

Insights informing strategies for optimizing the collection of clinical trial data

Kenneth Getz

kenneth.getz@tufts.edu

Tufts University School of Medicine

Emily Botto

Tufts University School of Medicine

Ana Calduch Arques

Novo Nordisk, Capital Region of Denmark

Laura Galuchie

Merck & Co., Inc

Natalia Camargo Sanmiguel

Merck & Co., Inc

Nicole Sheetz

Gilead Sciences Inc

Zak Smith

Tufts University School of Medicine

Research Article

Keywords: protocol design, protocol complexity, protocol data, clinical research data, clinical trial design, clinical protocol, protocol endpoints

Posted Date: September 14th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7527216/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Insights informing strategies for optimizing the collection of clinical trial data

Kenneth Getz* ^a, Emily Botto ^b, Ana Caldusch Arques ^c, Laura Galuchie ^d, Natalia Camargo Sanmiguel ^e, Nicole Sheetz ^f, Zak Smith ^g.

Affiliations

^a Tufts Center for the Study of Drug Development, Tufts University School of Medicine, Boston, MA, USA

^b Tufts Center for the Study of Drug Development, Tufts University School of Medicine, Boston, MA, USA

^c Novo Nordisk, Copenhagen, Capital Region of Denmark, Denmark

^d Merck & Co., Inc., Rahway, NJ, USA

^e Merck & Co., Inc., Rahway, NJ, USA

^f Gilead Sciences Inc., Foster City, CA, USA

^g Tufts Center for the Study of Drug Development, Tufts University School of Medicine, Boston, MA, USA

Abstract

Although past research has quantified the proportion and types of procedures that support clinical trial endpoints, little is known about the volume and nature of data collected by these procedures and their impact on participant and site burden. Tufts CSDD and 15 TransCelerate Biopharma sponsor companies convened to collect data from 105 multi-therapeutic protocols with a primary completion date after 2018. Overall data volume by study and data volume by participant were analyzed and procedures were categorized into core, standard and non-core based on the endpoints they supported. Study results show that total data volume continues to grow with 5.9 million datapoints now collected on average per phase III protocol, up 11% annually since 2020. Non-core procedures comprise 17.8% of total phase II and 16.2% of total phase III procedures per protocol, a downward trend from the proportion of non-core procedures observed in 2020. Nearly half (46%) of phase II and 35% of phase III non-core procedures gather data for exploratory and future use purposes. On average, 6.6% and 12.6% of procedures supporting core and standard endpoints from phase II and phase III protocols are deemed non-essential (i.e., procedures determined by the clinical team or protocol authors as being performed in excess of the number of times required to demonstrate a clinical outcome), and their associated data represents 8.2% to 17.1% of the total datapoints collected. Combined, nearly one-third of all procedures and datapoints collected per protocol is classified as non-core or are non-essential procedures supporting core standard/required endpoints with more than half of this data associated with clinical and patient-reported questionnaires. As much as 30% of participant and site burden is associated with noncore procedures or are non-essential procedures supporting core, standard/required endpoints. The implications of these results and potential strategies to optimize the collection of protocol data are discussed in this analysis.

*Corresponding Author (kenneth.getz@tufts.edu)

Keywords: protocol design, protocol complexity, protocol data, clinical research data, clinical trial design, clinical protocol, protocol endpoints

Introduction

Clinical trial protocols are arguably the most intensive and extensive plans prepared by clinical research professionals in pharmaceutical and biotechnology companies. As the blueprint directing the administration of study requirements, procedures and assessments; the screening, randomization and enrollment of study participants; and the collection of clinical research data in support of study objectives, endpoints and the clinical development plan; the protocol represents the culmination of an unprecedented number of inputs and intent from clinical teams, operations and data management functions, external collaborators and senior leadership. Finalization of a late-stage clinical protocol, for example, requires dozens of clinical team meetings, an average of seven internal protocol review meetings, and input from an average of four external collaborators (e.g., scientific advisors, key opinion leaders, regulatory consultants and patient advocacy groups), over the course of several months and up to half a year.ⁱ

Protocol design is carefully considered and highly intentional. Based on the maturity of the compound, indication, treatment and available scientific evidence, the trial design balances study objectives, prioritizes participant safety, avoids unnecessary design and executional complexities, while ensuring scientifically sound methodology. Recently released regulatory guidelines and technical requirements harmonized under international standards (ICH) aim to improve the quality of clinical trials.^{ii,iii}

Protocols accommodate a growing number of objectives and endpoints^{iv}. Phase III protocols, for example, have an average of 19 total endpoints at database lock, a 30% increase over the average observed 15 years ago^v. During this same period, the number of procedures performed per participant visit has risen 37% in step with the growth in the number of endpoints, and the number of countries and sites engaged per protocol have increased 44% and 63% respectively^{vi}. Protocols are also collecting a substantially higher volume of clinical research data. Studies conducted by the Tufts Center for the Study of Drug Development (Tufts CSDD) show that between 2012 and 2020, the average total number clinical research datapoints collected in phase III protocols increased nearly four-fold, from 929,203 to 3,560,201^{vii}.

