Synthetic Data and Predictive Data Modeling: Benefits in Driving Standards and Quality Outcomes

W. Stetson Line, Clinventive LLC

INTRODUCTION

The clinical trials professional community is constantly seeking to bring new therapies to the point of a 'fill or kill' decision and potentially a regulatory submission in the most efficient manner possible. Standards, budget, study design, regulatory compliance, and the goodwill of patients and partnered clinicians are all critical success factors. Of these It could be argued that one of the most important factors is time. Why? Simply put, unexpected delays will negatively impact and increase risks to all other aspects of the trial. This is especially true if competition for new markets or 'first in human' status is in play. The key to avoiding needless delays during the conduct of a trial is to ensure the highest operational quality and fitness for study objectives in the data collection, review, and analysis systems before the study 'goes live'. With so much at stake quality control and testing are rigorously employed. Too often though such testing fails to fully vet all the potential issues in the study design and implementation. This is due to the challenges in obtaining test data in sufficient quality and quantity.

To most creating test data is an expensive and tedious task. There is ample incentive to obtain only the bare minimum amount in the most expedient way possible. For example, one commonly employed approach is to create a few records of data for each case report form in order to satisfy individual test cases. Such data is sometimes manually entered by a programmer or quality control analyst. At the other end of the spectrum entire sets of subject case report forms are 'mocked up' and sent to data entry services to be entered into a test instance of the study. But this approach can be limited by the expertise of those authoring the data and the care given in ensuring internal consistency. Another common method is to copy data from a similar trial, manually plugging the gaps where any design aspects differ. To varying degrees all these methods produce test data compromised in scope, utility and fitness for identifying potential issues. Why? Because such data is often fragmented in significant ways that violate the integrity naturally present in real data. There is also an opportunity cost in that such data is limited in its potential to serve other downstream use cases that could add significant value to other organizational quality assurance goals.

Consider though, what if instead of restricting test data to serving specific test cases the goal was to create a holistic study data model? What if it were possible to efficiently create synthetic study data as precise, rich, and accurate as real data? If synthetic study subjects aligned with assessments, events, and outcomes dynamically in a manner consistent with the trial design and chronology? In that case, how could synthetic data inform trial design decisions? In what ways could this data accelerate, or even automate data acquisition and qualification processes? This paper will seek to address these questions by first examining aspects that often differentiate test data from real data. This will help illustrate both the challenges and the benefits of producing synthetic data in a holistically integrated fashion. It will explore techniques that can generate synthetic data ideal for establishing quality, both inexpensively and in any desired quantity. Additionally, how holistic synthetic data 'pays it forward' to benefit the qualification of downstream processes in clinical trials data acquisition, analysis, and submission will be considered.

TEST DATA THAT FAILS

To aid in understanding the value of such synthetic data it is important to consider various 'axes of integrity' in real study data. These axes are markers of implicit congruity in study data that are a natural consequence of real subjects participating in a study protocol within a shared epoch in time. In clinical trials data quality systems are in place to identify where such expected markers of integrity are missing or violated.

The following table (Figure 1) is a simplified version of what is commonly referred to as a schedule of assessments or evaluations found in clinical study protocols. A protocol documents how the study is to be conducted, guides data collection, and explains the study rationale and objectives among other things. The table is called a schedule because it lays out across the x axis the timeline for when a subject will be expected to visit the clinician's office or study site to undergo various tests or evaluations. The visit labels are represented as column headers where common abbreviations such as 'Screen', 'Wk 1', '3 Mo' denote the shortened form for the visit name. The column footnote documents the study day or number of days elapsed from the start of the study timeline for the individual subject. The row headers describe the various tests, evaluations or questionnaires that can be administered at a clinical visit and usually correspond the shortened form of a case report form, the collection format for the assessment. Any cell marked with an 'X' indicates that the assessment on the cell row is planned for the visit on the cell column.

Figure 1: Simplified Schedule of Assessments Example

Assessment	*Screen*	*Day 1*	*Wk 1*	*Wk 2*	*Wk 4*	*Wk 8*	*Wk 12*	*3 Mo F/U*
Informed Consent	X							
Inclusion/Exclusion Criteria	X							
Demographics	X							
Medical History	X							
Adverse Events		X	X	X	X	X	X	
Concomitant Medication	X	X	X	X	X	X	X	
Physical Exam	X							
Vital Signs	X	X	X	X	X	X	X	X
IP Admin		X	X	X	X	X		
Chemistry	X		X		X	X	X	
Hematology	X		X		X	X	X	
ECG	X		X		X	X		
End of Study							X	
Study Day	*0*	*1*	*7*	*14*	*28*	*56*	*84*	*168*

The example study has a planned length of three months plus a follow up visit three months after the end of study. Not all subjects begin the study on the same day but are enrolled over a short period of time i.e. the 'enrollment period'. Most subjects are expected to complete all visits and associated assessments. Inevitably some subjects will fail the screening process and others will be withdrawn early from the study for various reasons. In a real study such enrollment factors are carefully planned for so that data will be sufficiently populated in most assessment forms and their corresponding datasets to enable analyses from which conclusions can be made with a high degree of confidence. The sufficiency of data to enable reliable analyses is termed the study's 'power'.

