PharmaSUG2010 - Paper CD11

Quality in a World of "Standards" with Differences!

Sandra VanPelt Nguyen, i3 Statprobe

ABSTRACT

Even with the use of data standards such as CDISC, obtaining quality data and analyses can still be quite challenging, particularly when there is great variance in the interpretation and usage of standards.

How many varieties of SDTM have you seen? Have you heard the descriptions "SDTM-like", "SDTM-friendly", "SDTM-plus", or "SDTM-minus"? How many varieties of data specifications have you seen? Does anyone use controlled terminology? How many different file formats have you received from vendors? Do you ever struggle with how to map data to SDTM? So do others – and they may end up with different results!

This presentation will take a look at how standards are implemented, where there are opportunities for improving quality when using standards, and the challenges that still exist with the use of standards, with focus on the CDISC SDTM standard. Additionally, it will cover how process and tools can be used to enhance the use of standards and assist towards improving overall quality.

INTRODUCTION

Ideally, the use of standards such as CDISC SDTM will minimize quality issues and lead to more effective and efficient design, creation, implementation, and usage of data review and quality control checks and programs.

Realistically, achieving clean, quality data and analyses still presents many challenges! This is further complicated by the variety of means and degrees of standards implementation as well as differences in interpretation within and between companies.

With the many challenges which exist, it is all the more important to take steps to ensure the quality of clinical trial data. These steps must be robust enough to handle the challenges and do so in an efficient and effective manner.

Data is ultimately validated when accepted by a regulatory agency. For SDTM datasets submitted to the FDA, this means that the agency must be able to load the data into WebSDM™ [1]. WebSDM™, besides being a review tool, consists of a set of rules [2] which validate the data against the SDTM Implementation Guide [3] and FDA's Study Data Specifications [4].

Study data typically needs to be cleaned for accuracy and completeness prior to submission, in order to accurately summarize, present, and interpret results. Several data cleaning and quality control steps will be performed throughout the processes of collecting, organizing, and analyzing the data. The use of data standards such as CDISC SDTM may allow for more efficient review and cleaning procedures, resulting in earlier detection and resolution of issues, which in turn may lead to shorter project timelines and reduced costs.

CHALLENGES

Even with the use of standards such as CDISC SDTM, many challenges to obtaining clean, quality data still exist and will continue to exist as standards develop and are implemented. Several of these are a part of the growing pains and learning curves associated with changes. Others are inherent to the nature of the business or to human nature itself. Additionally, some challenges are actually are brought on by partial adoption of standards or by differences in interpretation of standards.

Some of these challenges include:

- Site or vendor data reconciliation/logic issues

  The use of standards does not remove the potential for data entry errors, plus the entry of data to multiple sites or forms will always create an opportunity for discrepancies between data sources (e.g. sample identification information collected on the CRF versus what is recorded on the samples sent to the lab).

- Poor and/or inconsistent CRF design

  A complicated or confusing form may lead to incorrect completion of the form or incorrect interpretation of the fields. If CDISC is not considered when designing the CRF, it may later be difficult to map the data into the standards. Furthermore, differences in the forms between studies will make it more difficult to integrate the data and will cause inefficiencies in programming.

- Legacy data not set up for use with current standards

  Older CRFs may not have collected information which is required in SDTM, may have missing or incomplete documentation, and may contain data issues, which could make it difficult to determine the mapping to the SDTM domains and variables and could cause WebSDM™ findings. Additionally, depending on the source data format, the mapping to SDTM domains and variables may not be so straightforward.

- Variety of source data formats, structures, and contents

  Data continues to be transferred in a variety of formats and structures, depending on the operational databases used, the data provider's level of standards implementation, and their ability to provide CDISC-compliant files (the vendor may have a standard, but it may not be CDISC; or it may be based on CDISC, but not fully adherent to the model). Even when standard fields are used, standard terminology may not be used, making it difficult to match data records up, tabulate data, and analyze data. Limited or no metadata may cause additional time trying to identify and interpret the contents.

- Study-specific issues inherent to study design, nature of study (indication, subject population, etc.)

  Due to the nature of a study (indication, subject population, etc.), it may not be possible to acquire all data required for SDTM, and the inclusion or exclusion of certain fields or values may cause WebSDM™ findings. Additionally, there may be issues associated with patient-reported data, IVRS, and other methods of data collection utilized for a study.

