PhUSE US Connect 2018

Paper TT13

A statistics-based tool to inform risk-based monitoring approaches

Laura Williams, CROS NT, Chapel Hill, NC, USA
Giulia Zardi, CROS NT, Verona, IT

ABSTRACT

In light of the latest guidelines for good clinical practice (ICH E6(R2) Integrated Addendum), sponsors of clinical trials should implement a "systematic, prioritized, risk-based approach" to monitoring activities, potentially including both on-site and centralized monitoring. CROS NT has developed a tool, which uses a statistically-controlled scoring method in conjunction with principal component analysis (PCA) to identify potentially problematic sites, rather than subjective thresholds. Data-driven risk factors, mainly computed as p-values, are used to assess the quality of data. CROS NT is currently integrating it with a customizable dashboard that visualizes the workflow, allowing sponsors or CROs to receive alerts and track corrective actions. Users can drill down through the visualizations from a global view to comprehensive profiles by site or by subject. The platform can provide real-time visibility of multiple sources of clinical and operational data.

INTRODUCTION

In light of the latest guidelines for Good Clinical Practice (ICH E6(R2) Integrated Addendum)[1], sponsors of clinical trials should implement a "systematic, prioritized, risk-based approach" to monitoring activities. The traditional approach to clinical trial monitoring is to deploy monitors to sites on a regular basis, however the latest guidelines suggest centralized monitoring or a combination of centralized and on-site monitoring may be more effective and efficient. The guidelines highlight that you can use a more comprehensive statistical approach to identify issues that can impact subject safety and reliability of results. Even non-critical data can indicate problems at a site, so a statistical approach that looks at the totality of the data is an important key risk indicator. This type of risk indicator can be used as part of a centralized monitoring process, which may reduce the frequency and/or scope of on-site monitoring.

The ultimate responsibility for the quality and integrity of the trial data always resides with the sponsor, even if a study is fully outsourced to a CRO. Therefore, sponsors need tools to quickly and easily assess the quality of data from the clinical trial sites. Statistics-based tools can summarize large clinical databases into meaningful visualizations, which is particularly important for multi-center and international studies. With the application of electronic case report forms (CRFs), it is possible to develop tools to support risk-based monitoring (RBM). Sites can be monitored as data is collected, taking into account risk factors, tracking study progression, and proactively addressing potential critical situations.

CROS NT has developed a tool which uses a statistically-controlled scoring method in conjunction with principal component analysis (PCA) to identify potentially problematic sites, rather than subjective thresholds. The computational core of the analysis consists of identifying data driven risk factors, many as p-values, for data quality and analyzing these risk factors with different PCA approaches. Therefore, sites with issues such as repeated values or digit preference become apparent. These are difficult to detect by monitors reviewing data at a single site or a single patient visit. CROS NT is currently integrating it with a customizable dashboard that visualizes the workflow, allowing sponsors or CROs to receive alerts and track corrective actions. Users can drill down through the visualizations from a global view to comprehensive profiles by site or by subject. The platform can provide real-time visibility of multiple sources of clinical and operational data.

This paper describes technology solutions to enable a risk-based approach in accordance with the guidelines set out in the new ICH GCP E6 (R2) Addendum. It also demonstrates how you can harness the power of data visualization and statistics to enhance the collaboration between clinical operations, data management and biostatistics.

MONITORING APPROACHES

The clinical phase is the most complex part of the drug development process. It requires efficient planning, conducting, and monitoring of clinical trials to produce reliable study data. Quality study data is crucial for submission to regulatory agencies. The monitoring of clinical trials ensures that protocols are followed and data is as complete and accurate as possible. As complexities in clinical trials continue to increase, the clinical monitoring cost has also risen significantly, in order to not only achieve higher data quality but to protect subject safety. Monitoring is one of the largest expenses in a clinical trial, accounting for 9-14% of the overall budget.[2] Therefore, any efforts to fortify and streamline the process will be beneficial to both monitors and clinical trial sponsors.

