

Database, 2019, 1–10 doi: 10.1093/database/baz017 Original article



Original article

A cross-source, system-agnostic solution for clinical data review

Michael A. Farnum^{*}, Mathangi Ashok, Daniel Kowalski, Fang Du, Lalit Mohanty, Paul Konstant, Joseph Ciervo, Victor S. Lobanov and Dimitris K. Agrafiotis¹⁰

Covance, the Drug Development Division of LabCorp, 210 Carnegie Center, Princeton, NJ 08540, USA

*Corresponding author: Tel: 609-419-2514; Email: mafarnum@gmail.com.

Citation details: Farnum, M.A., Ashok, M., Kowalski, D. *et al.* A cross-source, system-agnostic solution for clinical data review. *Database* (2019) Vol. 2019: article ID baz017; doi:10.1093/database/baz017

Received 9 November 2018; Revised 24 December 2018; Accepted 19 January 2019

Abstract

Assembly of complete and error-free clinical trial data sets for statistical analysis and regulatory submission requires extensive effort and communication among investigational sites, central laboratories, pharmaceutical sponsors, contract research organizations and other entities. Traditionally, this data is captured, cleaned and reconciled through multiple disjointed systems and processes, which is resource intensive and error prone. Here, we introduce a new system for clinical data review that helps data managers identify missing, erroneous and inconsistent data and manage queries in a unified, system-agnostic and efficient way. Our solution enables timely and integrated access to all study data regardless of source, facilitates the review of validation and discrepancy checks and the management of the resulting queries, tracks the status of page review, verification and locking activities, monitors subject data cleanliness and readiness for database lock and provides extensive configuration options to meet any study's needs, automation for regular updates and fit-for-purpose user interfaces for global oversight and problem detection.

Introduction

The goal of clinical trials is to demonstrate the efficacy, safety and comparative effectiveness of new investigational treatments or existing therapies that warrant further study. Data management plays a crucial role in this process, ensuring that the data collected is complete, accurate and delivered in a timely manner for analysis, submission and disclosure (1). That process involves many different groups, such as clinical investigators and site staff, patients, laboratory and imaging vendors, device and technology companies and more, entering data into a multitude of disparate systems. Of central importance is the electronic

© The Author(s) 2019. Published by Oxford University Press.

Page 1 of 10

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/),

which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

data capture (EDC) system used to capture patient observations at investigational sites. That data undergo a significant amount of scrutiny, from ensuring that site personnel enter the information in a timely and complete manner, to supporting queries and other forms of communication to address concerns for accuracy and consistency, to facilitating multiple rounds of review, particularly for verification of correct transcription from the original source. To aid in this process, the data are inspected through a number of programmed checks, which fall into two categories: (i) those implemented and managed within a particular source system (typically the EDC) and (ii) those implemented and managed outside of any individual source. The latter usually require data to be combined from multiple sources and are implemented in SAS, R or Microsoft Excel, as they may require fairly complex logic. In both cases, the output of these checks is reviewed by data managers who delegate any suspect data to the appropriate parties for verification and correction.

Managing this process is very inefficient and error prone. First, as non-EDC (laboratory, biomarker, imaging, patientreported, etc.) data account for the majority of the information collected during a clinical trial, queries against those systems or vendors involve a different workflow from that used in the EDC (often relying on emails or telephone calls.) Second, whereas the EDC checks are typically implemented by the data managers who set up the electronic case report forms, the more complex checks and discrepancy logic are implemented by statistical programmers.

One approach to addressing these problems is to bring all the data into the EDC and implement both simple and complex checks within that platform. While this may simplify code management and reuse, it is constrained by the performance and scalability of today's EDC systems, which were developed many years ago and designed for much smaller volumes of data. More critically, this solution increases the EDC setup time protracting its go-live date and does not solve the communication problem with external vendors who do not use an EDC to capture their data, thus complicating query management and resolution.