One could apply this concept of 'power' in a similar way to test data. One could ask: Is the test data sufficient to enable reliable conclusions to be drawn about the quality of its implementation? Are the test datasets populated with enough subjects and records to gauge for example, system performance under load, the suitability of medical review listings, cross-form and log form edits, and downstream activities related to analysis dataset creation and the testing of analysis tables, listings, and figures? Often the answer to these questions is 'No'. Figure 2 below illustrates why this is the case. In using a traditional approach to User Acceptance Testing (UAT) of electronic data collection the team has drafted test cases. In each test case data has been scripted to address a particular 'dirty' or 'clean' scenario where an assessment form is completed for a particular visit and the expectation is either that the system will behave normally (clean case) or handle the dirty data in an appropriate manner. The team has decided to assign each test case to a different test subject in order to be able to efficiently track the results. In Figure 2 a

different color has been given to each test subject to illustrate how the data was populated in the schedule of assessments after the UAT is complete. In this simplified example it is assumed that all test subjects completed the first three assessments.

Figure 2: Simplified Schedule of Assessments Example UAT Data Coverage

Assessment	*Screen*	*Day 1*	*Wk 1*	*Wk 2*	*Wk 4*	*Wk 8*	*Wk 12*	*3 Mo F/U*
Informed Consent	X							
Inclusion/Exclusion Criteria	X							
Demographics	X							
Medical History	X							
Adverse Events		X	X	X	X	X	X	
Concomitant Medication	X	X	X	X	X	X	X	
Physical Exam	X							
Vital Signs	X	X	X	X	X	X	X	X
IP Admin		X	X	X	X	X		
Chemistry	X		X		X	X	X	
Hematology	X		X		X	X	X	
ECG	X		X		X	X		
End of Study							X	
Study Day	*0*	*1*	*7*	*14*	*28*	*56*	*84*	*168*

As we see in this somewhat exaggerated example the data entered for the UAT is intentionally sparse with many forms not being used for data entry at all. Further no test subject completed all the forms. This was deemed unnecessary by the test case authors because the scope of the test cases did not require it. At the end of the UAT the resulting test database was so fragmented and limited as to be unusable for any other downstream qualification activities.

Would completing all the forms for several test subjects address these limitations? Not necessarily. When data is 'mocked up' out of thin air it is very difficult to adhere closely to the many axes of data integrity that characterize real data. There are many markers that characterize well-formed clinical data, in fact, too many to consider in this brief discussion. So, to narrow our scope only five axes of integrity related to temporal congruity in well-formed clinical data will be cited and illustrated by example. Let's examine how manually created test data can violate various temporal axes of integrity:

> *Cohort Temporal Alignment*: All subjects should be part of a group that begins their participation within the enrollment period and completes or withdraws from the study before the end of the enrollment period plus the length of the trial. In the example schedule if the enrollment period were set at two months then the last subject data collection should be complete by the date that is equal to the enrollment start date plus the enrollment period plus six months. A lack of temporal alignment in test data subjects may not have a great impact on reports or analyses that aggregate data by study day. However, this issue would compromise testing for monitoring and clinical trial conduct reports that tie to actual dates.
>
> *Visit Sequence and Chronological Agreement*: Visits take place according in a sequential manner that aligns with a chronological order, i.e. the second visit takes place after the first visit and the visit dates support this natural order. Test data that does not enforce visit sequence and chronological agreement will present problems for most reports and analyses, e.g. plots, patient profiles, changes from baseline.

Visit Assessments Synchronicity: All assessments assigned to a visit should have an associated date that falls within a contemporaneous target window i.e. within a day or two of the planned day for that visit. Test data that does not have this internal consistency will frustrate testing of cross form edits and can make certain outputs lack coherence e.g. graphs, patient profiles.

Study Data Collection Assumptions: On-study assessments should have dates of data collection before study completion but after study enrollment. Historical data domains should have event or medication start dates before the data of enrollment.

Event Duration Assumptions: All records that record a duration should have an end point of the duration that is greater than or equal to the starting point if the end point is non-missing.

Would copying real data from a similar study resolve such issues with internal consistency? Yes. To the extent that the prior study design matches the new study data could be leveraged in this way. But this approach poses its own problems. For example: How to best identify and 'patch' test data for forms and fields that do not match? How to obtain and use the real data without compromising patient confidentiality or unmasking restricted or potentially unblinding data? It also may be very difficult to identify the study that is the 'best fit' for harvesting test data. This, of course, is an argument in favor of having and using data standards and for leveraging a global library to manage the standards' metadata. Curating a standard or model study with its own associated test data has advantages and this is considered by many a best practice. However, this may still be insufficient in quantity and prone to the problem areas identified in our discussion so far. Fortunately, there is a better way.