In addition to the benefits to the sponsor of clinical trials, the most recent update to the ICH Guideline for Good Clinical Practice (Integrated Addendum E6(R2) – November 2016)[1] encourages "implementation of improved and more efficient approaches to clinical trial design, conduct, oversight, recording and reporting while continuing to ensure human subject protection and reliability of trial results." One such approach, is risk-based monitoring. FDA guidance for industry[3] and an EMA reflection paper[4], both published prior to the Integrated Addendum, support this approach. Risk-based monitoring can include on-site monitoring (performed at the site where a clinical trial is being conducted) and centralized monitoring (remote review of the data accumulated at each site, usually by data managers and biostatisticians).

TRADITIONAL MONITORING

Traditional monitoring practice has involved Clinical Research Associates (CRAs) visiting each site in a clinical trial on a regularly, scheduled basis (e.g. every 4-6 weeks). During these site visits, a CRA will typically perform 100% Source Data Verification (SDV). This task means the CRA is reviewing every page of the Case Report Form (CRF) (electronic and/or paper) and comparing it to primary health records. Alternatively, certain fields of the CRF could be selected to be reviewed at random, instead of 100% at each visit. The strategy is usually outlined in a Data Monitoring Plan. The purpose of this task is to verify that there are no data entry errors or fraud happening at a site. However, this is a time consuming task and usually leaves little time for the CRA to be potentially helpful in other areas, such as investigational product accountability, site recruitment, checking training logs and protocol deviations.

RISK-BASED MONITORING

Risk-Based Monitoring (RBM) aims to assist CRAs and clinical trial sponsors, promoting risk mitigation and early issue detection, by looking at the data as collected. A traditional approach to RBM involves setting metrics to be reviewed throughout the study. The study sponsor and Data Manager (DM) set thresholds to flag sites or subjects (e.g. > 3 Serious Adverse Events (SAE) per site; drug compliance < 80%) and corrective actions for exceeding the thresholds. The DM reviews the data (or it is done programmatically) as it is collected and alerts the CRAs/sponsor when a subject or site meets or exceeds a threshold. The data review and preventative/corrective actions are tracked and documented. Then, critical results are discussed with the study team, so everyone is up to date as the study progresses. This approach allows CRAs to focus their activities specifically on high risk sites identified by outcomes of RBM instead of equally distributing efforts over all the sites. Additionally, the data cleaning process is streamlined as the DM is reviewing data on an ongoing basis. Also, the sponsor has a broader view of how the study is progressing. However, since thresholds are subjective (as based on clinical experience) and fixed, they may not be effective in all areas; for example, excessive SAEs and low drug compliance can easily be flagged, but it may be difficult to detect fraud or instrument mis-calibration.

CENTRALIZED STATISTICAL MONITORING

Centralized Statistical Monitoring (CSM) is a type of central monitoring, where instead of metrics being set by the study sponsor and DM, they are data driven. Data is analyzed as collected with appropriate statistical methodologies, therefore additional team members (Biostatistician, Statistical Programmers) are involved in the monitoring process. With CSM, each site is assigned a type of "risk score" based on the results of the statistical models. Therefore, sites deemed to be high-risk can be the focus of monitoring efforts, including corrective actions. This type of data-driven approach creates exceptional data quality checks and can flag "risky" sites much sooner than using traditional monitoring or RBM Metrics. Statistical models can also easily detect fraud and instrument mis-calibration or malfunction. However, each protocol is different and requires statistical input to determine critical data. The appropriateness of the model must be carefully reviewed by the Biostatistician on an ongoing basis. Also, a significant amount of data needs to be collected before certain statistical models can be considered appropriate. Thus, this type of monitoring may not be feasible for small studies.

