflects the diverse nature of the clinical research administrator. With the breadth and depth of clinical research administration's scope established, specific duties will be focused on for analysis and application of data management best practices and automation potential. These duties will be categorized into tasks before study start and after study start and are selected based on their subject matter focus' presence in ICH GCP.

It is important to note that selected duties could include various subtasks that must be fulfilled in order for a duty to be completely fully. This paper's scope is establishing a conceptual foundation of applying data management principles and automation to clinical research administration and does not seek to cover each possible subtask within or

[21] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5198475/

permutation of the above-mentioned duties. Therefore, to ensure that example use cases are comprehensive and practical, duties will be further narrowed to specific subtasks aimed at specific goals. The below section defines the duties to be focused on and their narrowed scope.

6.2.1 Common Duties Before Study Startup

6.2.1.1 Protocol Development

The protocol development stage is where the processes and procedures for executing a clinical trial are fully fleshed out in the form of a document called a protocol. Multiple entities can be involved in this process, where sponsor and government personnel work together to create what can be seen as the "manual" for how to complete a study. Al-Jundi and Sakka published an article reviewing common components and benefits of a clinical protocol[22]. Of interest to this paper is the Methodology section's data collection components. Al-Jundi and Sakka indicate that the Methodology section "defines the variables and demonstrates in detail how the variables will be measured." As per this paper's definition (and the information sources used to establish that definition), a clinical research administrator may be tasked with contributing to a protocol's development by identifying relevant data types for capture during the study. This can involve in-depth literature reviews or reviewing past protocols to identify data commonly captured for study's similar in nature.

6.2.1.2 IDMC Organization

IDMCs can be known by a host of other names such as Data and Safety Monitoring Board (DSMB) or Safety Monitoring Committee (SMC). In trials such as those under the purview of NIAID DMID, these committees must meet and provide their recommendations

[22] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5198475/

on a study's design from a safety perspective before they can continue. Clinical research administrators tasked with organizing such committees can expect to be involved in identifying possible members by their professional expertise, drafting invitation materials, organizing conflict of interest reviews, and providing administrative support to committee members from the first organizational meeting that allows a study to begin through to the study's end.

6.2.2 Common Duties After Study Startup

6.2.2.1 Protocol Deviations

The first post-study startup activity this paper will focus on relates to protocol deviations. Protocol deviations include any actions that occurs outside the defined parameters of the protocol. Deviations can occur for a number of reasons, such as study staff intentionally deviating from protocol-defined procedures out of concern for a participant's safety. Clinical research administrators involved in monitoring or processing protocol deviation reports from study sites may find themselves assigned as stewards of deviation information, having to extract data from report forms or design deviation collection methods.

6.2.2.2 Study Status Reporting

Whether stakeholder or federal agency, specialized study operations and administrative reports may be requested as regularly submitted deliverables. The frequency and subject matter of these reports can greatly vary. To illustrate, consider a clinical research administrator assigned to safety oversight coordination and support for ten clinical trials. For each trial, they must provide a weekly report on meeting activity between the study sites and federal agencies, a monthly report on data reviews by IDMCs, a monthly report on number of deviation and serious adverse events that occurred per study, and a general study benchmark

progress report. As the trend of ever increasing clinical trials identified in the literature review continues, reporting duties become increasingly work heavy.

6.3 Clinical Research Administration and Good Clinical Data Management Practice

GCDMP are a set of guidelines and methods of best practices for clinical data and electronic records management set forth by the Society for Clinical Data Management (SCDM). The primary focus of the document is providing recommendations related to ensuring regulatory compliance in the realm of clinical data management. However, the document also contains practical suggestions and methods by which readers can meet said recommendations. Covering a total of 20 sections, GCDMP encompasses best practices for all clinical data-related components of a clinical trial in technical depth. This paper will focus on three specific sections containing recommendations and best practices relevant to clinical research administration; Data Entry and Processing, Dictionary Management, and Assuring Data Quality.

6.3.1 Data Entry and Processing

GCDMP states that the objective of data processing is to "efficiently produce quality data for analysis." Typical clinical data analysis involves processing data through statistical models and programs (e.g. SAS). Data quality directly impacts the reliability of clinical data analysis. Clinical research administrators should pay particularly close attention to the concept of data entering a model and exiting the model as a calculated result. This concept, known as Input/Output (I/O), establishes the importance of good data management in clinical research administration. Instead of manually completing deliverables, such as IDMC invitations, clinical research administrators can think about their task from an I/O model perspective, where data can be input into a program, script, or template, with output being a

final product ready to deliver to needed stakeholders. Higher quality data results enable high quality automated output with little to no need for time consuming quality checks. I/O is a vital conceptual foundation for clinical research administrators seeking to apply data management and automation to their work.