STRATEGIES FOR SYNTHETIC STUDY DATA GENERATION

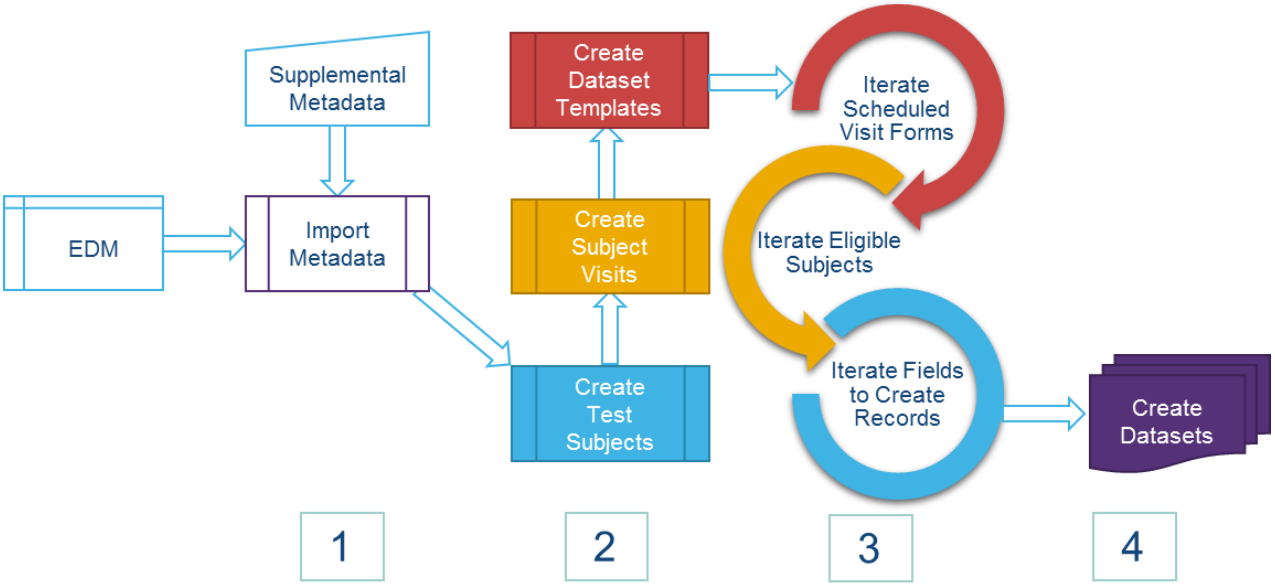
A high-quality synthetic study is populated with test subjects that 'experience' the schedule of assessments in the same manner as that expected of real subjects, and therefor satisfies all the axes of data integrity discussed earlier. The test subjects can withdraw from study early and/or participate in sub-studies and other dynamic study design features e.g. cohorts, treatment arms, based on 'events' or triggers in their test data. In a synthetic study there is no functional limit to the number of sites, subjects or data records that can be generated. To those repeatedly frustrated in their pursuit of quality test data this probably sounds too good to be true. Conceptually it is very difficult to understand how such a synthetic study could be produced. In the next section the framework for successfully building and using synthetic study data will be explained. It is hoped the discussion will help simplify the problem set into workable components and illustrate the value of this approach.

Metadata is data about data, in other words, descriptive information about how data should be collected, qualified, and stored. A familiar example of metadata would be the domain specifications found in published CDISC standards documents. In clinical trials Electronic Data Capture (EDC) systems are built using electronic data specifications called study design metadata. This metadata specifies the study schedule of assessments, visits, forms, fields, associated code lists and other controlled terminology. It may also include lab textbook ranges, derivations, unit conversion factors, and quality edit checks. Clinical Trial Management systems, Medical Coding and Thesaurus software, and external data transfer processes also use metadata in one form or another.

Leveraging these various types of design metadata is the key to producing synthetic study data efficiently. This descriptive information about how real data should be collected, qualified, and stored can be 'reverse engineered' to direct test data production. That said, there are factors that apply to test data that do not apply to real data and vice-versa. For example, in the production of synthetic data edit checks or rules are prescriptive, not proscriptive, i.e. they define what should be created, and importantly how it should be created, whereas normal edits define erroneous or insufficient data. This distinction is one of many that can require supplementary metadata to enable accurate implementation.

The following graphic is a high-level process flow for creating synthetic study data. The process is broken down into four phases, numbered in sequential order at the bottom of the graphic. The objectives and requirements for each phase will be discussed and a rationale for the approach used provided.

Figure 3: Synthetic Data Generation Process



Phase 1 - Metadata Acquisition: Data standards, in a large part driven by CDISC standards, are today more the rule than the exception. Global standards libraries supported by electronic metadata repositories are likewise gaining adoption. EDC systems commonly support authoring new studies in an off-line manner using electronic design metadata (EDM) that is exported and imported as needed. These developments have made EDM suitable for generating test data more widely and easily obtainable.

Supplemental metadata is used to augment the acquired EDM in order to fill-in 'gaps' with an end goal to make the synthetic data more realistic. For example, in a real study subjects' medical conditions and treatments are recorded during clinical interviews and then standardized through a coding process using medical dictionaries. These medical dictionaries may be available for use in generating test data but how would one determine what medical conditions or terms out of the thousands available would be the most likely to occur in the subject population under study? Supplemental 'tailored' medical dictionaries can be provided to address this gap. Categorical or discrete data fields usually have associated code lists that enumerate the acceptable values for each field. Supplemental metadata can be used to provide expected values for fields that do not have associated code lists and/or provide guidance for the expected frequency distribution of the values in those lists. Continuous numeric data fields may not have guidance for a generally accepted range of values, decimal precision, or standard units of measure. Supplemental metadata can fill in the gaps to guide creation of reasonable data values where needed. Creating this metadata may seem a difficult task, and it does require careful consideration. Keep in mind though, that the majority of this 'setup' cost is an investment that can be reused. Data standards metadata repositories have tools to identify where study designs differ, and these can simplify the task of updating supplemental metadata for new study-specific implementations.