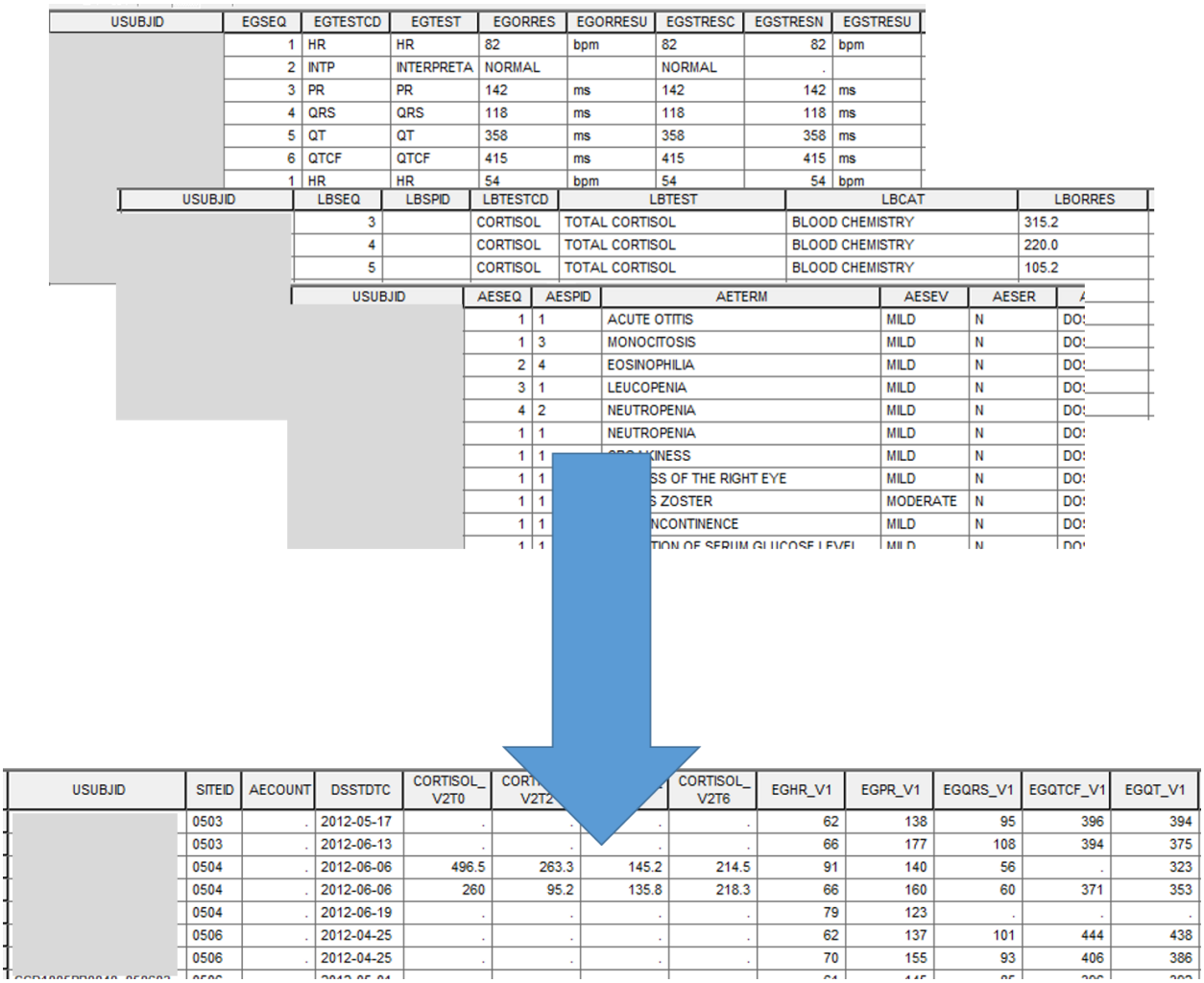
Both RBM metrics and CSM can reduce the SVD done by the CRAs, by highlighting data that seems to be problematic. Together, they can identify issues such as data entry errors, treatment non-compliance, enrollment issues, instrument mis-calibration or malfunction, and fraud. We have developed a tool, called Central Statistical Intelligence (CSI), that is based on Principal Component Analysis (PCA) and the results of which can be used to inform CSM.

AN APPROACH TO RBM: CENTRAL STATISTICAL INTELLIGENCE

Our CSI tool uses a statistical methodology developed around PCA, the results of which are used to compute a risk score. The input to the PCA is a matrix of indicators, based on the data determined to be critical by the Biostatistician. Generally, the distribution of data at one site is compared to all the other sites. This type of model requires a suitable number of sites with at least 10 subjects per site to be meaningful. The risk scores are ranked and the sites with high risk scores are red-flagged.

INDICATORS

The first step to computing various indicators is to aggregate subject data. Our tool has been optimized using CDISC standards, particularly SDTM datasets. The use of SDTM allows for the reusability of code, due to a standardized data structure. Figure 1 below show the process of preparing the data. Subject data is collapsed by visit and time point into a single row per subject, representing all the critical data in the database, as identified by the Biostatistician and agreed with the trial sponsor.