Whatever clinical research administration data one may utilize, GCDMP provides relevant best practices to ensure the data's integrity in the course of its use. Most relevant is the best practice of ensuring quality control in one's data. The GCDMP provided example of a quality control measures is to periodically inspect samples of data then take corrective action if errors are found. Note that a data source does not need to be a large web-based database. A data source is simply where data is located, such as a spreadsheet. Whether a data source is handled by one or more clinical research administration professionals in a working environment, such a measure helps ensure data quality. There are two best practices for data entry and processing that, when adjusted to fit the correct context, are relevant to clinical research administration:

- Use written procedures that describe data processing steps and required quality level.
- Address the purpose, characteristics, and complexity of each data source.

6.3.2 Data Dictionaries

Data dictionaries define specialized coding criteria typically used for clinical events, medical history, and diagnosis data types. While clinical research administrators are unlikely to be involved in this technical component of clinical research, the concept of defining data types is useful for clinical research administrators. To ensure high quality data is maintained, it is important to define what data is being collected, why that data is collected, where is the data coming from, and what the data will be used for. Establishing these definitions helps

maintain data integrity and ensures the right data is reliably present and usable. There are two best practices for data dictionaries that when adjusted to fit the correct context are relevant to clinical research administration:

- Establish a process for evaluating a change in a dictionary.
- Store all versions of dictionaries for future reference.

6.3.3 Maintaining Data Quality

As previously stated, high quality input (data) ensures high quality output (results). If data is managed poorly, the benefits of its use in clinical research administration will not be realized and can negatively impact quality. Specific processes are needed to ensure data quality is maintained from the moment it is collected. This concept again relates to the concept of I/O. There are two best practices for data quality during collection that when adjusted to fit the correct context are relevant to clinical research administration.

- Include as few steps as possible in the data collection and data handling processes to eliminate the chances for error.
- Collect only data essential for intended use.

6.4 Assessing GCDMP Suitability in Clinical Research Administration Duties

Consistency and dependability are continually present themes throughout GCDMP. These concepts are applicable to clinical research administration, as categorizable data flows through clinical research administration work processes. Furthermore, this categorizable dtaa can have a dependable source related to it. Patterns and themes emerge when stepping back to assess clinical research administration work, how it is achieved, and what is required to achieve it. Doing so allows one to exploit the patterns in a systematic way, enabling the confident use of automated processes to handle certain tasks, thus easing workload burden.

To identify these patterns and categories, tasks should be considered with the data management concept of I/O in mind. Answering the questions in the below tool can help one identify whether a task could easily utilize GCDMP principles and thus immediately (or near immediately) be automated:

Question	**If Yes**	**If No**
Is the input required for the task in a structured (e.g. tabular protocol deviation data) format?	Continue	Continue
If not structured (e.g. a narrative protocol amendment), could the input be easily reformatted to be structured?	Continue	Stop
Is the output's structure consistently the same (e.g. a report with a significant amount of boilerplate language)?	Continue	Stop
Are the output's constants (e.g. boilerplate language) and variables (e.g. form field data) consistently located in the same place in the final product?	Continue	Stop
Are the steps in the task needed to get from input to output product always the same?	Continue	Stop

It should be noted that many of the "stop" points above are not an impassable barrier to automation given enough experience. For example, even if a task's input contains unstructured data, the data can be reshaped into a structured format programmatically. However, doing so would require knowledge and experience most clinical research administrators will not have or will need time to develop. For immediately implementable and effective data management and automation in clinical research administration, data should be structured until such time that a clinical research administrator expands their skill set to include scripting languages such as Python, R, JavaScript, or Julia.

6.5 Examples of Applying GCDMP

To present how GCDMP can be utilized for clinical research administration tasks, consider the following use cases in common clinical research administration tasks established earlier. These examples are presented both to illustrate how GCDMP best practices can be applied and the benefits of applying them. The benefits focused on are expedient informed analysis and task automation. Informed analysis using previously completed studies can benefit clinical research administrators working within the steps before a clinical trial begins by providing insight from structured and easily parsed data sources. Process automation can ensure that studies and clinical research administrators are not hindered by the increased workflow accompanying the increasing number of clinical trials entering the lifecycle pipeline.