Several factors have contributed to the increasing requirements, scope and data volume of clinical trial protocols. Global drug development pipelines are targeting more challenging disease conditions – including more narrowly defined and rare disease patient communities — requiring more assessments and data to demonstrate treatment effects. Historical requirements from global regulatory and oversight agencies have increased the volume of safety, quality and compliance data collected. Additionally, growing numbers of procedures are conducted, and data is collected to support the product label and reimbursement^{viii}.

As protocol requirements, scope and data volume have increased, clinical trial speed and efficiency have worsened. This inverse relationship has been consistently shown during the past two decades: protocols with a higher relative number of endpoints, procedures, countries and sites have longer mean enrollment durations, higher incidence and mean number of protocol deviations and amendments, and lower participant enrollment and completion rates^{ix}.

Most clinical research professionals have found – and empirical evidence shows – protocols collect a large amount of data which may have strategic value and inform clinical teams but is not relevant to support assessment of primary and key secondary endpoints demonstrating efficacy and safety^x. A 2012 study conducted by Tufts CSDD indicated that 25% of procedures performed per phase III protocol and 18% of those performed per phase II protocol supported supplementary secondary, tertiary and exploratory (i.e., non-core) endpoints^{xi}. A subsequent 2020 study by Tufts CSDD saw a proportional increase in non-core procedures in phase II.

Research suggests that rising data volume and the collection of tertiary and exploratory data not only contributes to clinical trial costs but also jeopardizes study quality and places considerable burden on sites and participants. Nahm, Pieper and Cunningham, for example, found that higher data volume adversely impacts data quality due to higher error rates ^{xii}. Research by Getz et al. demonstrated that rising protocol complexity and scope had corresponding levels of participant and site burden resulting in longer clinical trial timelines, higher levels of inefficiency and poorer recruitment and retention performance ^{xiii,xiv}.

Procedures and their associated data were classified as ‘non-core’ by the clinical teams and protocol authors from sponsor companies in the 2012 and 2020 Tufts CSDD studies. Additionally, a 2023 case study asked reviewers at the Food and Drug Administration (FDA) to classify protocol procedures based on the endpoints they supported. FDA reviewers classified a higher percentage of procedures as extraneous or ‘non-core’ (26% vs 18%) and they deemed that half the data collected was ‘Not Relevant’ to informing the clinical meaningfulness of outcomes for patients^{xv}.

Taken together, the collective body of evidence suggests a compelling need and opportunity to optimize the volume of protocol data collected and its ensuing relationship with site and participant burden. Although past research has quantified the proportion and types of non-core procedures, little is known about the volume of data collected, reasons for collection, how it is used, and its impact on participant and site burden.

Recognizing this need, TransCelerate BioPharma, Inc. and Tufts CSDD conducted a collaborative research study in 2024 and early 2025 to shed light on the underlying drivers of clinical trial data volume and to inform internal discussions and strategies to optimize protocol designs. Several major goals framed this collaborative study including: collection of the most relevant and necessary data to answer the key scientific objectives and questions; data to inform primary and key secondary endpoints; use of sound

research methodologies; compliance with regulatory requirements; and minimization of site burden and participant burden.

This TransCelerate BioPharma, Inc. – Tufts CSDD collaborative study is very timely. The recently finalized ICH E6 R3 guidelines encourage more fit-for-purpose trial design and data acquisition to support flexible protocol design and execution that optimizes the volume and nature of data collected while improving participant convenience and access with appropriate quality management.^{xvi}

Methods

TransCelerate Biopharma, Inc. and Tufts CSDD convened a working group of 15 sponsor companies — each a TransCelerate member — in April 2024. The primary aims of the working group were to characterize the volume, reason and usage of data from non-core procedures as well as to assess procedures that are conducted more often than necessary (i.e., non-essential) to support core, standard and required endpoints. To our knowledge, this study is the first to collect and analyze data on ‘non-essential’ procedures.

Working group members provided extensive input during development of the data collection instrument including which protocol variables to collect and how to prioritize them (whether the variables were “need to have” or “nice to have”). When variables of interest included metrics that have not been or rarely collected in the past, working group members provided input into how the variables may be collected and formatted, or if necessary, what appropriate proxies may be used.

Among the variables collected were

- background characteristics such as phase, therapeutic area, molecule type (small molecule or biologic), orphan designation (orphan or non-orphan),
- participant population (patients or healthy volunteers),
- route of administration,
- protocol design (traditional or master),
- adaptive design (yes or no), and
- whether health authority feedback was received during protocol design.

Data on several scientific and operational characteristics were also collected:

- number of countries,
- number of sites,
- number of participants screened,
- number of participants enrolled,
- number of eligibility criteria,
- number of study arms,
- total number of procedures performed (for an average participant),

- number of unique procedures included,
- number of datapoints per participant,
- number of endpoints,
- number of adverse events of special interest.

Working group members were asked to provide additional information on each procedure performed including its core, required, standard, or non-core classification, who was responsible for performing the procedure (PI or Specialist, Other Site Staff, or Patient Reported), and the minimum number of times the procedure was required to be performed by the highest (most important) endpoint it supported.

Working group members collectively agreed on the following sampling frame to guide the collection of protocols that would be analyzed in this study:

- Retrospective protocols with a database lock (or primary completion date) in 2018 or later.
- Phase II and Phase III protocols.
- Excludes pediatric and device trials.
- Target of 10 or more protocols for each company.
- Therapeutic areas included in each company's sample should be representative of their respective portfolios.