USUBJID	EGSEQ	EGTESTCD	EGTEST	EGORRES	EGORRESU	EGSTRESC	EGSTRESN	EGSTRESU
	1	HR	HR	82	bpm	82	82	bpm
	2	INTP	INTERPRETA	NORMAL		NORMAL	.	
	3	PR	PR	142	ms	142	142	ms
	4	QRS	QRS	118	ms	118	118	ms
	5	QT	QT	358	ms	358	358	ms
	6	QTCF	QTCF	415	ms	415	415	ms
	1	HR	HR	54	bpm	54	54	bpm

USUBJID	LBSEQ	LBSPID	LBTESTCD	LBTEST	LBCAT	LBORRES
	3		CORTISOL	TOTAL CORTISOL	BLOOD CHEMISTRY	315.2
	4		CORTISOL	TOTAL CORTISOL	BLOOD CHEMISTRY	220.0
	5		CORTISOL	TOTAL CORTISOL	BLOOD CHEMISTRY	105.2

USUBJID	AESEQ	AESPID	AETERM	AESEV	AESER	A
	1	1	ACUTE OTITIS	MILD	N	DO
	1	3	MONOCITOSIS	MILD	N	DO
	2	4	EOSINOPHILIA	MILD	N	DO
	3	1	LEUCOPENIA	MILD	N	DO
	4	2	NEUTROPENIA	MILD	N	DO
	1	1	NEUTROPENIA	MILD	N	DO
	1	1	KINESS	MILD	N	DO
	1	1	SS OF THE RIGHT EYE	MILD	N	DO
	1	1	S ZOSTER	MODERATE	N	DO
	1	1	NCONTINENCE	MILD	N	DO
	1	1	TION OF SERUM GLUCOSE LEVEL	MILD	N	DO

USUBJID	SITEID	AECOUNT	DSSTDTC	CORTISOL_V2T0	CORTISOL_V2T2	CORTISOL_V2T4	CORTISOL_V2T6	EGHR_V1	EGPR_V1	EGQRS_V1	EGQTCF_V1	EGQT_V1
	0503	.	2012-05-17	.	.	.	.	62	138	95	396	394
	0503	.	2012-06-13	.	.	.	.	66	177	108	394	375
	0504	.	2012-06-06	496.5	263.3	145.2	214.5	91	140	56	.	323
	0504	.	2012-06-06	260	95.2	135.8	218.3	66	160	60	371	353
	0504	.	2012-06-19	.	.	.	.	79	123	.	.	.
	0506	.	2012-04-25	.	.	.	.	62	137	101	444	438
	0506	.	2012-04-25	.	.	.	.	70	155	93	406	386

Figure 1: Combining SDTM to a merged data set ready for the CSI tool.

Once the data is in the format above (bottom of Figure 1), indicators can be computed. Based on literature[6-16], we have selected the following indicators. The general rule for the computation of indicator variables is that they are unit-less and have a range of 0 to 1, where 0 indicates a very problematic site and 1 indicates no problems at a site. Table 1 below describes the type of error and method to calculate the indicator. Table 1 also includes some error detection approaches that will only be used for visual review and not in the CSI overall risk computations.

Table 1: Indicators

Errors	Approach/Description	Type	Overall Risk
Dates	Detecting visits occurring on weekends and national holidays	proportion	
Univariate Outliers	Identified by using a certain rule (e.g. 3σ rule)	proportion	
Visit Scheduling	Difference between planned study day and actual study day	p-value	
Rounding to Integers	Frequency of rounded values compared among sites	proportion	X
Digit Preference	Trailing digit from all continuous variables – one site compared to all others; P-value for CMH Row Mean Scores Differ; Max percent difference	p-value	

Errors	Approach/Description	Type	Overall Risk
SAE Rate	Frequency per site/number of subjects/time period of reporting	rate	
Systematic Errors	Test on means of one site compared to all others	p-value	X
Repeated Measures	Minimum within-subject variability for each test with repeated measures (standard deviation) at each site.	normalized value	X
Multivariate Inliers/Outliers	Mahalanobis distance compared to Chi-squared statistic to classify observations	proportion	X
Missing Values	Proportion of missing values among sites through a Chi-squared test	p-value	X
Neighborhood	Density-based clustering (Minimum Distance to Nearest Neighbor)	normalized value	X