6.5.1 Identifying Necessary Data Points During Protocol Development

Clinical trials have a variety of characteristics that differentiate how they operate and what their expected outcomes are. These characteristics include study phase, target population, target disease, and study endpoints. While this data is not necessarily tabular in format, it can be captured on an ongoing basis in a table. Alternatively, if data is (a) located together on one page and (b) all contain a delimiter (e.g. ":" in Disease: Influenza), one could easily paste this text into a spreadsheet and use the "Text to Column" function to split text by a shared delimiter into multiple columns. Once done, the data can be transposed with an additional copy and paste to fit a table format. This in turn creates a "protocol profile," painting a picture of a specific protocol by key characteristics (see below example).

Protocol Number	Phase	Study Design	Recruitment Goal	Disease	Study Length	…

The more characteristics utilized, the more specific one can make their comparative analysis of the protocol in development and similar previously approved protocols. Protocol specific data points and types could then be directly matched with their study.

The most important point to bare in mind when executing this approach is that all characteristics must be linked to a single unique identifying data type (or column). In database development, this data type is referred to as the primary key. The key uniquely identifies an individual record or row in a table, allowing the row to be referenced programmatically to other tables quickly and reliably. For this example, the key would be a clinical trial's identifying ID, such as a protocol number. If an institution uses a study ID scheme that differs from other involved entities, then it is important to include all other IDs in a table matched to the institution's ID for a study. This will allow collected data to be easily referenced regardless of entity source. This approach ensures that protocol specific characteristics are kept to their protocols and enables reliable assignment of additional types of data, such as data capture points, for future comparative analysis. This approach can provide a foundation which clinical research administrators can use in identifying needed data types to capture in the study without having to start from scratch with a literature review.

6.5.2 IDMC Formation Organization

Organizing an IDMC requires inviting medical professionals with expertise in the medical field of the study that the IDMC is intended to oversee. For example, DMID requires a welcome letter, contact information sheet, conflict of interest form, and cover email for IDMC candidates. While these output items may seem extremely varied, they are not beyond application of the GCDMP best practices mentioned above. Specifically, when considering what data is essential for each output in this task and the processing steps required to go from

data to formatted output, modeling an automated solution is relatively easy. The input for this task consists of variables (e.g. names, addresses, protocol numbers, etc) placed within large chunks of constants (templated language). The language, structure, and general makeup of the output without the variables included can be templatized, leaving placeholders for needed variables. Functions found in spreadsheet and word processing programs can enable the automatic generation of all needed email language and documents customized for their intended recipients in mere seconds. This function, more colloquially known as mail merge, utilizes word processor templates and spreadsheet columns to quickly generate documents.

Columns are named within a word processing program's mail merge function that match columns in a data source containing necessary information (e.g. contact's email address, phone number, full study title). Once the tags are placed in the template, it can be linked to the spreadsheet data source. Each row in the source acts as a unique instance of the document to be generated from the template, allowing users to generate as many documents as necessary. This approach can be executed without any programming.

This form of automation can be taken one step further using a scripting language. An example of a Python-based implementation of automatic generation of IDMC candidate invitations can be found in Appendix A. Using data contained in a spreadsheet, the script takes contact information and trial information and places them into templated documents and emails for as many recipients as needed. When run from a command prompt in the folder containing the main script, the following process occurs:

1. User is prompted to enter "clinical trial number, first name, last name, role, noprior/prior-start/mid."

a. Subject matter experts can be invited under a number of different scenarios requiring unique language. To account for this, the "noprior/prior-start/mid" identifies which type of templated communication the script will populate.
b. “noprior”=intended recipient has not yet been contacted by someone (e.g. a federal employee) about this invite ahead of it being sent.
c. “prior”=intended recipient has been contacted by someone (e.g. a federal employee) about this invite ahead of it being sent.
d. “start”=intended recipient is being invited at the beginning of the clinical trial.
e. “mid”=intended recipient is being invited after the clinical trial has started.

2. User can enter as many contacts as they desire. When finished, they may press "Enter" on an empty prompt.
3. Data source sheets are parsed for the appropriate information. The final product is placed into a dictionary to allow for the fastest package generation possible.
4. Script will select and populate Word and email templates based on their file name.
    a. Word templates are populated by pre-placed mail merge tags and dictionary data.
    b. Email templates are populated by string format tags and dictionary data.
5. Outlook email is generated and fully populated with email address, attachments, and email body fully formatted and ready to send.