A second approach, and the one adopted here, is to integrate all clinical trial data into a data warehouse and use that to drive data management activities. This approach is best exemplified by Oracle's Data Management Workbench (DMW), (2) which relies on Oracle's Life Sciences Hub (LSH) (3) as the underlying data repository. DMW offers a number of important features, including integrated access to clinical trial data, configurable data management workflows, extensive query, discrepancy and library management capabilities, tight integration with Oracle's InForm EDC, (4) and more. While DMW does, in principle, enable integration with other EDC systems, we are not aware of any sponsor who has actually reduced this configuration to practice. Perhaps the greatest limitation of DMW is its reliance on LSH, a clinical data warehouse that consists of thousands of tables, which makes it challenging to deploy and maintain.

Recently, we introduced a comprehensive application suite, known as Xcellerate, that uses advanced data integration, analytics and visualization capabilities to improve patient safety, data quality and protocol compliance throughout the clinical development process and enable greater transparency and oversight of study conduct and performance (5, 6). The system consists of a number of enduser applications connected to a clinical data repository that supports near-real-time acquisition, mapping and integration of clinical trial data from any germane source (7, 8). The solution described herein leverages this clinical data repository to streamline and automate the process of delivering high-quality study data and associated metrics for downstream use. It consists of the following four webbased applications: (i) the 'Discrepancy Manager', which facilitates review of programmed discrepancy check output and bulk handling of the resulting queries to sites and thirdparty data vendors; (ii) the 'Page Tracker', which facilitates reporting of missing EDC pages and the backlog of outstanding work for page review, verification and locking; (iii) the 'Query Tracker', which provides reporting and analytics for query management and root cause analysis; and (iv) the 'Subject Tracker', which tracks subject data cleanliness and readiness for database lock and helps data management teams manage outstanding work.

Methods

Page Tracker

Page Tracker facilitates the execution and reporting of data entry into the EDC. One of the core functionalities of the tool is the ability to define the visit, page entry and source data verification (SDV) schedules, and the rules that determine which data should be expected when and under what conditions. We use the following three types of rules:

(i) Visit date rules, which determine the time when forms related to a specific study visit should be expected. These rules allow the use of multiple baselines and reference points to support complex study designs that are becoming increasingly prevalent in drug development, such as oncology treatment cycles that may have different schedules for subjects assigned to different study arms.

(ii) Page entry rules, which determine data entry conditions for EDC pages. These rules are used to define the subset of patients and/or pages applicable to a given visit and may utilize any type of data coming from either EDC or non-EDC sources to set the appropriate flag. The most common example is to specify pages applicable to only one gender.

(iii) SDV rules, which define conditions for source data verification. These conditions are originally defined in the EDC system when the study is set up, and we use the EDC metadata to populate the corresponding entries in the SDV schedule.

Once all the rules are defined, the visit, page entry and SDV schedules are initialized from the EDC metadata and refined interactively using the Page Tracker user interface (UI). In addition, page categories and page actions are specified to help review the status of data cleanliness and completeness (i.e. show how many expected pages are pending, completed, verified, reviewed, frozen and locked). The page entry scheduler is illustrated in Figure 1 and a representative visit date rule is shown in Figure 2.

By applying these predefined rules, schedules, categories and actions on the latest snapshot of the data in our clinical data repository [which is refreshed nightly from the source systems though automated extraction, transformation and loading (ETL) pipelines], the system tracks the amount of completed and remaining work and allows data managers and clinical research associates (CRAs) to identify sites, subjects and forms with pending actions. The actual calculations can be scheduled to run automatically at any desired frequency or triggered manually through the UI at any time. When completed, the results (including form, page, site, subject and action summaries) become available through the UI and the appropriate personnel are notified by email.

The tool also allows users to define and track arbitrary cohorts that may be of special interest or priority at some point during the trial. These subsets, commonly referred to as data cuts, may be limited to specific subjects, visits, dates or any combination thereof.