Phase 2 – Synthetic Study Set-up: As is the case in a real study, a synthetic study should be thoughtfully set-up prior to the 'go-live' of data creation. The recommended first step is to consider the desirable attributes of the test subjects that will populate the synthetic study. How many test sites and subjects are needed? What are the study enrollment criteria for age and sex if any? When will the study start? How long is the target enrollment period? How long is the on-study duration? What site and subject identification conventions should be used? These are questions that will inform the creation of a test subject roster that mirrors the planned study population. It is also critical to set-up the test subjects with enrollment dates that are aligned with study start and enrollment period assumptions in order to prevent some of the internal integrity issues discussed earlier.

An optional but recommended step is to create an inventory of planned subject visits. This is the assignment of the study schedule of assessments to each test subject saved in the form of a normalized dataset. The advantage of this approach is that it enables the customization of each test subject's journey through the synthetic study. Do several test subjects withdraw from the study early? Their visit inventory can be reduced accordingly to reflect this. Do other subjects participate in an optional study arm or sub-study? Their visit inventory can be augmented appropriately.

Another important step is to create template datasets or domains to store the test data. This can be done using the data specification metadata that normally defines the field names, data types, field lengths and formats, any associated permissible value code lists, and the field labels. Once populated with test data these template datasets can be written to a desired location or used to support other data structures and use cases.

Most studies have libraries of approved and expected terminology that is to be used for categorical data collection in various fields. These can range from extensive lists e.g. questions and response options in questionnaires, to a single value e.g. 'Y'. The expected terminology code lists or dictionaries need to be electronically loaded and available for use in the test data generation process. The data specifications specify the name of the code list used for a given field where applicable. This code list name indicates the terminology requirements for the field and is a link to the code list library. Some code lists have values that are in themselves links to other code lists. This type of relationship is used in defining 'value-level' terminology requirements. Value-level metadata can be used to drive complex test data creation. However, this is a topic best left for a 'deep-dive' discussion.

Phase 3 – Synthetic Study 'Conduct': A clinical trial is conducted over a planned study length by investigators who administer planned assessments to active subjects at prescribed intervals or visits. A synthetic study can be 'conducted' in a similar manner. To elaborate, study visit assessments are administered as scheduled in chronological order. Later visits can be impacted by data collected at earlier visits. For example, a marker or symptom of a condition may prompt further testing for an individual subject or may require that subject withdraw from the study. A synthetic study can enforce chronological integrity by creating data in the same order as expected in the real study. This approach affords the opportunity to introduce dynamic elements into the data creation process, where the subject visit inventory can be adjusted based on certain key 'events' occurring in the data. Obviously, the data creation engine must incorporate elements of randomness in its algorithms for this approach to have value.

In Figure 3 there are three circular glyphs that denote types of iteration in this third phase. More accurately these are nested not sequential iterations or loops. Plainly stated the test data is created by 'looping' through the schedule of assessments in chronological visit order, then stepping through every assessment that is assigned to that visit. Next the roster of test subjects that have that have that visit/assessment in their inventory are identified. For each qualifying subject a visit date is determined by using their individual enrollment date and incrementing it by the number of days equal to that visit's planned 'study day'. The data creation 'engine' then iterates through the design specifications for each field in the assessment's associated dataset, using field-level metadata to populate the test data appropriately. This process is repeated until each expected synthetic data record is completed and written to its corresponding dataset.

The routines that populate the data take into consideration each field's data type, format, terminology requirements, and any 'rules' that may apply to it. Defining and implementing rules for test data creation is another topic best addressed in a longer form discussion. Suffice it to say that it is possible, with less effort than would be expected, to define complex criteria for the creation of test data, such as: Populating a field only when another field is populated, selecting a value randomly from a code list, creating a random number or date based on a range of values, selecting a value based on another field's value selections, deriving a field's value based on another field or fields' values, and so on. Nearly any real-world data scenario can be simulated using a straightforward set of rules. (See APPENDIX for a simplified example)

Phase 4 – Synthetic 'End-of-Study' Processing

Sourced EDM for the test data may define datasets targeting STDM, ADaM, SEND, or other CDISC standards. It could be sourced from an EDC system or an external data transfer plan. Whatever the case it is likely that the target output format and structure may be different that defined in the available metadata. For example, if the metadata defines unique form variants e.g. 'Vitals Signs at Baseline', 'Vital Signs On-Study', 'Vital Signs at Follow-up', it is likely that the desired result would to coalesce the three forms into a single dataset, aligning the common fields while allowing for conditional population of the variant fields. It could also be that the test data is needed in an alternate format e.g. SAS transport files, CDISC ODM or RDF, or MS Excel. The final stage of the test creation process addresses these remaining requirements.

TEST DATA THAT SUCCEEDS

Congratulations for making it through the dense technical discussion above. One may wonder if the benefits of creating synthetic data in this way is worth the effort. It could be said that the true value of this approach accrues over time and over a wide range of potential use cases. In addition to the more obvious returns of having robust test data there are additional 'hidden gems' that may take some digging to uncover.