6.5.3 Protocol Deviation Reporting and Monitoring

Clinical research administration professionals tasked with processing and maintaining protocol deviation data will deal with data from multiple deviation cases simultaneously, and in some cases from multiple studies or study sites. The method by which deviations are

reported can vary. However, depending on regulatory requirements for a study, standardized forms are efficient ways to capture protocol deviation data. Specifically, table style forms allow for simplified data capture and tracking via spreadsheet. Whether or not the form is contained in a word processing program-related file, table formatted forms can be placed from any source are more readable by spreadsheet programs without formatting corruption. Data fields will reliably be present in the same spreadsheet cell every time the table is placed into a spreadsheet program, allowing the use of spreadsheet formulas to retrieve data and place it in the appropriate column. Much like IDMC organization, standardized tabular forms allow for automated data extraction and tracking through programmatic means. An example program implemented in Microsoft Excel using its built in the Visual Basic (VB) scripting language interpreter can be found in Appendix B. The program is simply pasted into a macro script editing window for a spreadsheet and run directly from the spreadsheet.

6.5.4 Study Operations and Administration Reporting

This particular duty is an excellent example of a clinical research administration duty where GCDMP best practices can be applied. Recall the GCDMP best practice regarding dictionaries that define data types, sources, and purposes of each. Identifying reports repeatedly requested, their content, and sources that contribute to each report, then using those to create a data dictionary can prove to be a boon to administrators. Doing so can identify common data points and sources across reports, allowing for relevant columns to be called from different spreadsheet sources, generating a report. With standardized report formats, columns, and identification of key columns that uniquely identify records within disparate reports, clinical research administrators can generate ad hoc reports of any kind when requested through features built into spreadsheet programs (e.g. VLOOKUP), so long

as the data the report is intended to focus on is part of a clinical research administrators overall data management approach. A more advanced approach to leveraging benefit from applying GCDMP to this task can once again by demonstrated using the Python programming language. In Appendix C, an example program is provided that generates a weekly report subset taken from a much larger reporting source that is attached to an automatically formatted and written email ready to send the report to relevant stakeholders.

Chapter 7. Recommendations and Discussion

7.1. Introduction

The trends shaping the future clinical research environment necessitates the need for new approaches to be considered in executing clinical research administration duties. Specifically, a multidisciplinary approach embracing core principles from GCDMP and applied to clinical research administration can facilitate easily implemented automation of specific duties, thus easing workloads on administrative and operation staff and ensuring clinical trials are able to pass through the pipeline without administrative hindrances despite the increasing volume of studies within the pipeline.

7.2. Recommendations and Discussion of Recommendations

This paper sets forth the following recommendations for policy makers and leading entities within the clinical research enterprise.

1. Consider developing an Artificial Research Administrator (ARA) A.I. to process the more mundane yet high volume administrative and operational tasks and slowly expand its capabilities the longer it remains operational. This A.I. would utilize natural language processing techniques to read, understand, and execute requests via email from entities within a clinical research environment. A prototype for such an

A.I. was developed for this paper in the form of a simple machine learning algorithm (MLA). This MLA uses term frequency-inverse document frequency (tf-idf) to stem terms and calculate relevance scores, which are then used in a logistic regression-based model. As of this paper, the model can process email-based text requests and predict appropriate categorization and actions to take with 93% accuracy. A new model is currently being developed utilizing a tensorflow-based artificial neural network.

2. Engage in a research study to identify all clinical research administration duties throughout the entire clinical research lifecycle. Duties must be common to all clinical trials.
3. Collaborate with policy makers to identify each duty's expected output format, anticipated input sources, and the process used today by which input would be transformed into acceptable output.
4. Identify duties that prove to be the most significant administrative bottlenecks in the clinical trial life cycle across all clinical trial types and phases.
5. Federal entities should collaborate with industry and academic institutions to develop open-source easily implementable automated solutions to provide clinical research entities with.

Chapter 8: Conclusion

The clinical research process is seen by both industry and government as outdated and in dire need of modernization. Failing to do so will delay the release of medical products that could significantly improve quality of life for patients, and in some cases save a patient's life. The regulatory and legal complexity of overhauling the clinical research pipeline translates to

a significant amount of time passing before the process can be modernized to account for the increasing number of clinical trials entering the pipeline.

Clinical research administrators throughout the clinical research lifecycle are familiar with their unique part of the pipeline and could facilitate the modernization process at a grass-roots level by applying principles and best practices from other professional fields to their work (namely, GCDMP best practices). Applying these practices will enable clinical research administrators to leverage the data flowing around them everyday in enabling leadership to make informed decisions as well as automating various processes necessary for a clinical trial to continue through the pipeline in its lifecycle. Should enough clinical research administrators apply the best practices and methods for leveraging clinical research administration data set forth in this paper, the modernization of the clinical trials process may be completed faster than thought possible.