- Unanticipated data

  One must always expect the unexpected. Even with the use of standards and the best planning, the data is not always as you anticipated and may require special handling. It is just not possible to anticipate every possible data scenario which may arise during the course of a clinical trial. Furthermore, forms may not allow for the appropriate collection of specific data scenarios, so sites may get creative.

- Differing degrees of implementation of standards

  Many companies have not yet adopted CDISC standards and others have only partially adopted them. The varieties of levels of implementation and compliance make it difficult to quickly and easily obtain and use clinical trial data.

- Customer-specific requirements

  If you work with multiple customers, you need to be prepared to create or use data in different formats. Different companies have implemented standards in their own ways – different processes, tools, and documents – and may have different expectations or requirements based on their own interpretation of the standards.

- Differing interpretations of CDISC SDTM guidance

  Since there is still some level of interpretation within SDTM, there is the possibility for different usage of particular domains and variables. SDTM also allows for multiple ways to handle particular data situations (e.g. handling of an "Other, Specify:" free text field). Additionally, the modeling of custom domains and usage of SUPPQUAL and RELREC may vary greatly between companies.

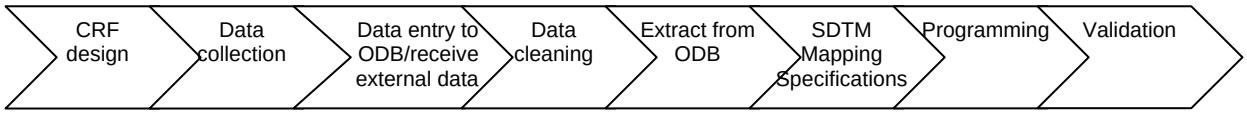
- Inexperience with standards (internally and externally)

  Lastly, inexperience can make it more difficult and time-consuming to incorporate standards. It can be difficult to persuade people of the benefits of standards, particularly when there has been a great investment in development and training associated with legacy systems. Plus, there is a natural reluctance to change. Inexperience furthermore creates a reliance on those who do have the necessary expertise, which may mean looking outside of your team or your company and may impose additional costs. Working with inexperienced clients or vendors may lead to a lot of questions and confusion and slow down the process if there are difficulties getting answers or information needed to work with the data.

Challenges related to differing customer requirements, interpretation of standards, and level of standards implementation can make it difficult, but not impossible, to incorporate standards. On the other hand, the use of standards should help alleviate some of the challenges associated with study design, data entry and reconciliation errors, and receipt of vendor or other external data, but will not eliminate them entirely. Standards can, however, be used to help minimize the impact of these types of challenges.

IMPLEMENTATION OF STANDARDS

Typical process from CRF through SDTM:



*ODB = Operational Database

When standards are not implemented upstream, there is potential for several problems or complications:

- There may be extra mapping work at the end of a project, and at a point where it is difficult to go back and address issues in the data.
- The lack of controlled terminology may mean a disconnection between the SDTM datasets and the terminology used in the protocol and the case report forms.
- There may be complications with joining and reconciling data sources when coming in a variety of formats and using discrepant terminology (differing visit identifiers, for instance).
- You may not collect data which is required for SDTM and then must decide if you can or should derive it later.
- If SDTM is not used to generate the analysis datasets, then there may be a disconnection between the datasets included in the submission and extra documentation may be needed in order to provide the connection for the reviewers.

OPPORTUNITIES TO IMPROVE QUALITY AND STREAMLINE PROCESSES

Within the processes leading towards generation of SDTM datasets, there are several opportunities to implement standards and take advantage of the resulting benefits:

- Potentially fewer data entry errors

  There are clearer expectations for the data once all parties gain familiarity with the standards. Additionally, there is opportunity to build checks in up front (in Electronic Data Capture systems) so that erroneous values are immediately flagged and disallowed for entry.

- Greater ability to detect issues earlier

  Beyond building immediate checks into Electronic Data Capture (EDC) systems, standard edit checks can be created and run early and often during the data capture and cleaning processes.

- Reduced set-up time

  By using standard case report forms, a standard database set-up can be used for multiple projects, along with standard edit checks.