A pain point that organizations may encounter when implementing synthetic data is in the acquisition of metadata from sources outside of centralized control. It is often the case that providers of specialty labs or evaluation centers have their own standards for specifying datasets and associated expected values. This lack of standardization in metadata can stymie efficient communication and can require bespoke programs to qualify and on-board the external data transfers. Programmers and data managers may need to translate such bespoke specifications into a useable format, and this not only increases overhead but can become a maintenance nightmare as requirements change. Synthetic data can drive home the value of data standards and, significantly, of *metadata* standards. Even when data can't be standardized, if the metadata follows a rigorous standard then it becomes possible to streamline or even automate certain tasks.

In one instance a sponsor that had several dozen external data sources invested in standardizing the metadata in the templates used to define these sources. Eliminating dependence on outsourced test data, developing qualification tests and reports, review listings, and data transfer processes earlier and more reliably were the initial goals. However, it was only after this exercise was successfully completed that additional opportunities to leverage this metadata became clear. Since the metadata defined the transfer schedule, data variables, expected values or code lists, and data blinding requirements in a globally consistent manner it was possible to deprecate hundreds of bespoke on-boarding and qualification programs. These were replaced by a single global program that automated data on-boarding, conversion, and quality control by using the same metadata as that used to create the synthetic data. The effort and resources conserved were then applied to increasing the scope and depth of the quality checks leading to identification of risks and issues that had previously gone un-noticed. This example illustrates how synthetic data 'rules' can be leveraged to both create test data and to aid the validation of real data. It also demonstrates how prioritizing the organization, standardization, and availability of metadata is a key to unlocking other unanticipated and beneficial use cases.

It is often a race against time setting up a clinical trial. Early qualification of medical review safety data listings is a priority, as these need to be available by the time the first data is collected. This means that using real data to test these is not an option. Another challenge is in preparing the analysis tables, figures, and listings (TFL) for the first Safety Update Report (SUR) or Interim Analysis (IA), when early results are first shared with regulatory agencies. It is ideal to have these reports consume datasets that are conformant to the Analysis Data Model (ADaM) industry standard, and those in turn having been derived from datasets conforming to the Submission Data Tabulation Model (SDTM) standard. This is so because data is required to be in this format when it is ultimately submitted as part of an application for marketing approval. Additionally, TFL standards, a common-sense and worthy goal, are only practical when built on a stable underlying data standard. However due to time constraints sponsors will often create analysis datasets directly from the operational data, postponing the conversion of operational data to SDTM and then to ADaM until there is more time and data available. This is not only a duplication of effort, but also requires repeated reconciliation between various sources to ensure that no data has been corrupted or 'left behind' by the various conversion and derivation programs.

Synthetic data offers a better approach to a consistent and efficient production of TFLs by enabling the production of SDTM and ADaM datasets before the study goes 'live'. In contrast to most data generated in UAT efforts, synthetic study data is holistically produced and as such enforces the axes of internal congruity discussed earlier. This makes it possible to develop and test conversion programs rigorously and with a high degree of confidence in their fitness for purpose. Mapping synthetic data to the next required 'down-stream' standard well in advance of when it is needed for production provides teams with valuable time to develop and quality assure dependent TFL programs and outputs. It also eliminates the need for 'stop-gap' bespoke programs and datasets and supports adoption of TFL standards.

As data and metadata standards are implemented further 'upstream' opportunities to automate aspects of data conversion become more feasible. Synthetic metadata can be augmented to include conversion instructions for the next 'downstream' standard so that, for instance, operational field synthetic metadata can 'map' SDTM field destinations, SDTM field synthetic metadata can 'map' ADaM field destinations, etc. Having a global metadata standard and toolkit for specifying synthetic data means that nearly any kind of data can be modeled effectively. In practice this can free development teams from any dependence on the availability of data from an 'upstream' process. In one case a sponsor modeled and produced synthetic ADaM data for a trial before the EDC system was even in production.

One use case that may be of special interest to the standards community: Synthetic data can be used to evaluate proposed and published standards, where it enables analysis of fitness for purpose and the assessment of likely impacts to the organization's processes and tools. These are important considerations in decisions about if and when such proposed standards should be implemented.

CONCLUSION

Although test data is a necessary evil to many, holistically produced synthetic study data has an enormous potential for good. To realize this potential the prevailing mindset must shift from viewing test data as only serving specific test cases to seeing its utility in a host of downstream processes. This utility is present when synthetic data is precise, rich, and accurate, when synthetic study subjects 'experience' assessments, events, and outcomes dynamically in a manner consistent with the trial design and chronology, conforming to the markers of internal integrity expected of real data. Such robust synthetic data can inform design decisions, and can help accelerate, or even automate data acquisition and qualification processes, including mapping of data to CDISC submission standards. Indeed, this paper only provided a brief outline of the various techniques useful in synthetic data generation and how their application has great potential in advancing clinical trials data acquisition, mapping, analysis, and regulatory submission activities.

APPENDIX: A SIMPLE SYNTHETIC DATA GENERATOR

The following illustrates a very basic approach to synthetic data generation. The code in the example is written in SAS and is given to illustrate the technical feasibility of some of the techniques discussed in this paper. No warranty is made of the code or of the techniques discussed.