Bibliography

Al-Jundi, A., Sakka, S. (2016). *Protocol Writing in Clinical Research*. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5198475/

Division of Microbiology and Infectious Diseases. (n.a.). *Microbiology and Infectious Diseases Clinical Research - Safety Oversight*. Retrieved from: https://www.niaid.nih.gov/research/safety-oversight-clinical-research

The Society of Clinical Research Associates. (n.a.). *SOCRA. For Clinical Research Excellence.* Retrieved from: https://www.socra.org/about/clinical-research-education-certification-networking-opportunities/

Bian, J. et al. (2014). *CLARA: an integrated clinical research administration system.* Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4173184/

Moffitt Cancer Center. (n.a.). *Clinical Trials Administration*. Retrieved from: https://moffitt.org/clinical-trials-research/clinical-trials/clinical-trials-administration/

Association for Academic Health Centers. (n.a.). *Association for Academic Health Centers*. Retrieved from: http://www.aahcdc.org/

Society of Research Administrators International. (n.a.). *Clinical Trials Research Administration 101*. Retrieved from: https://www.srainternational.org/learn-0/certificate-programs-0/ctra101-cert

Rush University. (n.a.). *Clinical Research Administration*. Retrieved from: https://www.rushu.rush.edu/research/office-research-affairs/clinical-research-administration

Society for Clinical Data Management. (n.a.). *Good Clinical Data Management Practice.* Retrieved from: https://www.scdm.org/publications/gcdmp/

ICH. (2016). *ICH Good Clinical Practices (E6)*. Retrieved from: https://ichgcp.net/

Division of Microbiology and Infectious Diseases. (2017). DMID Safety Information Sheet. From: https://www.dmidcroms.com/Shared%20Documents/Safety%20Oversight%20and%20SAE%20Reporting%20Info%20Sheet.pdf

Food & Drug Administration. (2006). Guidance for Clinical Trial Sponsors. From: https://www.fda.gov/downloads/regulatoryinformation/guidances/ucm127073.pdf

Smuck, B. et al. (2010). Ontario Protocol Assessment Level, Clinical Trial Complexity Rating Tool. doi: 10.1200/JOP.2010.000051.

Malikova, M.A. (2016). Optimization of Protocol Design: A Path to Efficient, Lower Cost Clinical Trial Execution. doi: 10.4155/fso.15.89

Regan,B., (2017). Can Technology Foster Better Collaboration? Applied Clinical Trials. from: http://www.appliedclinicaltrialsonline.com/can-technology-foster-better-collaboration

Dilts, D., Sandler, A.B. (2006). Invisible Barriers to Clinical Trials. doi: 10.1200/JCO.2005.05.0104

Getz, K., Campo, R. (2017). Trial Watch, Trends in clinical trial design complexity. doi: 10.1038/nrd.2017.65.

Getz et al. (2011). Measuring the Incidence, Causes, and Repercussions of Protocol Amendments. doi: 10.1177/009286151104500307

Dilts et al. (2010). Phase III Clinical Trial Development: A Process of Chutes and Ladders. doi: 10.1158/1078-0432.CCR-10-1273

Getz, K., Campo, R. (2017) New Benchmarks Characterizing Growth in Protocol Design Complexity. doi: 10.1177/2168479017713039

National Institutes of Health. (2018). ClinicalTrials.gov Trends, Charts, and Maps. From https://clinicaltrials.gov/ct2/resources/trends