- Reuse of code

  Standard data extraction and/or mapping programs can be created to generate the SDTM datasets. Additionally, standard programs for loading, reconciling, and joining external data can be used with vendors or partners who have also adopted standards.

- Reduced study-specific modifications

  With the use of standard case report forms and external data standards, there is little study-specific modification needed in set-up and programming (mostly just for efficacy).

- Streamlined validation procedures

  With standard data structures, standard programs may be validated in a general fashion and then used for multiple projects. You may only need basic validation steps at the project level to ensure that programming is functioning as expected. Additionally, automated systems of data checks can be built to help identify data and compliance issues.

- Easier to transition between projects

  With the use of standards, everyone should be familiar with the data collection, quality checks, data structures, and programming and should easily be able to move from one project to another.

- Opportunity to develop a comprehensive process

  Based on the standard inputs and outputs which will be generated and used, it is possible to build an end-to-end process which may be used from one project to another. This again makes it easier to transition between projects and may further make it easier to project timelines and gather metrics.

- Opportunity to develop systems and tools

  Yes, many companies already have their internal systems and tools built around their legacy data structures, however CDISC will be required for data submissions in the near future, and with the benefits of upstream standards implementation, it is worth the initial investment to implement CDISC-based systems and tools at the earliest possible stage. Once systems and tools are in place which can be used with standard data structures, costs may be greatly reduced (set-up, data collection and entry, programming, and analysis).

With the use of standards comes the opportunity to develop and incorporate quality checks at multiple levels and timepoints within a clinical trial. By leveraging these opportunities, costs and timelines may be reduced and will ultimately result in greater preparedness for submission of data.

HOW TO ADDRESS THE CHALLENGES

1. **Despite the challenges, use standards and incorporate them as far upstream as possible**.

   As discussed, using standards as early in the process as possible provides for additional benefits and reduced costs and aids to reduce the potential for and increase the detection of errors and issues in the data. Invest time up front to determine how to address non-standard data, special challenges or requirements, etc.

2. **Develop or acquire the expertise; however, not everyone needs to be an expert**!

   By creating a core team of expertise, the process and tools can be built such that anyone may use them. A core team can provide the training and consultation needed to ensure appropriate execution, usage, and application of the data standards. Once the process and tools are in place, it is mostly a matter of following and using them. A basic understanding of the fundamental structures and rules of the standards is sufficient for most users to be able to produce the desired outputs, under the guidance of those more experienced and knowledgeable in the standards.

3. **Develop standard processes**.

   Create an end-to-end process (or as far upstream as possible) to realize the maximum benefits. The process must be robust enough to handle the potential challenges – allowing for different source data formats which may be received, different levels of standards implementation, different customer requirements, etc. Theoretically, the process would be the same from project to project, however certain components within each step may require special attention due to variance from the defined standards.

4. **Develop standard systems and tools.**

   Try to automate the steps within the process to the extent possible. Apply standard programs and checks around the standard data formats. This will reduce costs and reduce the potential for errors due to study-specific modifications.

5. **Incorporate standard quality checks at every opportunity.**

   The use of standards allows an opportunity to create standard quality checks and apply quality checks at early stages in the process. Keeping in mind the WebSDM™ validation rules, the data may be checked early and often for potential issues, and allow for early resolution or acknowledgement* of potential findings.

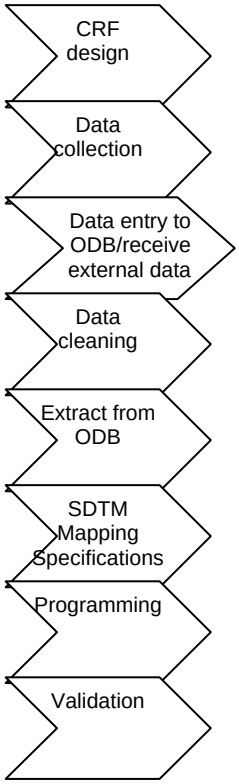
   *Note that some WebSDM™ findings may not necessitate changes to the data, but serve as warnings of potential reviewability issues.

6. **Allow for flexibility.**

   Given all of the potential challenges, one must allow for some degree of flexibility. The process is critical here – the overall process can generally be followed under most circumstances, with potential variance within particular steps as needed to accommodate non-standard data.