Query Tracker

Queries are an essential component of any modern EDC, indicating potential issues in the data collected. The number of queries and the elapsed time before a query is acknowledged and closed are important determinants of study and site risk and performance (9). Inadequate visibility into such metrics can result in poor data quality and delayed deliverables.

The Query Tracker (Figure 3) is an application that provides query metrics (listings, counts and percentages, cycle times) organized by site, subject, form, marking group



Figure 1. Page entry schedule editor in Page Tracker.

Date Definition Name V7_012 Description Phone Visit 1 expected date Color Shape Amber Rectangle \sim Phone visit 1 happens 12 weeks after ref_date if the subject is still in study 11 2 3 var numdays = 12*7; 4 var dtDate = [ref_date]; 5 var eos = [Date_EoS]; 6 if (dtDate != null && dtDate.HasValue && eos == null) 7 . { 8 return dtDate.Value.AddDays(numdays); 9 } 10 else if (dtDate != null && dtDate.HasValue && eos.HasValue && eos > dtDate.Value.AddDays(numdays)) 11 -{ 12 return dtDate.Value.AddDays(numdays); 13 } 14 else 15 -{ 16 return null; 17 3 18

Figure 2. An illustrative visit date rule defined in the Page Tracker. This particular rule specifies that phone visit 1 must occur 12 weeks after the reference date if the subject is still in the study. The rule utilizes two dates: the reference date for this specific visit and the end-of-study (EOS) date. If the reference date is not null and the EOS date is null, phone visit 1 is expected 12 weeks after the reference date. If EOS is not null but happens after the expected phone visit date, the rule returns this expected date, otherwise it returns null to indicate that a phone visit is not expected for this subject because the study has already finished.

(system, user, etc.), status (open, answered, closed), age (days open) and recipient. The tool provides drill-down to individual query details, extensive searching, sorting, filtering, exporting capabilities and direct integration with the Xcellerate Risk and Issue Management (RIM) system (10) to escalate any problematic findings into issues and actions for mitigation.

Besides keeping track of the workload, these query metrics can help users identify systemic problems and trends, and possible solutions. For example, looking at the total counts and percentages of open queries per page or form allows users to identify potential misfiring checks due to logic or programming errors, issues with EDC page design or gaps in site training. Likewise, an excessive number of open queries per site or subject, a high ratio of open queries to pages entered or excessive aging of queries at a given site may be a sign of poor site performance. While spikes should be expected due to the periodic nature of data review, sustained trends over a prolonged period of time could indicate quality issues and higher site risk.

Discrepancy Manager

While EDC edit checks can work well for simple data entry issues, more advanced checks are still needed to ensure overall data quality. These include data reconciliations, where information is compared across different data sources (e.g. EDC vs. Interactive Voice Recognition System (IVRS), central laboratory or ePRO), programmable protocol deviations and other complex checks that are difficult or impossible to program within the EDC itself. These checks are typically implemented and executed in external software packages like SAS, and their output is delivered to data managers for review and follow up. As this output usually comes in the form of Excel listings, tracking and updating the discrepancy list is very labor intensive and error prone.

The Discrepancy Manager (Figure 4) is a web-based, multi-user, collaborative tool that can import discrepancies identified through any external application or process and manage their entire lifecycle within the system.