Appendix A. – IDMC Candidate Invitation Generator

An example program utilizing word processor templates and spreadsheet data sources to generate IDMC invitations and associated documents automatically.

```
import pandas as pd
from mailmerge import MailMerge
from datetime import date
import win32com.client as win32
from datetime import datetime, timedelta
import email_templates
path1 = 'C:/path/to/data/input/'
path2 = 'C:/path/to/template/input/'
outpath = 'C:/path/to/attachment/output/folder/'

protocol = []
frst_names = []
lst_names = []
temp_type = []
name_role = []
i=0
while 1:
    i+=1
    name=input("Enter candidate %d's protocol#, first name, last name, role, and prior/noprior-mid/start : " %i)
    if name=='':
        break
    info = name.split(',')
    frst_names.append(info[1].lower())
    lst_names.append(info[2].lower())
    name_role.append(info[1:])
    protocol.append(info[0])
    protocol = list(set(protocol))

d0 = {'Committee_Type': ['SMC', 'DSMB'],
      'Committee': ['Safety Monitoring Committee (SMC)', 'Data and Safety Monitoring Board (DSMB)'],
      'Sentence': ['an SMC','a DSMB']}
df0 = pd.DataFrame(data=d0)
df01 = pd.DataFrame(name_role, columns=['First_Name','Last_Name','Role','doc_temp'])

df01 = df01.replace({'Role': {'Chair': 'the Chair', 'Member':'a Member', 'Biostatistician':'the Biostatistician'}})
df01['etype'] = df01['doc_temp'].str.split('-',expand=True)[0]
df01['name_key'] = df01['First_Name'] + ' ' + df01['Last_Name']
df01['name_key'] = df01['name_key'].str.lower().apply(lambda n: n[:3] + '.' + n[n.find(' '):])
df01 = df01[['name_key','Role','doc_temp','etype']]
xls1 = pd.ExcelFile(path1+'CMSReport-welcomepackage-14-09-18 13-36.xlsx')
df1 = xls1.parse(xls1.sheet_names[0])
df1.fillna('',inplace=True)
df1['first_key'] = df1['First_Name'].str.lower().apply(lambda n: n[:3] + '.')
df1['name_key'] = df1['first_key'] + ' ' + df1['Last_Name'].str.lower()
df02 = df01.merge(df1, on='name_key', how='left')
```

```
df02['combined'] = df02.astype(str).add('_').sum(axis=1).str[:-1]
df02['count'] = df02['combined'].str.len()
df02 = df02.sort_values('count', ascending=False).drop_duplicates('name_key').sort_index()
df02 = df02[['First_Name', 'Last_Name','Contact_Credentials', 'Title', 'Organization', 'First_Address_Line',
             'Second_Address_Line', 'City', 'Contact_State', 'Postal_Code','Email_Address1', 'Email_Address2',
             'Work_Telephone', 'Phone','Fax_Number1', 'Fax_Number2', 'Assistant_First_Name','Assistant_Last_Name',
             'Alernate_Phone', 'Assistant_Fax','Assistant_Email','Role', 'doc_temp', 'etype']]
df02.reset_index(inplace=True)
df02.drop(['index'],axis=1,inplace=True)
xls2 = pd.ExcelFile(path1+'CMSReport-prots-20-09-18 10-08.xlsx')
df2 = xls2.parse(xls2.sheet_names[0])
df2 = df2[df2['Protocol_Number'].str.lower().isin(protocol)]
df2.fillna('',inplace=True)
df2.reset_index(inplace=True)
df2.drop(['index'],axis=1,inplace=True)
df02['joincol'] = 1
df2['joincol'] = 1
df3 = pd.merge(left=df02,right=df2, on='joincol', how='outer')
df3.drop(['joincol'],axis=1,inplace=True)
df4 = df3.merge(df0,on='Committee_Type',how='left')
df4['letter'] = df4['Protocol_Number']+' '+df4['Committee_Type']+' Welcome Letter - '+df4['Last_Name']+'.docx'
df4['cif'] = df4['Protocol_Number']+' Contact Information Form - '+df4['Last_Name']+'.docx'
df4['subj'] = 'Safety Oversight, Protocol '+df4['Protocol_Number']+', '+df4['Committee_Type']+' Membership Invitation - '+df4['Last_Name']
contacts = df4.to_dict('records')

email_types = {'prior': email_templates.prior, 'noprior': email_templates.noprior}
outpath = 'C:/path/to/attachment/output/folder/'
cif = path2+'Template Contact Information Form.docx'
for d in contacts:
    welcome_letter = path2+'Template Welcome Letter - {}.docx'.format(d['doc_temp'])
    doc_letter = MailMerge(welcome_letter)
    welcome_email = email_types[d['etype']]
    doc_cif = MailMerge(cif)
    name1 = outpath+d['letter']
    name2 = outpath+d['cif']
    print('Package complete for',d['Last_Name'])
    doc_letter.merge_pages([d])
    doc_cif.merge_pages([d])
    doc_letter.write(name1)
    doc_cif.write(name2)
    doc_letter.close()
    doc_cif.close()
    outlook = win32.Dispatch('outlook.application')
    recipient = d['Last_Name']
    prot = d['Protocol_Number']
    titl = d['Protocol_Full_Title']
    cmt = d['Committee']
    cmt_abrv  = d['Committee_Type']
```

```
    due_date = datetime.now() + timedelta(days=7)
    due_date_formatted = due_date.strftime('%d-%b-%y')
    mail = outlook.CreateItem(0)
    mail.To = d['Email_Address1']
    mail.cc = 'SOCS@dmidcroms.com; '+d['Assistant_Email']
    mail.Subject = d['subj']
    mail.Attachments.Add(outpath+d['letter'])
            mail.Attachments.Add(outpath+d['cif'])
            mail.HtmlBody = welcome_email.format(prot=prot,titl=titl,recipient=recip
ient,cmt=cmt,due_date_formatted=due_date_formatted,cmt_abrv=cmt_abrv)
            mail.Display(False)
```