Metadata for SDTM v3.2 DM Domain – Includes Supplemental Metadata in Varlogic Variable

```
data dm;
 length varorder logicorder 8 varname $8 varlabel $52 vartype $4 varcodelist $10 varformat $10 varlogic $200;
 infile datalines dlm='|' dsd;
    input varorder varname $ varlabel $ vartype $ varcodelist $ logicorder $ varformat $ varlogic ;
 datalines;
 1|STUDYID|Study Identifier|Char||1|$20|:derive!_study
 2|DOMAIN|Domain Abbreviation|Char||2|$2|:derive!"DM"
 3|USUBJID|Unique Subject Identifier|Char||3|$20|:derive!catt(_study,'-',_site,'-',_subject)
 4|SUBJID|Subject Identifier for the Study|Char||4|$5|:derive!_subject
 5|RFSTDTC|Subject Reference Start Date/Time|Char||5|$25|:derive!put(dhms(_startdt,0,0,time()),e8601dt.)
 6|RFENDTC|Subject Reference End Date/Time|Char||6|$25|:derive!put(dhms(_enddt,0,0,time()),e8601dt.)
 7|DTHDTC|Date/Time of Death|Char||12|$25|:derive!RFENDTC:align!DTHFL="Y"
 8|DTHFL|Subject Death Flag|Char|NY|11|$1|:derive!substr('NNNNNNNNY',rand('integer',1,9),1)
 9|SITEID|Study Site Identifier|Char||13|$5|:derive!_site
 10|AGE|Age|Num||17|3|:derive!rand('integer',18,79)
 11|AGEU|Age Units|Char|AGEU|18|$5|:derive!"YEARS"
 12|SEX|Sex|Char|SEX|19|$3|:derive!substr('MF',rand('integer',1,2),1)
 13|RACE|Race|Char|RACE|20|$30|:derive!scan('ASIAN~WHITE~WHITE~WHITE~WHITE~BLACK OR AFRICAN AMERICAN~ASIAN',rand('integer',1,7),'~')
 14|ARMCD|Planned Arm Code|Char||22|$3|:derive!substr('AB',rand('integer',1,2),1)
 15|ARM|Description of Planned Arm|Char||23|$30|:derive!ifc(ARMCD='A','ARM A','ARM B')
 16|ACTARMCD|Actual Arm Code|Char||24|$3|:derive!substr('AB',rand('integer',1,2),1)
 17|ACTARM|Description of Actual Arm|Char||25|$30|:derive!ifc(ARMCD='A','ARM A','ARM B')
 18|COUNTRY|Country|Char|COUNTRY|26|$30|:derive!"USA"
 run;
```

Set-up Study Parameters and Subject Population Arrays

```
%let sites = 9;
%let subjs = 9;
%let study = 1234;
%let studystart  = 23MAY2016;
%let studymaxday = 300;
%let enrollperiod= 90;
%let subjpattern = 100n;
%let sitepattern = 0100n;

*** -- create array of subjects -- ***;
%do site = 1 %to &sites;
    %let si&site = %substr(&sitepattern,1,%index(&sitepattern,n)-1)&site;
    %do subj = 1 %to &subjs;
        %let su&site._&subj = %substr(&subjpattern,1,%index(&subjpattern,n)-1)&subj;
        %let fd&site._&subj = %eval(%sysfunc(inputn(&studystart,date9.)) + %sysfunc(rand(integer,1,&enrollperiod)));
        %let ld&site._&subj = %eval(&&fd&site._&subj + %sysfunc(rand(integer,100,&studymaxday)));
    %end;
%end;

*** -- create arrays for dataset specification metadata -- ***;
proc sql noprint;
    select count(distinct varname) into :vcnt from &ds where varname > ' ';
    %let vcnt = &vcnt;
    select varname into :fldarray separated by ' ' from &ds where varname > ' ' order by varorder ;
    select varname into :logarray separated by ' ' from &ds where varname > ' ' order by logicorder ;
    select varname,         vartype,            varlabel,
           varformat,       varcodelist,        varlogic
     into :vn1-:vn&vcnt,    :vt1-:vt&vcnt,      :vl1-:vl&vcnt,
          :vf1-:vf&vcnt,    :vc1-:vc&vcnt,      :vi1-:vi&vcnt
     from &ds
     where varname > ' ' order by logicorder;
quit;
```

Create DM Template then Generate Data using Variable and Supplemental Metadata

```
*** -- create template form for dataset and populate data-- ***;
data &ds._test(keep=&fldarray);
format &fldarray;
length _site _subject _study $10 _startdt _enddt tempnum 8 tempchr $200;
%do v = 1 %to &vcnt;
    attrib &&vn&v length=&&vf&v label="&&vl&v";
%end;
_study = "&study";
%do site = 1 %to &sites;
    %do subj = 1 %to &subjs;
        _site    = "&&si&site";
        _subject = "&&su&site._&subj";
        _startdt = &&fd&site._&subj;
        _enddt   = &&ld&site._&subj;
        _visitdt = _startdt + rand('integer',30,300);
        %do v = 1 %to &vcnt;
            %let tags = %sysfunc(count(&&vi&v,:));
            %let todo = &tags;
            %do t = 1 %to &tags;
                %let token  = %scan(&&vi&v,&t,:);
                %let tag&t  = %scan(&token,1,!);
                %let task&t = %scan(&token,2,!);
            %end;
            *** -- process valid tasks for variables -- ***;
            %do %while(%eval(&todo > 0));
                %do t = 1 %to &tags;
                    %if &&tag&t = align %then %let todo = %eval(&todo - 1);;
                    %if &&tag&t = derive %then %do;
                        &&vn&v = &&task&t;
                        %let todo = %eval(&todo - 1);
                    %end;
                %end;
                %let todo = 0;
            %end;
        %end;
        output;
    %end;
%end;
run;
```