BACK TO IMPLEMENTATION…

Taking a look back at the process for generating SDTM datasets, we can then envision how standard processes, systems, and tools can be incorporated to enhance the overall quality:



Standard CRFs

Better compliance with use of standard forms
Use of CDISC controlled terminology to limit variability in data values

Standard database set-up
Standard data entry conventions
Data standards established with vendors

Standard edit checks and data reconciliation programs

Standard extract programs

Standard specification format and mapping specifications based on standard source data and internal mapping conventions
Tools for creation of specifications and specification checks

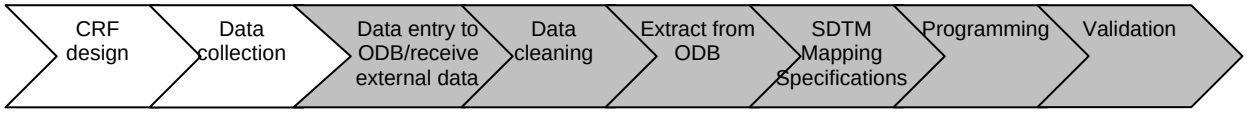
Standard SDTM programs (at least for safety and common domains)
Tools may generate or partially generate programs, map values to CDISC controlled terminology, generate define docs, etc.

Standard data, metadata, and compliance checks

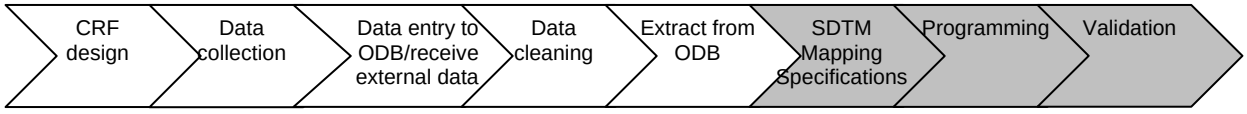
To handle variances in the data sources and other particular data challenges, it is best to make use of the standard process and tools to the fullest extent possible, moving data into the standards at the earliest possible point.

Examples:

a) If using non-standard CRFs, try to design the database as closely to the standard structure as possible so that standard edit checks and programming may be used.



b) If receiving legacy data (including external data) in non-standard data format, try to use standard SDTM mapping specifications so that tools for programming and validation of data and metadata can be used to the fullest extent possible.



RESULTS

TIME AND COSTS

As described, there is much potential to reduce overall timelines and costs by applying standards. A look at several case studies will demonstrate the benefits of using data standards:

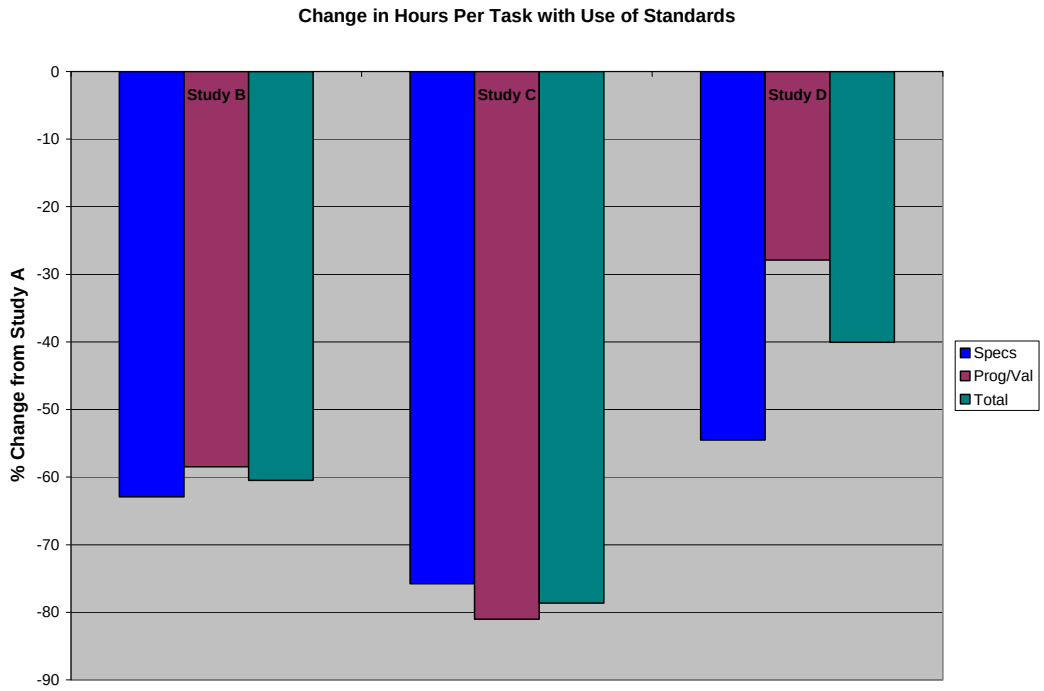
Case Study A: Did not fully utilize data standards, did not follow a standard process, and had many compliance issues identified at the end of the project which then required rework