Appendix B. – Automatic Protocol Deviation Form Extraction Program

Example implementation of data extraction from table style forms using Microsoft Excel Visual Basic scripting.

```
Sub NewPDNFExtractCells()

    Dim wb As Workbook
    Dim ws As Worksheet
    Dim MySheet As String
    Dim r1 As Range
    Dim r2 As Range
    Dim r3 As Range
    Dim r4 As Range
    Dim r5 As Range
    Dim r6 As Range
    Dim r7 As Range
    Dim r8 As Range
    Dim r9 As Range
    Dim r10 As Range
    Dim r11 As Range
    Dim r12 As Range
    Dim r13 As Range
    Dim r14 As Range
    Dim r15 As Range
    Dim i As Integer

    Dim OpenWorkbook As Workbook
    Dim OpenWorksheet As Worksheet
    Dim SheetName As String

    Dim Directory As String
    Dim FileSpec As String
    Dim MyFile As String

    Directory = "C:\example\path\to\table\formated\forms\"
    FileSpec = ".xlsx"
    MyFile = Dir(Directory & "*" & FileSpec)
    SheetName = "Table 1"

    ' set local vars
    Set wb = ThisWorkbook
    MySheet = "Sheet1"
```

```
    Set ws = wb.Worksheets(MySheet)


    Set r1 = ws.Range("A2")
    Set r2 = ws.Range("B2")
    Set r3 = ws.Range("D2")
    Set r4 = ws.Range("E2")
    Set r5 = ws.Range("F2")
    Set r6 = ws.Range("I2")
    Set r7 = ws.Range("K2")
    Set r8 = ws.Range("L2")
    Set r9 = ws.Range("M2")
    Set r10 = ws.Range("N2")
    Set r11 = ws.Range("O2")
    Set r12 = ws.Range("P2")
    Set r13 = ws.Range("Q2")
    Set r14 = ws.Range("R2")
    Set r15 = ws.Range("S2")
    i = 0
    Do While MyFile <> ""
        Set OpenWorkbook = Application.Workbooks.Open(Filename:="C:\example\path\to
table\formatted\Forms\" & MyFile, ReadOnly:=True)
        Set OpenWorksheet = OpenWorkbook.Worksheets(SheetName)
        With OpenWorksheet
            r1.Offset(i, 0).Value = .Range("C6").Value
            r2.Offset(i, 0).Value = .Range("C8").Value
            r3.Offset(i, 0).Value = .Range("C4").Value
            r4.Offset(i, 0).Value = .Range("C10").Value
            r5.Offset(i, 0).Value = .Range("C22").Value
            r6.Offset(i, 0).Value = .Range("C12").Value
            r7.Offset(i, 0).Value = .Range("C15").Value
            r8.Offset(i, 0).Value = .Range("C16").Value
            r9.Offset(i, 0).Value = .Range("C35").Value
            r10.Offset(i, 0).Value = .Range("C37").Value
            r11.Offset(i, 0).Value = .Range("C38").Value
            r12.Offset(i, 0).Value = .Range("C36").Value
            r13.Offset(i, 0).Value = .Range("C17").Value
            r14.Offset(i, 0).Value = .Range("C18").Value
        End With
        i = i + 1
        OpenWorkbook.Close SaveChanges:=False
        MyFile = Dir()
    Loop
End Sub
```