Align Variables that Populate based on Other Variables

```
*** -- now process for alignment instructions -- ***;
data spec.&ds._test(keep=&fldarray);
 set &ds._test;
    %do v = 1 %to &vcnt;
        %let tags = %sysfunc(count(&&vi&v,:));
        %do t = 1 %to &tags;
            %let token  = %scan(&&vi&v,&t,:);
            %let tag&t  = %scan(&token,1,!);
            %let task&t = %scan(&token,2,!);
        %end;
        %do t = 1 %to &tags;
            %if &&tag&t = align %then %do;
                if not (&&task&t) then do;
                    %if &&vt&v = Char %then %do;
                        &&vn&v = "";
                    %end;
                    %else %do;
                        &&vn&v = .;
                    %end;
                end;
            %end;
        %end;
    %end;
run;
```

Sample of Synthetic DM Data Generated

	STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC	RFENDTC	DTHDTC	DTHFL	SITEID	AGE	AGEU	SEX	RACE	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY
1	1234	DM	1234-01001-1001	1001	2016-08-19T14:43:50	2017-05-03T14:43:50		N	01001	22	YEARS	M	WHITE	A	ARM A	B	ARM A	USA
2	1234	DM	1234-01001-1002	1002	2016-05-27T14:43:50	2016-11-09T14:43:50		N	01001	18	YEARS	F	BLACK OR AFRICAN AMERICAN	A	ARM A	B	ARM A	USA
3	1234	DM	1234-01001-1003	1003	2016-05-25T14:43:50	2017-03-01T14:43:50	2017-03-01T14:43:50	Y	01001	18	YEARS	M	WHITE	B	ARM B	A	ARM B	USA
4	1234	DM	1234-01001-1004	1004	2016-06-15T14:43:50	2017-01-27T14:43:50		N	01001	33	YEARS	F	WHITE	B	ARM B	B	ARM B	USA
5	1234	DM	1234-01001-1005	1005	2016-05-25T14:43:50	2016-12-10T14:43:50		N	01001	45	YEARS	M	WHITE	A	ARM A	B	ARM A	USA
6	1234	DM	1234-01001-1006	1006	2016-07-03T14:43:50	2016-11-16T14:43:50		N	01001	34	YEARS	M	WHITE	B	ARM B	B	ARM B	USA
7	1234	DM	1234-01001-1007	1007	2016-08-04T14:43:50	2017-02-06T14:43:50		N	01001	43	YEARS	M	ASIAN	B	ARM B	B	ARM B	USA
8	1234	DM	1234-01001-1008	1008	2016-06-12T14:43:50	2017-01-11T14:43:50		N	01001	61	YEARS	M	WHITE	B	ARM B	A	ARM B	USA
9	1234	DM	1234-01001-1009	1009	2016-06-19T14:43:50	2017-01-19T14:43:50		N	01001	26	YEARS	F	WHITE	A	ARM A	B	ARM A	USA
10	1234	DM	1234-01002-1001	1001	2016-07-26T14:43:50	2017-02-02T14:43:50		N	01002	18	YEARS	F	WHITE	A	ARM A	A	ARM A	USA
11	1234	DM	1234-01002-1002	1002	2016-07-13T14:43:50	2016-12-01T14:43:50		N	01002	35	YEARS	F	WHITE	B	ARM B	A	ARM B	USA
12	1234	DM	1234-01002-1003	1003	2016-08-01T14:43:50	2016-11-23T14:43:50	2016-11-23T14:43:50	Y	01002	74	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	B	ARM B	USA
13	1234	DM	1234-01002-1004	1004	2016-06-05T14:43:50	2016-10-31T14:43:50	2016-10-31T14:43:50	Y	01002	45	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	B	ARM B	USA
14	1234	DM	1234-01002-1005	1005	2016-08-17T14:43:50	2017-04-23T14:43:50		N	01002	45	YEARS	F	ASIAN	B	ARM B	B	ARM B	USA
15	1234	DM	1234-01002-1006	1006	2016-07-20T14:43:50	2016-11-05T14:43:50		N	01002	57	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	B	ARM B	USA
16	1234	DM	1234-01002-1007	1007	2016-08-08T14:43:50	2017-02-19T14:43:50		N	01002	45	YEARS	F	BLACK OR AFRICAN AMERICAN	B	ARM B	B	ARM B	USA
17	1234	DM	1234-01002-1008	1008	2016-08-10T14:43:50	2017-04-07T14:43:50		N	01002	23	YEARS	M	ASIAN	A	ARM A	B	ARM A	USA
18	1234	DM	1234-01002-1009	1009	2016-08-20T14:43:50	2016-12-21T14:43:50		N	01002	51	YEARS	M	ASIAN	A	ARM A	A	ARM A	USA