Case Study B: First study fully utilizing internal data standards (based on CDISC) and standard process and tools

Case Study C: Later study fully utilizing internal data standards and standard process and tools, with enhancements based on previous experience.

Case Study D: Later legacy data conversion (non-standard source data and limited metadata), which used the standard process and tools to the extent possible.

And the winner is……

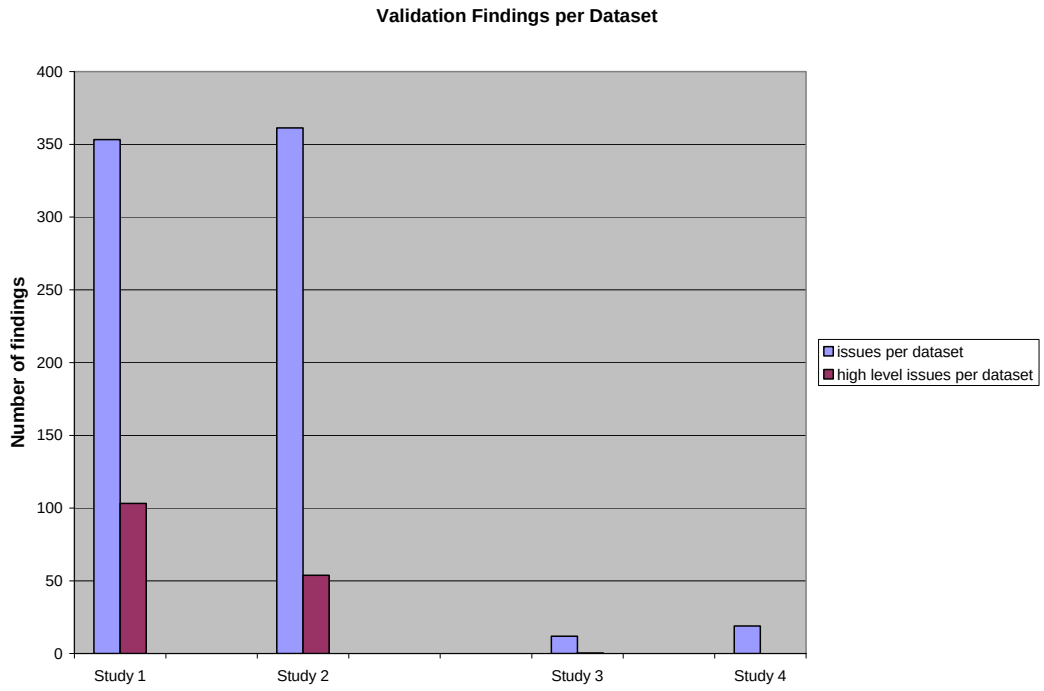


Case Study C!

All studies which utilized data standards in some manner experienced a reduction in hours for creation of SDTM mapping specifications and programming and validation of SDTM datasets. The greatest benefits were seen with studies which incorporated data standards from beginning to end, followed a standard process, and used standard data tools. Further reductions were seen after gaining experience with the data standards, process, and tools and using previous experience to finesse and improve on the process and tools.