Within each working group member company, clinical teams, protocol authors or others coded all of the variables in their protocols. Several important definitions were used to guide the data coding and collection activity. Procedure classifications were defined as follows:

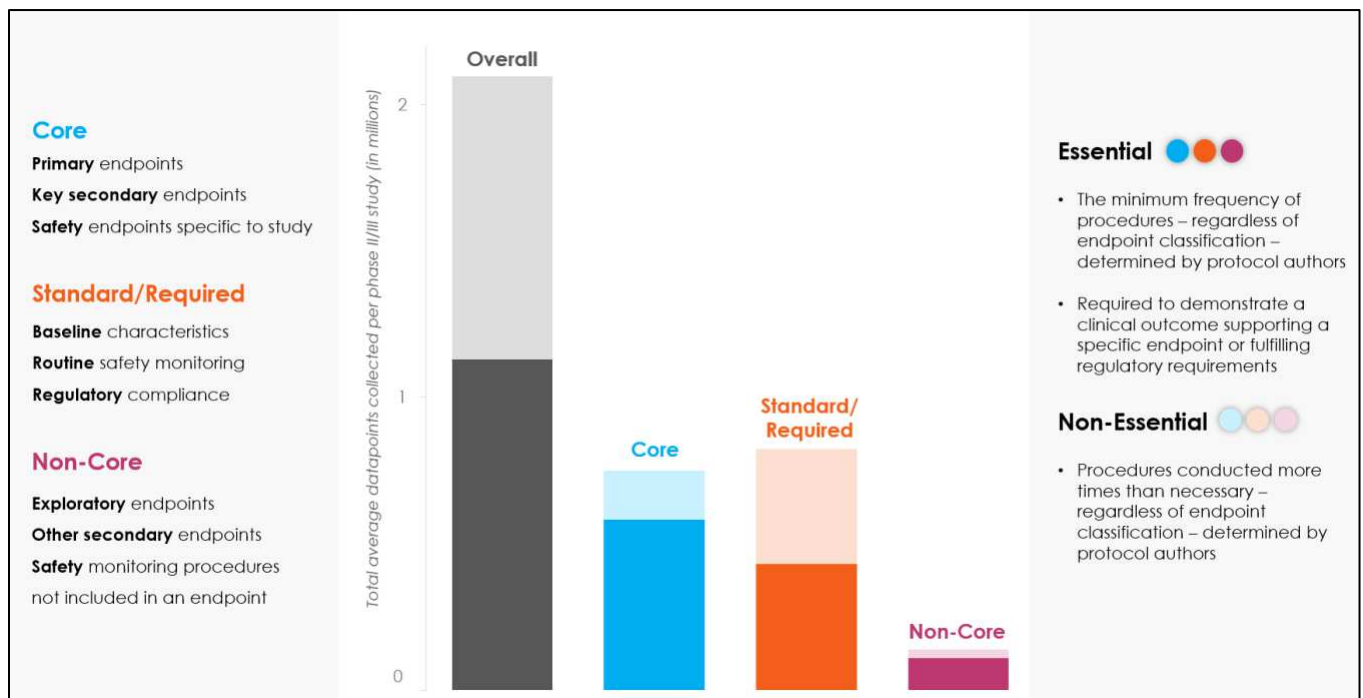
- Core Procedures –
 - Procedures supporting primary and/or secondary objectives
 - Procedures supporting primary, key secondary, and safety endpoints
- Required Procedures –
 - Procedures supporting screening requirements
 - Procedures supporting the informed consent process
 - Procedures supporting drug dispensing and compliance
- Standard Procedures –
 - Routine procedures including gathering data on baseline health, concomitant medications, demographics, adverse events, and adverse drug reactions
- Non-core Procedures –
 - Procedures supporting tertiary and exploratory objectives and endpoints
 - Procedures supporting supplementary (all non-key) secondary endpoints
 - Safety and efficacy procedures that are not included as an endpoint or objective

- Procedures not tied to an endpoint or objective¹

A hierarchy was used for procedures that supported more than one endpoint type. For example, if a procedure supported both a primary and tertiary endpoint, that procedure was coded based on the highest endpoint supported — in this case as ‘core’. Standard and required procedures were also considered higher in the hierarchy than non-core procedures. In all cases, when a procedure supported multiple procedure classifications, the highest type was noted.

Finally, working group member companies assessed the frequency of each procedure and determined the minimum number of times it could be conducted to address its highest associated clinical endpoint, requirement or standard. Any number of times the procedure was performed more than this amount was deemed ‘non-essential’. The following graphic describes the procedural subgroups that comprise the total data collected by a given protocol for the purposes of this analysis.

Figure 1. Procedures Generating Protocol Data by Endpoint Category



¹ Nothing in this hierarchy is intended to imply that data collection related to a tertiary or exploratory objective or endpoint, or procedures not tied to an endpoint or objective are without value or purpose to clinical research or clinical trial participants.