19	1234	DM	1234-01003-1001	1001	2016-07-11T14:43:50	2016-11-09T14:43:50		N	01003	68	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	A	ARM B	USA
20	1234	DM	1234-01003-1002	1002	2016-07-14T14:43:50	2017-02-09T14:43:50		N	01003	37	YEARS	F	WHITE	B	ARM B	A	ARM B	USA
21	1234	DM	1234-01003-1003	1003	2016-08-13T14:43:50	2017-04-29T14:43:50	2017-04-29T14:43:50	Y	01003	79	YEARS	F	WHITE	B	ARM B	A	ARM B	USA
22	1234	DM	1234-01003-1004	1004	2016-05-30T14:43:50	2016-10-19T14:43:50		N	01003	55	YEARS	F	WHITE	A	ARM A	B	ARM A	USA
23	1234	DM	1234-01003-1005	1005	2016-06-02T14:43:50	2017-01-22T14:43:50		N	01003	77	YEARS	F	WHITE	B	ARM B	B	ARM B	USA
24	1234	DM	1234-01003-1006	1006	2016-07-26T14:43:50	2017-04-03T14:43:50		N	01003	62	YEARS	M	ASIAN	B	ARM B	A	ARM B	USA
25	1234	DM	1234-01003-1007	1007	2016-07-18T14:43:50	2017-01-26T14:43:50		N	01003	79	YEARS	F	WHITE	B	ARM B	A	ARM B	USA
26	1234	DM	1234-01003-1008	1008	2016-07-27T14:43:50	2017-01-15T14:43:50		N	01003	78	YEARS	M	ASIAN	A	ARM A	A	ARM A	USA
27	1234	DM	1234-01003-1009	1009	2016-07-19T14:43:50	2016-11-19T14:43:50		N	01003	75	YEARS	F	WHITE	B	ARM B	A	ARM B	USA
28	1234	DM	1234-01004-1001	1001	2016-08-19T14:43:50	2017-04-15T14:43:50		N	01004	64	YEARS	F	ASIAN	B	ARM B	B	ARM B	USA
29	1234	DM	1234-01004-1002	1002	2016-07-25T14:43:50	2017-01-26T14:43:50		N	01004	72	YEARS	F	WHITE	B	ARM B	B	ARM B	USA
30	1234	DM	1234-01004-1003	1003	2016-06-11T14:43:50	2016-11-02T14:43:50		N	01004	50	YEARS	M	BLACK OR AFRICAN AMERICAN	A	ARM A	B	ARM A	USA
31	1234	DM	1234-01004-1004	1004	2016-06-29T14:43:50	2017-04-21T14:43:50		N	01004	73	YEARS	F	WHITE	A	ARM A	B	ARM A	USA
32	1234	DM	1234-01004-1005	1005	2016-07-11T14:43:50	2017-04-12T14:43:50		N	01004	33	YEARS	F	BLACK OR AFRICAN AMERICAN	B	ARM B	A	ARM B	USA
33	1234	DM	1234-01004-1006	1006	2016-08-09T14:43:50	2017-03-31T14:43:50	2017-03-31T14:43:50	Y	01004	29	YEARS	M	WHITE	A	ARM A	A	ARM A	USA
34	1234	DM	1234-01004-1007	1007	2016-08-12T14:43:50	2017-02-08T14:43:50		N	01004	20	YEARS	M	WHITE	A	ARM A	B	ARM A	USA
35	1234	DM	1234-01004-1008	1008	2016-07-21T14:43:50	2017-02-15T14:43:50		N	01004	53	YEARS	M	BLACK OR AFRICAN AMERICAN	A	ARM A	B	ARM A	USA
36	1234	DM	1234-01004-1009	1009	2016-06-22T14:43:50	2017-01-31T14:43:50		N	01004	42	YEARS	M	WHITE	B	ARM B	A	ARM B	USA
37	1234	DM	1234-01005-1001	1001	2016-07-16T14:43:50	2016-11-28T14:43:50		N	01005	62	YEARS	F	ASIAN	A	ARM A	B	ARM A	USA
38	1234	DM	1234-01005-1002	1002	2016-07-26T14:43:50	2017-03-16T14:43:50		N	01005	37	YEARS	M	WHITE	B	ARM B	B	ARM B	USA
39	1234	DM	1234-01005-1003	1003	2016-05-24T14:43:50	2017-01-03T14:43:50		N	01005	57	YEARS	F	WHITE	B	ARM B	B	ARM B	USA
40	1234	DM	1234-01005-1004	1004	2016-08-08T14:43:50	2017-03-17T14:43:50		N	01005	58	YEARS	M	WHITE	B	ARM B	B	ARM B	USA
41	1234	DM	1234-01005-1005	1005	2016-07-09T14:43:50	2017-03-29T14:43:50		N	01005	26	YEARS	F	ASIAN	A	ARM A	A	ARM A	USA
42	1234	DM	1234-01005-1006	1006	2016-07-14T14:43:50	2017-04-27T14:43:50		N	01005	62	YEARS	F	ASIAN	A	ARM A	B	ARM A	USA
43	1234	DM	1234-01005-1007	1007	2016-05-24T14:43:50	2017-01-24T14:43:50		N	01005	79	YEARS	M	ASIAN	B	ARM B	B	ARM B	USA
44	1234	DM	1234-01005-1008	1008	2016-06-23T14:43:50	2016-10-29T14:43:50		N	01005	24	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	A	ARM B	USA
45	1234	DM	1234-01005-1009	1009	2016-07-18T14:43:50	2017-03-15T14:43:50	2017-03-15T14:43:50	Y	01005	76	YEARS	M	BLACK OR AFRICAN AMERICAN	B	ARM B	A	ARM B	USA