One important aspect of computing these indicators is to capture the variables of interest for each test or proportion (e.g. values that have repeated measures, values that are collected as non-integers, etc.) automatically, in a non-study specific way. The Base SAS® software procedures PROC CONTENTS and PROC SQL are useful for this task. For example, if the data set shown at the bottom of Figure 1 is transposed so that there is one row per subject, per visit, per time point, the code below captures the names of the variables with repeated measures:

```
/** contents of our merged datasets, in one row per subject, per visit, per time point
format **/
	proc contents data=start out=start_contents noprint;
		run;

/** all numeric variables, not including visit, timepoint **/
	data cont1(where=(type=1 and name not in ('VISITNUM' 'TPTNUM')) keep=name type);
		set start_contents;
	run;

	proc sql noprint;
		select name into :vars separated by ' '
		from cont1;

		select name into :ctvars separated by '_x '
		from cont1;

		select name into :comvars separated by '_x, '
		from cont1;

		select count(*) into :nvars
		from cont1;
	quit;

	%let ctvars=&ctvars._x;
	%let comvars=&comvars._x;

/** counting number of repeats (visits), by subject **/
	data rm;
		set support.start2;
		by usubjid;

		array var[&nvars] &vars;
		array count[&nvars] &ctvars;

		retain &ctvars;
		do i=1 to dim(var);
			if first.usubjid then count[i]=1;
			else if not missing(var[i]) then count[i]=count[i]+1;
		end;

		if last.usubjid then output;
	run;
```

```
/** counting number of repeats (visits), by variable **/
       ods listing close;
       ods output OneWayFreqs=freqs;
       proc freq data=rm;
              tables &ctvars;
       run;
       ods listing;

       data freqs1;
              set freqs;

              var = tranwrd(scan(table, 2), '_x', '');
              reps = sum(&comvars);

              if reps>1;
       run;

/** list of variables that have repeated measures **/
       proc sql noprint;
              select unique var into :repvars separated by ' '
              from freqs1;
       quit;
```

After this code is run, the macro variable *&repvars* will contain a list of variables that have repeated measures, regardless if the repeats are by visit or by time point. This is extremely useful when the data is incomplete, as it does not rely on a list of the variables that are expected to be repeated (e.g. input from the user), but programmatically creates a list of variables that are actually repeated in the current data. This list can then be used to drop or keep variables from the original data set. As an example, here is how it could be used in the computation of the Normalized Minimum Within-Subject Standard Deviation:

```
ods output BasicMeasures=univ1(where=(varmeasure='Std Deviation' and not
missing(varvalue)));

proc univariate data=start2;
       by siteid usubjid;
       var &repvars;
run;
```

The output data set has the standard deviation, by site and subject, for each of the variables included in *&repvars*.

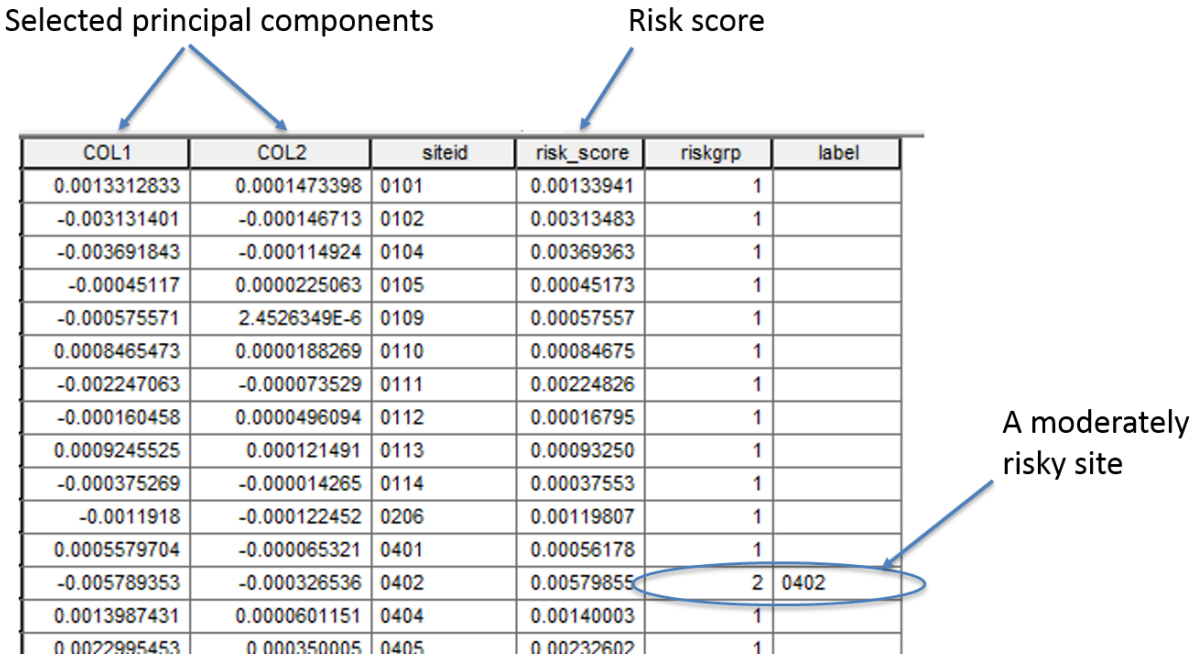
After all the indicators are derived, they are compiled into a matrix, with one row per site. Figure 2 below shows a partial example of the matrix of indicators. Note that not all indicators are shown in this figure. Indicators such as the test on means or repeated measures will have multiple indicators in this matrix; one indicator for each variable from the data shown in Figure 1.