Category	All Data		
	Core	Standard & Required	Non-Core
Definition	CORE Procedures supporting: <ul style="list-style-type: none"> Primary endpoints Key Secondary endpoints Safety endpoints specific to study <p>These procedures always create clinical data.</p>	STANDARD Procedures supporting: <ul style="list-style-type: none"> Baseline characteristics Routine safety monitoring REQUIRED Procedures supporting: <ul style="list-style-type: none"> Regulatory compliance 	NON-CORE Procedures supporting: <ul style="list-style-type: none"> Exploratory endpoints Other secondary endpoints Safety monitoring procedures not included in an endpoint <p>These procedures do not always create clinical data.</p>
Purpose Supported	<ul style="list-style-type: none"> Primary endpoints Key secondary endpoints Focused safety 	<ul style="list-style-type: none"> Good Clinical Practice (GCP) compliance Standard Safety Population characteristics 	<ul style="list-style-type: none"> Data supporting access and reimbursement Data supporting publications Future research Academic interests
Examples	<ul style="list-style-type: none"> Disease outcomes assessment <ul style="list-style-type: none"> Tumor/lesion assessment Specific analyte level Viral load Pharmacokinetics (PK), anti-drug antibodies, biomarkers (for primary or secondary endpoint) Survival data Medical events specific to study (e.g., adverse events of special interest) Specific safety procedures Specific clinical outcomes assessments 	<ul style="list-style-type: none"> Informed consent form Randomization information Visit information Screening requirements Drug compliance Adverse events Medical history Concomitant medications Demographics Disease diagnosis Vital signs (e.g., blood pressure, pulse rate) Laboratory rates 	<ul style="list-style-type: none"> Clinical outcomes assessments PK, anti-drug antibodies, biomarkers (not in primary or secondary endpoint) Contraceptive check Routine safety monitoring (e.g., physical examinations, temperature, respiratory rate)
Essential	<ul style="list-style-type: none"> The minimum frequency of procedures – regardless of endpoint classification determined by protocol authors Required to demonstrate a clinical outcome supporting a specific endpoint or fulfilling regulatory requirements 		
Non-Essential	<ul style="list-style-type: none"> Procedures conducted more times than necessary – regardless of endpoint classification – determined by protocol authors 		

Total clinical trial data volume was calculated using two separate methodologies. For the primary method, working group members were asked to report the number of datapoints collected each time individual procedures were performed, and these datapoints were classified according to the endpoint supported by the corresponding procedure. Total data volume for trials was then calculated by multiplying the number of datapoints collected per procedure by the average number of times each procedure was conducted for each participant over the course of the clinical trial. Data volume was calculated both per participant and for the overall trial, where the number of datapoints per participant was multiplied by the number of participants in the trial. For the alternative method, working group members reported the total datapoints collected per participant over the duration of each trial. These datapoints were separated into those that were auto collected (directly into a database without work from the site, such as wearable devices); collected by the site (site staff collected and entered into database); or collected by a secondary source (collected and entered on site's behalf by another source

such as a central laboratory; or where the site had access to the data). Although it was anticipated that both methodologies would result in similar data volume estimates, this was not the case. Therefore, throughout these analyses, the primary methodology was used unless specifically stated.

For the assessment of data usage, working group members indicated whether data collected for each procedure was included in the CSR (Clinical Study Report), and whether it was analyzed and/or discussed, or merely referenced. It was also noted whether the procedure data was part of the main body or appendix of the CSR.

The data collection instrument underwent several rounds of review and revision through the spring of 2024. Working group members then piloted the data collection instrument in summer 2024 by providing data on 1 or 2 protocols. After the pilot, adjustments to the workbook were made, and a finalized workbook was distributed. Remaining protocol data was provided between July and November 2024. Following a comprehensive QC process to ensure high data accuracy, validity, and completeness, the final database was locked in January 2025.

Descriptive statistics (counts, percentages, means, coefficients of variation, medians, and interquartile ranges) were calculated for the dataset overall and for important subgroups. Due to differences inherent in trial design and in order to simplify comparisons between therapeutic area subgroups, protocols targeting cancer-related illnesses were assigned the 'oncology' subgroup, all others were assigned the 'non-oncology' subgroup.

Tufts CSDD was able to calculate participant and site procedural burden estimates using data collected in previous but unrelated Tufts CSDD studies. These studies collected perceived burden scores on a scale from 0 – 100 for approximately 60 procedures commonly conducted during clinical trials. Perceived burden scores were collected from both participants and site staff (e.g., principal investigators, study coordinators). The procedures were grouped into several larger categories – medication administration, non-invasive, invasive, lab and blood, questionnaire, routine, imaging, diary, and other. Working group members were given a list of example procedures that fall within each procedure type based on the previous research and were instructed to use these as a guideline for categorization of the procedure in each protocol. Mean burden scores were calculated for each procedure type category using the procedures included in each group. Individual procedures from the current study were mapped to the same procedure type categories and assigned the corresponding mean participant and site burden scores. Participant burden scores were further adjusted to account for procedures which could be performed at home by discounting the burden of these procedures by 10%, as has been done in previous participant burden assessments^{xvii}. Participant and site burden score estimates were the sum of the procedural burden scores for the total number of times each procedure was performed. Total procedural burden estimates could be separated into essential and non-essential core, and essential and non-essential standard/required performances of procedures, and subgroups.

All data were collected and stored in Excel workbooks and analyzed using a statistical analysis program.