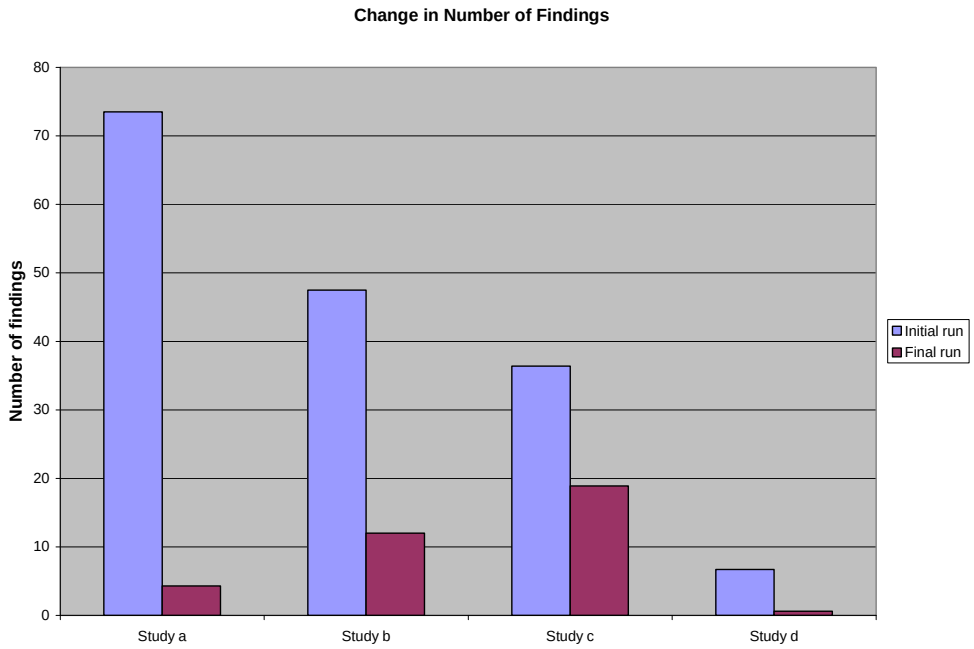
QUALITY AND COMPLIANCE

Additionally, as described earlier, data is ultimately 'validated' through WebSDM™. Several studies were run through the WebSDM™ data checks following database lock (note that studies referenced below may not correspond to case studies in previous example):



Case studies 1 and 2 were not run against the WebSDM validation rules until the end of the studies. There were many findings, requiring the programming team to go back and modify the SDTM mapping specifications and update and revalidate dataset programming, and all at the end of the project when a submission was imminent and there was little time to spare. Case studies 3 and 4 not only had the traditional dataset validation procedures performed, but were additionally processed through an automated system of data checks based off of the WebSDM validation rules at several points during the studies prior to and following database locks. At the end of the studies, there was minimal effort to review any new validation findings and confirm whether any changes were needed. The majority of findings had already been identified in previous data reviews and were explained and documented prior to database locks, allowing for quick turnaround of datasets and metadata at the end of the project.

Despite the benefits of using standard data checks and incorporating them early into the process, issues will still arise in the data due to the continuing data challenges which will be encountered. Each of the case studies examined below (note that studies referenced below may not correspond to case studies in previous examples) had numerous validation findings which required data cleaning and/or specification or programming modifications. Data issues could not be eliminated with the use of standards, but could be reduced over time as additional standards were implemented, both in data structures and processes, and additional quality checks were built into the processes and became more automated.



CONCLUSION

When using data standards such as CDISC SDTM, it is easier to implement and automate standard data checks, however this does not necessarily mean there will be fewer obstacles to achieving clean, quality data, as many data challenges will continue to exist even as standards are adopted more comprehensively and consistently within the industry. Additionally, although standard data checks can be implemented, it is never possible to foresee all possible data scenarios and, ultimately, data is never perfect. With the use of data standards, we can make additional efforts to limit the potential for and maximize the detection of errors and issues in the data. By creating and using standard processes and tools which can handle various data situations, and implementing these standards as far upstream as possible, it is possible to save time and costs and improve the overall quality of the data.

REFERENCES

[1] Phase Forward. *WebSDM™ data sheet*. Retrieved from http://www.phaseforward.com/products/clinical/ads/.

[2] Phase Forward. *WebSDM™ v3.0 Edit Checks*. Retrieved from http://www.phaseforward.com/products/cdisc/.

[3] CDISC. *SDTM Implementation Guide v.3.1.2*. Retrieved from http://www.cdisc.org/sdtm.

[4] FDA. *Study Data Specifications*. Retrieved from http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM199759.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Sandra VanPelt Nguyen
Enterprise: i3 Statprobe
E-mail: sandra.vanpeltnguyen@i3statprobe.com