Site ID (siteid)	Proportion of Non-Rounded Values (PNINT)	P-value of Wilcoxon test: QTCF (P2_WIL_QTCF)	Normalized Minimum Within-Subject Standard Deviation: (RM_CORTISOL)	Proportion of Non-Inliers (PNONIN)	Proportion of Non-Outliers (PNONOUT)	Chi-squared test p-value, missingness (RES)	Normalized Minimum Distance to Nearest Neighbor (NORMINDIS)
0504	0.7407407407	0.132932	0.719696278	1	1	0.0047216336	.
0506	0.85	0.878230	.	1	1	0.0122794105	0.0817430923
0507	0.53125	0.705781	.	0.9	1	0.0047420953	0.1222418915
0508	0.48	0.232541	.	0.9	1	0.9549921284	0.1478078928
0509	1	.	0.8717383626	.	.	9.6310114E-7	.
0510	0.7045454545	0.123497	0.790188968	1	1	1.583007E-17	0.1050098791
0601	0.5	0.324101	.	1	1	0.0343096884	0.9999999798
0605	0.5	0.213502	.	1	1	0.000045863	0.2336300407
0607	0.6363636364	0.228676	.	1	1	0.726669051	0.3063486568
0701	0.8791208791	0.566604	0.8982847439	0.966666666	1	3.8656677E-7	0.0782986536
0703	0.6	.	.	.	.	0.0008012062	.

Figure 2: Example matrix of indicators used in the CSI tool.

PRINCIPAL COMPONENT ANALYSIS

Once the matrix of indicators is prepared, it can be used for PCA. PCA was chosen as a method to rank each site because it takes the matrix of indicators and, with a graphical solution, it maps each multidimensional point (i.e. site) into a smaller set of dimensions so that as much as possible of the original variability is explained. The PCA has been implemented in the SAS/IML® language. First, cross validation is used to determine the appropriate number of principal components (PCs). Then, the PCA is run with the number of PCs determined by cross validation. The results must be reviewed by the Biostatistician to determine which PCs are appropriate to use for the risk score calculation. Some PCs may not be considered because they are dominated by the missing-value imputation effect. Once PCs are selected, the risk score is computed as the Euclidean distance from the center of the PC space. Risk thresholds classify those sites as low, medium or high risk. The method described allows for moving thresholds, so that with any level or completeness of data, some sites may be identified. Figure 3 shows an example of the results of the PCA analysis, including the loadings of the two PCs and the resulting risk score.