Results

Fourteen working group member companies coded and collected data on 105 protocols with primary completion dates after 2018. Table 1 provides detailed characteristics of the study data. A majority (59%) are phase III protocols, with a predominance of small molecule chemical entities represented (63%). The study data contains a mix of therapeutic areas, proportionally representative of the broad global drug development pipeline. One-quarter (26%) of the investigational treatments analyzed in the dataset target oncology indications while the remaining were non-oncology. Indications with less than 5 protocols in the dataset were grouped into “other” for blinding purposes. The study data also contains a mix of investigational therapies by route of administration with 38% oral therapies, 29% intravenous and 26% subcutaneous.

Table 1. Data Characteristics

	Number of Protocols	Percent of Total
Total N	105	100%
Phase II	43	41%
Phase III	62	59%
Small Molecule Chemical Entities	67	63%
Biologics	38	37%
Oncology	27	26%
Endocrinology	17	16%
Immunology	14	13%
Infectious Diseases	12	11%
Neurology	9	9%
Respiratory	7	7%
Other (e.g., Cardiology, GI, Hematology)	19	18%
Oral	40	38%
Intravenous	30	29%
Subcutaneous	27	26%
Other (e.g., intramuscular, ocular)	8	8%

On average 17.8% of phase II procedures and 16.2% of phase III procedures were classified as non-core. Between 35.2% and 38.8% of procedures were identified as standard or required and 43.3% and 47.1% of phase II and phase III procedures as core (see Table 2).

The incidence of non-core procedures is higher among non-oncology protocols at 19% of the total. Coefficients of variation around the mean values are also presented. Coefficients that approach and exceed 1 indicate that the individual protocol values are widely dispersed, indicating more variability between protocols of the same phase, indication and class.

Table 2. Distribution of Procedures by Endpoint Classification per Protocol, Phase and TA

	Phase II	Phase III		Oncology	Non-Oncology
Core	43.3%	47.1%		38.8%	48.0%
(CoV)	(0.67)	(0.59)		(0.63)	(0.61)
Standard/Required	38.8%	35.2%		49.5%	31.9%
(CoV)	(0.72)	(0.72)		(0.50)	(0.79)
Non-Core	17.8%	16.2%		11.2%	19.0%
(CoV)	(0.98)	(0.99)		(1.26)	(0.90)

As Figure 2 shows, the proportion of non-core procedures has declined since last measured in 2020. Phase II protocols have seen the proportion of non-core procedures drop from 26.8% to 17.8%; phase III protocols from 22.7% down to 16.2%.

Figure 2. Trends in the Distribution of Procedures per Protocol by Endpoint Classification

[To show as a segmented bar chart]

	Phase II				Phase III		
	2012	2020	2025		2012	2020	2025
Core	55.6%	47.1%	43.3%		46.0%	50.8%	47.1%
Standard/Required	31.3%	26.1%	38.8%		30.4%	26.5%	35.2%
Non-Core	13.1%	26.8%	17.8%		23.6%	22.7%	16.2%

Table 3 presents company-reported reasons for conducting non-core procedures. The most common reasons for both phase II and phase III studies were to gather safety, future use and exploratory data. When combined, 46% of phase II and 35% of phase III non-core procedures were conducted for exploratory and future use purposes. The most common reasons for conducting non-core procedures differed when assessed by therapy area group rather than phase. For oncology studies, the most common reasons were for exploratory (43%) and future use (23%) purposes. For non-oncology studies, safety (23.5%) and future use (19.8%) were the most common reasons reported.

Table 3. Reasons for Conducting Non-Core Procedures

	Phase II	Phase III		Oncology	Non-Oncology
Safety	19.7%	21.4%		10.9%	23.5%
Future Use	23.7%	18.8%		23.4%	19.8%
Exploratory	21.9%	15.9%		43.1%	11.3%
Quality of Life/ Symptom Evaluation/PRO	5.3%	13.3%		2.9%	12.5%
Efficacy	5.7%	10.8%		10.2%	8.7%
Market Differentiation	9.6%	5.8%		2.9%	8.3%
Support Reimbursement	4.8%	5.1%		1.5%	5.9%
Regulatory Request	0.4%	3.1%		1.5%	2.4%

PK/PD	2.2%	1.2%		1.5%	1.6%
Screening	0.9%	0.0%		0.0%	0.4%
Other	5.7%	4.6%		2.2%	5.7%

Compared to benchmarks from previous studies, the total data volume per phase III protocol has risen 10.8% annually since 2020 (see Figure 3, calculated using second data volume calculation method). Between 2012 and 2025 the volume of data collected per protocol has increased 15.4% annually.

Figure 3. Growth in overall average data volume collected per phase III protocol

[Show as a histogram or line chart]

2012	929,203
2020	3,560,201
2025	5,957,122

Table 4 presents the distribution of datapoints collected per participant per protocol by procedures classification. The majority (61.2% and 61.5%) of total phase II and phase III data collected per participant supports core procedures. Only 5.2% of datapoints gathered per participant per protocol in oncology, and 15.7% of datapoints in non-oncology support non-core procedures.