Sponsor Account194	Protocol ✓ 000000111598	~									1	XLS	XLS ALL	
QUERY AGING REPORT	QUERY AGING REPORT BY SITE	FORMS MO			MOST OPEN UERIES	OPE 14-2			EN 60- 179	OPEN 180+		QUERIES		ED QUERIES TING
 Total metrics 	Open query metrics	Answere	ed query metri	cs 🛛	Closed query	metrics								
										Tot	al metrics			
		Q	uery status - a	ll queries		_	Query state	us - queries op	ened in last	7 days		Query stat	us - queries op	ened in last
uery Recipient	Total	Open	Answered	Closed	Cancelled	Total	Open	Answered	Closed	Cancelled	Total	Open	Answered	Closed
otal	97545	34702	27720	30162	4961	0	0	0	0	0	0	0	0	0
DM by ME	748	258	229	245	16	0	0	0	0	0	0	0	0	0
M by System	1600	670	144	520	266	0	0	0	0	0	0	0	0	0
linical Site by DM/SSG	31789	11665	8609	8754	2761	0	0	0	0	0	0	0	0	0
linical Site by CRA	16551	5757	5059	5095	640	0	0	0	0	0	0	0	0	0
uery Recipient	Count	Med	lan A.											
otal	count				May	0.3	3.7	7-28	28.00		90-190	19	0.365	365+
	5277	775		/erage 75.2	Max 1516	0-3 0	3-7	7-28 0	28-90 84		90-180 141	18 33	0-365 9	365+ 4489
DM by ME	5277 16		77		1000000			10.000				1000		200 Scott
		775	77 0.4 11	75.2	1516	0	0	0	84		141	33		4489
	16	775 121 616	77 0.4 11	75.2 133.5 56.4	1516 1386	0 0 0	0	0	84 0		141 0	339 0		4489 16
OM by System	16 185	775 1211 616.	77 0.4 11 9 65 ed query metri	75.2 133.5 56.4 cs	1516 1386 1087 Closed query 1	0 0 0 metrics	0 0 0 Answere	0 0 0 td query metric	84 0 12		141 0	339 0		4489 16
DM by System	16 185	775 1211 616.	77 0.4 11 9 65 ed query metri Open till Answe	75.2 133.5 56.4 cs	1516 1386 1087 Closed query I	0 0 0 metrics	0 0 0 Answere till Now(outs	0 0 0 ed query metric standing)	84 0 12 s	Da	141 0 7 iys outstane	339 0 25 ding answer	9 red	4489 16 128
M by System Total metrics Query Recipient	16 185 Open query metrics	775 1210 616. Maswere Time from C Average	77 0.4 11 9 65 ed query metri Open till Answe	75.2 133.5 56.4 cs	1516 1386 1087 Closed query f Time from Count	0 0 0 metrics Answered Average	0 0 0 Answere till Now(outs Median	0 0 cd query metric standing) Max 0-3	84 0 12 s	Da 7-28	141 0 7 ys outstand 28-90	339 0 25 ding answer 90-180	9 red 180-365	4489 16 128 365+
M by System Total metrics Query Recipient Total	16 185 Open query metrics Count 27720	775 1210 616. Answere Time from C Average 24.3	77 0.4 11 9 65 ed query metri Open till Answer Median 5.6	75.2 133.5 56.4 cs	1516 1386 1087 Closed query P Time from Count 2 234	0 0 metrics Answered Average 521.8	0 0 0 Answere till Now(outs Median 1 712.7	0 0 sd query metric standing) Max 0-3 1250 0	84 0 12 s 3-7 0	Da 7-28 : 0 0	141 0 7 ays outstand 2 8-90 0	339 0 25 ding answer 90-180 0	9 red 180-365 0	4489 16 128 365+ 207
DM by System Total metrics Query Recipient Total DM by ME	16 185 Open query metrics Count 27720 229	775 1211 616. Answere Time from 0 Average 24.3 88.7	77 0.4 11 9 65 ed query metri Open till Answer Median 5.6 82.1	75.2 133.5 166.4 cs red Max 800 504	1516 1386 1087 Closed query P Time from Count 2 234 6 0 0	0 0 metrics Answered Average 521.8 0.0	0 0 0 Mnswere till Now(outs Median 1 712.7	0 0 td query metric standing) Max 0-3 1250 0 0 0	84 0 12 s 3-7 0 0	Da 7-28 : 0 (0 (141 0 7 ays outstand 28-90 0 0	334 0 25 ding answer 90-180 0 0	9 red 180-365 0 0	4489 16 128 365+ 207 0