COL1	COL2	siteid	risk_score	riskgrp	label
0.0013312833	0.0001473398	0101	0.00133941	1	
-0.003131401	-0.000146713	0102	0.00313483	1	
-0.003691843	-0.000114924	0104	0.00369363	1	
-0.00045117	0.0000225063	0105	0.00045173	1	
-0.000575571	2.4526349E-6	0109	0.00057557	1	
0.0008465473	0.0000188269	0110	0.00084675	1	
-0.002247063	-0.000073529	0111	0.00224826	1	
-0.000160458	0.0000496094	0112	0.00016795	1	
0.0009245525	0.000121491	0113	0.00093250	1	
-0.000375269	-0.000014265	0114	0.00037553	1	
-0.0011918	-0.000122452	0206	0.00119807	1	
0.0005579704	-0.000065321	0401	0.00056178	1	
-0.005789353	-0.000326536	0402	0.00579855	2	0402
0.0013987431	0.0000601151	0404	0.00140003	1	
0.0022995453	0.000350005	0405	0.00232602	1	

Figure 3: PCA result data set.

One consideration for using PCA to evaluate data in an ongoing study is that there will be some degree of missing data. Traditional PCA is not tolerant to missing data, because it is based on eigenvalue decomposition of the covariance matrix. Therefore, we've taken two approaches: Bayesian PCA and PCA with single-value decomposition (SVD) based imputation. Both of these methods allow for a greater amount of missing data. In practice, both methods will be run and the Biostatistician should review the results, and select which model to use based on the fewest number of components required to explain the variability of the indicator matrix, with a more reasonable result in terms of the cross-validation.

VISUALIZATION

The final step is to provide sponsors and CRAs a report of the results that is easy to review and interpret. Currently, different visualization approaches are being evaluated, but the final product will have the following attributes: integration of RBM Metrics with CSI results, easy to navigate for clinical staff, drill-down capabilities, email alerts, and tracking of corrective/preventative actions. The visualization of results from RBM Metrics and CSI is the main way this tool will help streamline monitoring for sponsors and CRAs. Figure 4 shows an example of some ways the risk score could be visualized, including graphics created with SAS/VA® software. In addition to these visualizations, some of the indicators described in Table 1 that are not used in the CSI overall risk tool can also be plotted for review by the sponsor and CRAs (e.g. a bar graph of the frequency of visits occurring on each day of the week, by site).