Table 4. Distribution of Datapoints Collected per Participant per Protocol

	Phase II	Phase III		Oncology	Non-Oncology
Mean Total Datapoints	2,824	3,512		3,302	3,224
(CoV)	(0.87)	(0.69)		(0.66)	(0.80)
Median	1,770	2,596		2,163	2,181
IQR	1,291 – 3,969	1,771 – 5,278		1,716 – 4,311	1,274 – 4,638
Core	61.2%	61.5%		54.2%	64.4%
Standard/Required	22.7%	27.1%		40.6%	19.9%
Non-Core	16.1%	11.4%		5.2%	15.7%

When examined by phase, between 6.6% and 12.6% of core and standard/required procedures are deemed non-essential; and 8.2% to 17.1% of datapoints collected is from non-essential procedures (see Table 5). Combined, nearly one-third of all procedures and data collected per phase III protocol is classified as non-core or non-essential; 24.9% to 27.3% of phase II procedures and data is classified as non-core or non-essential combined. Almost four-out-of-ten (36.5%) datapoints collected per participant in non-oncology phase III protocols are classified as non-core or non-essential core, standard/required combined.

Table 5. Proportion of Total Procedures and Datapoints per Participant per Protocol Classified as Non-Core or Non-Essential

	Phase II	Phase III		Oncology	Non-Oncology
Non-Essential Standard/Required Procedures	6.6%	12.6%		11.4%	9.9%
(CoV)	(1.85)	(1.16)		(1.16)	(1.44)
Non-Essential Standard/Required Datapoints per Participant	8.2%	17.1%		11.8%	14.9%
(CoV)	(1.70)	(1.10)		(1.10)	(1.29)
Non-Core and Non-Essential Core, Standard/Required Procedures (Combined)	24.9%	31.3%		20.4%	30.6%
(CoV)	(0.87)	(0.70)		(0.85)	(0.73)
Non-Core and Non-Essential Core, Standard/Required Datapoints per Participant (Combined)	27.3%	32.5%		14.9%	36.5%
(CoV)	(1.15)	(0.58)		(0.95)	(0.65)

Data from non-core procedures and non-essential procedures is concentrated around specific procedure types (see Table 6). Questionnaires have the highest proportion of non-core and non-essential data, totalling 55.4% of data collected by questionnaires. A high relative percentage of data collected by patient diaries is also non-core or non-essential (41%), with most of this attributed to non-core procedures.

Invasive procedures, lab and bloodwork, imaging procedures and non-invasive procedures also collect a high relative proportion of total data from non-core and non-essential procedures.

Table 6. Proportion of Total Datapoints Collected from Non-Core or Non-Essential Core, Standard/Required Procedures per Protocol

(Phase II and III combined) Procedure Type	Percent of Total Procedures within Type that are Non-Core (Column B)	Percent of Total Procedures within Type that are Non-Essential Core or Standard/Required (Column C)	Total (Sum of Columns B + C)
Questionnaires	37.1%	18.3%	55.4%
Patient Diaries	37.5%	3.5%	41.0%
Invasive Procedures	25.9%	7.1%	33.0%
Lab & Blood	16.0%	13.4%	29.4%
Non-Invasive Procedures	10.0%	17.1%	27.1%
Imaging	14.9%	6.1%	21.0%
Routine Procedures	4.5%	11.9%	16.4%

Other	12.1%	1.9%	14.0%
Medication Dispensing	0.0%	3.1%	3.1%

Between 25% and 30% of participant burden is attributed to non-core and non-essential core, standard/required procedures with the higher relative burden found in non-oncology phase II and III protocols (see Table 7). A similar proportion of site burden is attributed to non-core and non-essential core, standard/required procedures. Non-core and non-essential core, standard/required procedures in phase II and III protocols in non-oncology contribute the highest relative burden for sites.

Table 7. Proportional Contribution to Participant and Site Burden by Endpoint Classification per Protocol, by Phase and TA

	Phase II	Phase III		Oncology	Non-Oncology
PARTICIPANT BURDEN					
Core procedures	39.6%	44.4%		36.0%	44.8%
Standard/Required Procedures	40.6%	37.2%		51.8%	33.7%
Non-Essential Core + Standard/Required Procedures	5.9%	12.7%		10.9%	9.7%
Non-Core Procedures	19.8%	16.7%		11.5%	20.3%
SITE BURDEN					
Core Procedures	41.1%	47.3%		38.9%	46.9%
Standard/Required Procedures	40.7%	34.6%		49.6%	32.5%
Non-Essential Core + Standard/Required Procedures	7.0%	12.9%		11.3%	10.3%
Non-Core Procedures	18.1%	16.6%		11.0%	19.5%

Similar to findings in earlier Tufts CSDD studies, most data collected (77%) which support non-core procedures ultimately appear in the Clinical Study Report (CSR). Nearly half of the data is discussed and 15.3% is referenced in the main body of the CSR. An additional 14% of data collected from non-core procedures will be either discussed, referenced or included in the CSR's appendix.

Table 8. Data Usage in the Clinical Study Report (CSR)

	Main Body		Appendix			TOTAL
	Discussed	Reference	Discussed	Reference	Included but not Discussed or Referenced	
Core	58.9%	16.7%	2.3%	12.8%	1.2%	92%
Standard/Required	45.0%	14.8%	0.8%	11.7%	4.7%	77%
Non-Core	46.5%	15.3%	1.5%	8.0%	2.9%	74%

Figure 4 indicates the top reported reasons why data from non-core procedures were not used in the CSR. The majority — 56% — of the data was exploratory in nature (28.3%) or intended for future use (27.7%); 14% is safety data that was collected but not core to the study.