Appendix C. – Automated Report Subset and Email Generator

Generates a subset report from a larger report source based on date ranges. Generates a formatted email with the subset report attached, ready to send to relevant stakeholders.

```
import pandas as pd
import numpy as np
import glob
import datetime as dt
import win32com.client as win32
```

```
pd.options.mode.chained_assignment = None
pd.set_option('display.max_colwidth', -1)
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

today = dt.date.today()
last_week = today - dt.timedelta(days=7)
tomorrow = today + dt.timedelta(days=1)

path = 'C:/path/to/main/report/source'

for mtracker_name in glob.glob(path+'Document Inventory Tracker*'):
    mtracker_file = mtracker_name
df = pd.read_excel(mtracker_file,header=None,usecols="A:O,W")
df[15].replace('Other', np.nan, regex=True, inplace = True)
df[15].replace('Date Updated', np.nan, regex=True, inplace = True)
df[15] = pd.to_datetime(df[15])
df1 = df[(df[15] >= last_week) &  (df[15] < tomorrow)]
df0 = df.iloc[:2]
all_dfs = [df0,df1]
df2 = pd.concat(all_dfs).reset_index(drop=True)
del df2[15]
# df2.drop(15,axis=1,inplace=True)
df2.fillna('',inplace=True)
df2[3].iloc[2:] = pd.to_datetime(df2[3].iloc[2:]).dt.strftime('%m/%d/%y')
df2.iloc[:,5:] = df2.iloc[:,5:].replace(1, 'Posted', regex=True)
df2.rename(columns={ df.columns[4]: "type" },inplace=True)
df2.rename(columns={ df.columns[6]: "m1" },inplace=True)
df2.rename(columns={ df.columns[7]: "m2" },inplace=True)
df2.rename(columns={ df.columns[9]: "m3" },inplace=True)
df2.rename(columns={ df.columns[10]: "m4" },inplace=True)
df2.rename(columns={ df.columns[12]: "m5" },inplace=True)
df2.rename(columns={ df.columns[13]: "m6" },inplace=True)
df2.loc[df2['type'] == 'E-REV', ['m1','m2','m3','m4','m5','m6']] = "N/A"
out_path = 'C:/path/to/report/output/folder/Study Tracker Report
Update {date:%d%b%y}.xlsx'.format(date=dt.datetime.now())
writer = pd.ExcelWriter(out_path , engine='xlsxwriter')
workbook = writer.book
df2.to_excel(writer, sheet_name='Sheet1',header=False,index=False)
worksheet = writer.sheets['Sheet1']
formater = workbook.add_format({'bold':True})
merge_format = workbook.add_format({
    'align': 'center',
    'text_wrap': True,
    'border': 1,
    'valign': 'vcenter'})
main_format = workbook.add_format({
    'align': 'left',
    'text_wrap': True,
    'border': 1,
    'valign': 'bottom'})
blank_format = workbook.add_format({
    'bg_color': '#fff2cc',
    'border': 1
})
worksheet.merge_range('A1:C1', 'Protocol Info', merge_format)
worksheet.merge_range('D1:E1', 'Meeting Info', merge_format)
```

```
worksheet.merge_range('F1:H1', 'M', merge_format)
worksheet.merge_range('I1:K1', 'CMS', merge_format)
worksheet.merge_range('L1:N1', 'DL', merge_format)
worksheet.set_column('A2:O2',28,merge_format)
worksheet.set_column('A:B',10)
worksheet.set_column('C:C',11)
worksheet.set_column('D:P',9)
worksheet.conditional_format('A1:O2',{'type':'no_blanks','format':formater})
worksheet.conditional_format('A3:O'+str(df2.shape[1]+2), { 'type' : 'no_blanks' , 'f
ormat' : main_format} )
worksheet.conditional_format('A3:O'+str(df2.shape[0]), { 'type' : 'blanks' , 'format
' : blank_format} )
writer.save()
m_email = (r"""
<body>
<font face = "Times New Roman" size="3">
<p>Dear Mr. Smith,</p>
<p>Per your request, please find attached this week's Study
Tracker Status Update.</p>
<p>If you have any questions or concerns, please contact me directly
at vpoonai@exampledomain.com or 301-444-5555.</p>
<p>Sincerely,<br>
Victor Poonai, BA<br>
Safety Oversight Coordinator<br>
Safety Oversight Support Team<br>
</font></body>
""")
outlook = win32.Dispatch('outlook.application')
mail = outlook.CreateItem(0)
mail.To = 'johnsmith@mail.nih.gov'
mail.Subject = 'Study Tracker Weekly Status Update ({date:%d%b%y})'.format(date=dt.d
atetime.now())
mail.Attachments.Add(out_path)
mail.HtmlBody = m_email
mail
```

Author Biography

Victor Poonai is a data scientist and clinical research administration professional. He has provided his expertise in process automation, data analysis, and clinical operations to Phase I-III clinical trials under various NIH divisions and offices for over seven years. He has experience in a broad range of clinical research activities which include serving as scientific lead for a project assessing the impact of protocol development policy changes on the clinical trial lifecycle, assessing recruitment method efficacy for multi-site studies, designing artificial intelligence powered automated solutions for clinical research administration activities, and identifying and convening independent data monitoring committee member candidates for clinical trials