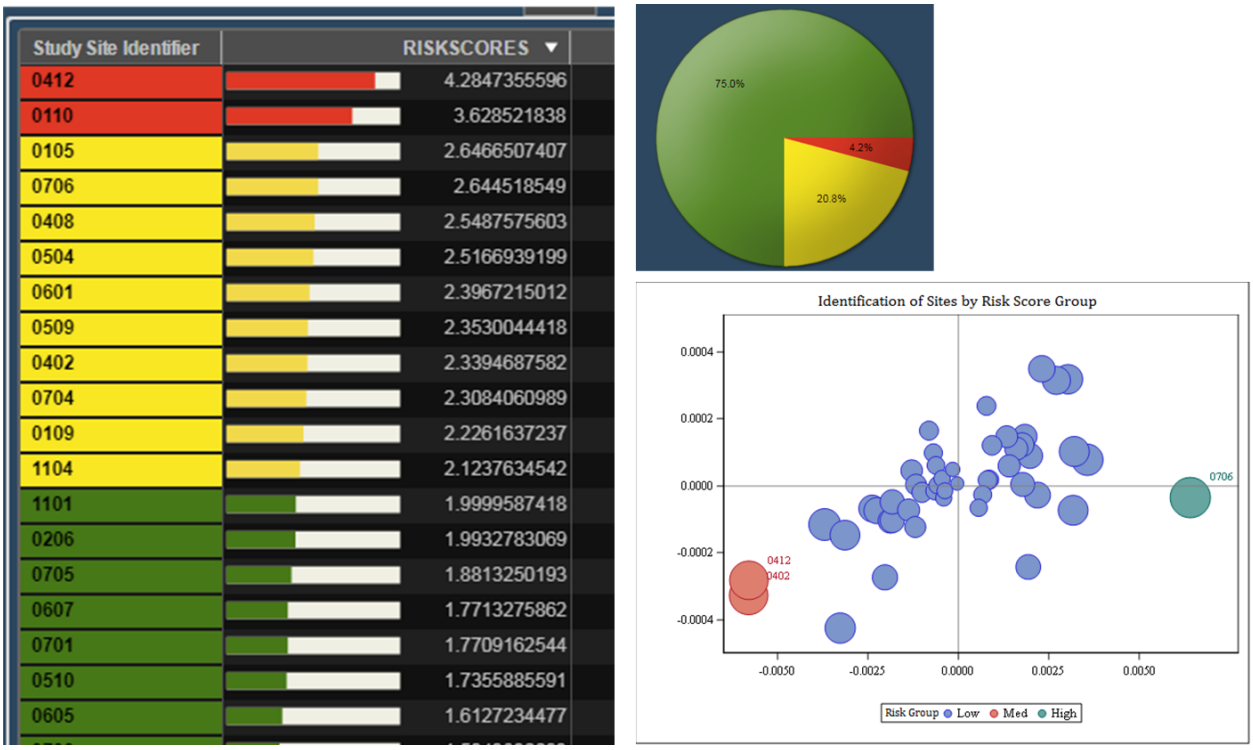


Figure 4: Examples of visualization of risk scores.

CONCLUSIONS

Guidance from regulatory bodies around the world suggest taking a risk-based approach to the monitoring of clinical trials. One way to implement a risk-based approach is to use centralized monitoring tools, such as review of the data against RBM metrics or grouping sites into risk categories using statistical methodologies, in order to focus monitoring efforts on sites that may compromise the safety of the trial subjects or the overall quality of the study data. Furthermore, the application of statistical models can streamline this process, providing a scientific basis for targeting monitoring efforts. The results of complex statistical models can be presented in a visual way to clinical staff, keeping all members of the study team updated on the study's progress. This data-driven approach provides a cost-effective way of meeting regulatory requirements, enhancing data review, and protecting subject welfare.

REFERENCES

1. International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use. (2016). E6(R2): Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Step_4.pdf
2. Sertkaya, A., et al. (2014). Examination of clinical trial costs and barriers for drug development. Submitted to U.S. Department of Health and Human Services. https://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development
3. US Food & Drug Administration. (2013). Guidance for industry: Oversight of clinical investigations – a risk-based approach to monitoring.
4. European Medicines Agency (2013) Reflection paper on risk based quality management in clinical trials.
5. TransCelerate BioPharma Inc. (2013). Position paper: Risk-based monitoring methodology.
6. Weir C. & Murray G. (2011). Fraud in clinical trials: detecting it and preventing it. Significance 8: 164-168.
7. Zink R. C. (2014). Risk-based monitoring of Clinical Trials using JMP® Clinical. White Paper. SAS Institute, Inc.
8. Zink R. C., Wolfinger R. D. & Mann G. (2013). Summarizing the incidence of adverse events using volcano plots and time windows. Clinical Trials 10: 398-406.
9. Venet D., Doffagne E., Beckers F., et al. (2012). A statistical approach to central monitoring of data quality in clinical trials. Clinical Trials 9: 705-713.
10. Kirkwood A., Cox T., Hackshaw A. (2013). Application of methods for central statistical monitoring in clinical trials. Clinical Trials 10: 783-806.
11. Wu X., Carlsson M. (2010). Detecting data fabrication in clinical trials from cluster analysis perspective. Pharm Statistics 10: 3435-3451.
12. Taylor R. N., McEntegart D., Stillman E. (2002). Statistical techniques to detect fraud and other data irregularities in clinical questionnaire data. Drug Information Journal 36: 115-125.

13. Akhtar-Danesh N. & Dehghan-Kooshkghazi M. (2003). How does correlation structure differ between real and fabricated data-sets? BMC Medical Reasearch Methodology 3(18): 1-9.
14. Pogue J. et al. (2012). Central Statistical Monitoring: Detecting fraud in clinical trial. Clinical Trials 13: 225-235.
15. Buyse M. et al. (1999). The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Statistics in Medicine 18: 3435-3451.
16. Bakobaki JM, et al. (2012). The potential for central monitoring techniques to replace on-site monitoring: findings from an international multi centre clinical trial. Clinical Trials 9: 257-264.

ACKNOWLEDGMENTS

The authors would like to acknowledge Lisa Comarella for her support of this project and review of this paper.

RECOMMENDED READING

Zink, Richard C. (2014) Risk-Based Monitoring and Fraud Detection in Clinical Trials Using JMP® and SAS®.
SAS Institute Inc. (2013) SAS/IML® 13.1 User's Guide.

CONTACT INFORMATION

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

Laura Williams
CROS NT, LLC
Chapel Hill, NC, USA
+1 (919) 929-5015
laura.williams@crosnt.com
www.crosnt.com