Figure 4. Top Reasons Non-Core Procedure Data was Not Included in the CSR

[Present as a histogram in descending order]

Exploratory	28.3%
Future Use	27.7%
Safety	13.8%
Market Differentiation	7.5%
QoL, Symptom Eval, PRO	7.5%

Discussion

The results of this study provide valuable updated benchmarks as well as new baseline measures that companies can use to compare against their own experience. The study found that protocol data volume has continued to rise since the last analysis performed five years ago. Each phase III protocol gathers, on average, 5.9 million datapoints, up 67% since 2020. This trend is due to many factors including the growing specialization of participant segmentation, rising use of biomarkers to target therapies, increasing complexity and scope of protocol designs, anticipation of questions from regulators and payers, and growing use of digital data collection technologies. Additionally, the processing power of artificial intelligence and machine learning-enabled analytical tools may be serving as a disincentive to reduce data volume. ^{xviii, xix}

Phase II protocols collect a higher average proportion of total data from non-core procedures (16%) compared to phase III protocols (11%). This finding may be due in part to heightened efforts to gather data in earlier phases to support the emerging safety and efficacy profile of the molecule and to inform the decision to enter later-stage clinical trial activity.

The proportion of non-core procedures has declined for both phase II and III protocols since 2020 while the proportion of standard/required procedures has increased. Required procedures (e.g., informed consent, adverse events, etc.) are anchored in ICH and will remain. However, there is an opportunity to re-assess data collection from standard or routine clinical procedures based on the disease, compound, stage of development and other factors.

The results of this study indicate that 7% of phase II and 13% of phase III protocol core, standard/required procedures are conducted more times than minimally necessary to support the highest endpoints. Non-essential core, standard/required procedures contribute 8% of phase II and 17% of phase III datapoints collected. A relatively high percentage of total participant burden and site burden

is associated with non-core and non-essential core, standard/required procedures. Burden contribution from non-core and non-essential core, standard/required procedures is highest in non-oncology.

Taken together, approximately one-third of all phase III and one-quarter of all phase II procedures and data collected are classified as non-core or are considered 'non-essential' to supporting its highest associated clinical endpoint, requirement or standard. The percentage of datapoints per participant that relate to non-core procedures or are deemed non-essential is highest in non-oncology where the combined proportion is nearly 40%. Procedures frequently classified as non-core and associated with a high volume of non-essential data collection represent key areas to optimize. Questionnaires were the most common procedure in this category. They are commonly used to assess quality-of-life impact of the treatment studied and are required in some countries. The number and frequency of assessments should be balanced with their purpose and burden. Other procedures in this category include patient diaries, invasive procedures, lab and bloodwork, imaging and non-invasive procedures. These procedures all had high relative proportions of total data collected that supported non-core and non-essential core, standard/required procedures. Clinical teams can discuss, prioritize and weigh the impact of keeping these identified procedures (e.g., impact on enrollment timelines, site and participant burden) against not keeping them, or adjusting the frequency of the procedure.

The results of this study found that most non-core data collected (77%) ultimately appeared in the Clinical Study Report (CSR). This is not surprising as ICH E3 provides clear guidance on content. Data collected for exploratory and future use purposes was the most likely not to be included in the CSR and represents an opportunity to delay or reduce data collection activity. Non-essential and non-core data that does not inform primary and key secondary endpoints and that may ultimately not be used in the clinical study report represents another optimization opportunity.

This study had a number of limitations. Participating companies faced considerable difficulty coding, accessing and compiling the workbook data. As a result, some areas of analysis had relatively small sample sizes. Comparisons between results from the current and past studies should be viewed with some caution as the datasets contained different proportions of protocols by phase, therapeutic area, and protocol scope. In addition, participant and site burden are largely a function of the type and volume of protocol procedures. Future research will look at burden more expansively to include non-procedural factors (e.g., distance required to travel to the site, duration of each study visit) to better assess the impact of activities associated to non-core procedures. Furthermore, the classification and usage of individual data points was beyond the scope of this study since the evaluation was conducted at the procedure level. There may be additional opportunities for optimizing data collection through a more detailed assessment of the data points needed for each individual procedure. Future research will also look to conduct more granular assessments of procedures and data supporting non-core and non-essential core, required/standard procedures by individual therapeutic areas.

The findings from this TransCelerate Biopharma, Inc. and Tufts CSDD collaborative study offer valuable benchmark and baseline metrics and new insights into opportunities to optimize the dramatically rising

volume of protocol data collected. While the study focused on Phase II and III trials, the implications extend across the entire clinical development journey, reinforcing the importance of designing studies whose data needs are complementary and aligned throughout a program. The results of this study not only inform the development of deliverables by TransCelerate Biopharma, Inc. to help sponsors consider how to balance scientific objectives, operational feasibility, site and participant burden, and regulatory requirements, but also support other organizations in evaluating and evolving their own internal strategies and practices. Importantly, this paper does not advocate for including or removing specific procedures. Rather, it underscores the importance of mindful data collection and the identification of areas that may contribute to the burden. Clinical trials are not broken; however, there are clear opportunities to improve what data is collected, evaluated, and used. The intent is not to eliminate all "non-core" or "non-essential" procedures, but to ensure their inclusion is carefully considered deliberate. As the complexity and volume of clinical trial data continue to grow, the time is now for industry-wide reflection and action: stakeholders across the ecosystem must critically evaluate the value and purpose of every data point collected to reduce stakeholder burden and drive more efficient, patient-centric research.