OM by System Total metrics Query Recipient Total OM by ME OM by System	16 185 Open query metrics Count 27720	775 1211 616. Answere Time from O Average 24.3 88.7 0.8	77 0.4 11 9 65 ed query metri Open till Answer Median 5.6	5.2 33.5 56.4 ccs □ wered Max 800 504 168	1516 1386 1087 Closed query P Time from Count 2 234 6 0 0	0 0 0 metrics Answered Average 521.8 0.0 0.0 381 3	0 0 0 Mnswere till Now(outs Median 1 712.7	0 0 sd query metric standing) Max 0-3 1250 0	84 0 12 s 3-7 0	Da 7-28 : 0 (0 (141 0 7 ays outstand 2 8-90 0	339 0 25 ding answer 90-180 0	9 red 180-365 0	4489 16 128 365+ 207
DM by System Total metrics Query Recipient fotal DM by ME DM hv Svetem	16 185 Open query metrics Count 27720 229 144	775 1211 616. Answere Time from O Average 24.3 88.7 0.8	77 0.4 11 9 65 ed query metri Open till Answer Median 5.6 82.1 0 1	5.2 33.5 56.4 ccs □ wered Max 800 504 168	1516 1386 1087 Closed query 1 Time from Count // 234 6 0 0 0	0 0 0 metrics Answered Average 521.8 0.0 0.0 381 3	0 0 0 till Now(outs Median 712.7 0.0 0.0 0.0	0 0 td query metric standing) Max 0-3 1250 0 0 0	84 0 12 s 3-7 0 0	Da 7-28 : 0 (0 (141 0 7 ays outstand 28-90 0 0	334 0 25 ding answer 90-180 0 0	9 red 180-365 0 0	4489 16 128 365+ 207 0
DM by System Total metrics Query Recipient fotal DM by ME DM hv Svetem	16 185 Open query metrics Count 27720 229 144	775 1210 616. ✓ Answere Time from C Average 24.3 88.7 0.8 ○ Answere	77 0.4 11 9 65 ed query metri Open till Answer Median 5.6 82.1 0 1	75.2 33.5 56.4 ccs □ reed Max 800 504 168 504	1516 1386 1087 Closed query I Count / 234 0 0 0 5 s Closed query I	0 0 0 metrics Answered Average 521.8 0.0 0.0 381 3	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 8d query metric standing) Max 0-3 1250 0 0 0 0 0 0	84 0 12 s 3-7 0 0 0	Da 7-28 : 0 (0 (141 0 7 ays outstand 28-90 0 0	333 0 25 ding answer 90-180 0 0 0	9 red 180-365 0 0	4489 16 128 365+ 207 0 5
DM by System Total metrics Query Recipient Total DM by ME DM by ME DM by Vector Total metrics	16 185 Open query metrics Count 27720 229 144 Open query metrics	775 1211 616. ☑ Answere 24.3 88.7 0.8 △ Answere Time Average	77 0.4 11 9 65 9 65 9 65 9 65 9 06 10 0 10 0 10 10 0 10 0 1	75.2 33.5 36.4 cs max 800 504 168 cs Answered	1516 1386 1087 Closed query I Count / 234 0 0 0 5 s Closed query I	0 0 0 metrics Answered Average 521.8 0.0 0.0 381 3	0 0 0 Massere till Now(outs Median 1712.7 0.0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 standing) Max 0-3 1250 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	84 0 12 s 3-7 0 0 0 0 0	Da 7-28 : 0 (0 (141 0 7 ays outstand 28-90 0 0	333 0 25 ding answer 90-180 0 0 0	e red 180-365 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	4489 16 128 365+ 207 0 5
Query Recipient Total DM by ME DM by System	16 185 Open query metrics Count 27720 229 144 Open query metrics	775 1211 616. ☑ Answere 24.3 88.7 0.8 △ Answere Time Average	77 0.4 11 9 65 9 65 9 65 9 65 9 06 10 0 10 0 10 10 0 10 0 1	75.2 33.5 36.4 cs □ Max 800 504 168 cs ☑ Answered lian	1516 1386 1087 Closed query I Time from Count / 234 C 0 Cosed query Closed query	0 0 0 metrics Answered Average 521.8 0.0 881 3 metrics	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 standing) Max 0-3 1250 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	84 0 12 5 3-7 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Da 7-28 : 0 (1 0 (1 0 (1)	141 0 7 2 28-90 0 0 0	333 0 25 ding answer 90-180 0 0 0 0	ed 180-365 0 0 0 0 0 0 0 0 0 0 0 0 0	4489 16 128 365+ 207 0 5 ed