Acknowledgements

The authors thank the 15 TransCelerate BioPharma member companies that contributed data and insights to this study. While the individual companies remain anonymized to protect confidentiality, their participation was essential in enabling a robust and comprehensive analysis. For more information about TransCelerate Biopharma, its membership and mission: <https://www.transceleratebiopharmainc.com/>

TransCelerate also thanks Yumi Sugiura (BMS) for subject matter expertise on ICH guidelines. And a special thank you is extended to all workstream participants for their valuable contributions to the data analysis study, whether helping shape the study or supporting data contributions.

Author Contributions

KG contributed to all four aspects (substantial contribution to the conception, design, analysis, interpretation; drafting and revising the work; final approval of the version to be published; agreement to be accountable for all aspects in ensuring the accuracy and integrity of the work. EB contributed to all four aspects. ACA contributed to all four aspects. LG contributed to all four aspects. NCS contributed to all four aspects. NS contributed to all four aspects. ZS contributed to all four aspects.

Declarations

Conflict of Interests

Ana Calduch Arques, Novo Nordisk, declares that she is an employee and has financial holdings in the company; Laura Galuchie, Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, declares that she is an employee and has financial holdings in Merck & Co., Inc., Rahway, NJ, USA; Natalia Camargo Sanmiguel, Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA, declares that she is an employee and has financial holdings in Merck & Co., Inc., Rahway, NJ, USA;

Nicole Sheetz, Gilead Sciences Inc., declares that she is an employee and has financial holdings in the company.

-
- ⁱ Kaitin K editor. Rising protocol design complexity driving rapid growth in data volume. *Impact Report*. 2021; 23(1).
- ⁱⁱ International Council for Harmonisation. ICH Harmonised Guideline: Guideline for Good Clinical Practice E6(R3). 2025. https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf
- ⁱⁱⁱ International Council for Harmonisation. ICH Harmonised Guideline: General Considerations for Clinical Studies E8(R1). 2021. https://database.ich.org/sites/default/files/E8-R1_Guideline_Step4_2021_1006.pdf
- ^{iv} Getz K, Smith Z, Kravet M. Protocol design and performance benchmarks by phase, and by oncology and rare disease subgroups. *Ther Innov Regu Sci*. 2023; 57: 49-56.
- ^v Kaitin K editor. Rising protocol design complexity driving rapid growth in data volume. *Impact Report*. 2021; 23(1).
- ^{vi} Getz K, Smith Z, Kravet M. Protocol design and performance benchmarks by phase and by oncology and rare disease subgroups. *TIRS*. 2023; 57(1): 49-56.
- ^{vii} Getz K, Smith Z, Kravet M. Protocol design and performance benchmarks by phase, and by oncology and rare disease subgroups. *Ther Innov Regu Sci*. 2023; 57: 49-56.
- ^{viii} Gnanasakthy A, Levy C, Norcross L, Doward L, Winnette R. A review of labeling based on patient-reported outcome endpoints for new oncology drugs approved by the European Medicines Agency. *Value Health*. 2023; 26(6): 893-901.
- ^{ix} Getz K, Smith Z, Kravet M. Protocol design and performance benchmarks by phase, and by oncology and rare disease subgroups. *Ther Innov Regu Sci*. 2023; 57: 49-56.
- ^x Profit A. Dealing with data over-collection in clinical trials. *Clinical Research News*. 2024. <https://www.clinicalresearchnews.com/news/2024/12/10/dealing-with-data-over-collection-in-clinical-trials>
- ^{xi} Getz K, Stergiopoulos S, Marlborough M et al. Quantifying the magnitude and cost of collecting extraneous protocol data. *Am Journal of Thera*. 2015; 22(2):117-124.
- ^{xii} Nahm M, Pieper C, Cunningham M. Quantifying data quality for clinical trials using electronic data capture. *PLOS One*. 2008; 3(8): e 3049. Doi:10.1371/journal.pone.003049.
- ^{xiii} Smith Z, Botto E, Getz K. Enhancing the Measure of Participation Burden in Protocol Design to Incorporate Logistics, Lifestyle, and Demographic Characteristics. *TIRS*, 2021
- ^{xiv} Florez M, Smith Z, Getz K. Quantifying Site Burden to Optimize Protocol Performance. *TIRS*, 2024
- ^{xv} Smith Z, Getz K. A case study assessment on the rationale for, and relevance of, non-core protocol data. *Ther Innov Regu Sci*. 2024; 58(2): 311-315.
- ^{xvi} International Council for Harmonisation. ICH Harmonised Guideline: Guideline for Good Clinical Practice E6(R3). 2025. https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf
- ^{xvii} Smith Z, Wilkinson M, Carney C, Grove N, Qutab B, Getz K. Enhancing the measure of participation burden in protocol design to incorporate logistics, lifestyle and demographic characteristics. *TIRS*. 2021.
- ^{xviii} Madanchia M, Tahendost H. The impact of artificial intelligence on research efficiency. *Results in Engineering*. 2025: [dot.org/10.1016/j.rineng.2025.104743](https://doi.org/10.1016/j.rineng.2025.104743).
- ^{xix} Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, Abdullah S. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. *Biomedical Materials and Devices*. 2023; 8: 1-8.