Figure 3. Representative screenshots of the Query Tracker.

140

8403

5051

11000

2544

0.1

6.9

6

0.0

0.8

183

16.6

7.4

13.9

168

800

334

547

332

139

8454

5036

11086

2542

15.3

The tool automatically recognizes any new and updated records and utilizes the web services application programming interfaces (API) that are available in most leading EDCs to open, close or cancel queries, either individually or in bulk, directly from its UI. This obviates the need for data managers to navigate to a specific study, site, subject, visit, form and field within the EDC and manage each query individually, introducing tremendous efficiency in the query management process. For vendors who do not provide query management and API integration capabilities (laboratories, biomarkers, imaging, etc.), the system can send and receive emails with attachments containing the selected discrepancies and capture the entire sequence of every incoming and outgoing email communication for every query or discrepancy originating within the system. The system provides a full audit trail of every action taken within the tool and makes it readily available for inspection during an audit.

Subject Tracker

DM by System

Clinical Site by DM/SSG

Clinical Site by CRA

Clinical Site by ME

Clinical Site by System

Finally, the Subject Tracker (Figure 5) provides comprehensive patient-level summaries of data cleanliness and completeness to help assess readiness for database lock. The information displayed includes basic milestones and target dates (treatment allocation date, projected end of treatment date, last visit date, etc.) and aggregate statistics on page entry, review, verification, freezing and locking activities, queries and reconciliation activities, etc.

8754

5095

12949

2599

16.8

0.0

7.9

4.7

193

31.7

10.6

16.9

250

759

411

586

331

Architecture

0.7

0.0

1.8

426

350

586

288

Our solution consists of several components integrated into a complete, end-to-end system, as illustrated in Figure 6. All source data are initially staged, transformed and loaded into the Xcellerate Operational Data Warehouse (ODW) (7). Protocol, site, subject, query and page data are loaded from the Clinical Trial Management System (CTMS), Interactive Voice/Web Recognition System (IXRS) and EDC through automated ETL processes, whereas other data, such as SAS discrepancy listings and EDC coding and report data that is not available through the EDC APIs, are loaded using a custom file loader. Once the data is loaded, all necessary calculations, including those for page entry and SDV status, are performed by a scheduling engine based

ISCREPANCY DETAILS							1.	SUMMARY V DISCREPANCIE	S			
Subject# Subject#		Status	Assigned to me COLUMNS - BULK UPD									ULK UPDATE -
		Queried	Rec	Site	Subject	Status	Reviewer	Comments	Query status	Email	CHECKID	Quei
screpancy Id	Check fired	Update date	1	003	201762301	Queried	✓ Daniel Kowalski	~	Open		DA_021	The •
48	No	May 15, 2018, 12:07:27 PM	2	003	124002002	Queried	> Daniel Kowalski	✓ Programming upda <u>+1 m</u>	ore		BIOP_SCN_042	The -
IELDO1 FIELDO2 2 APR 2017 Specify number of tablets taken since last dispensation=Not Present IELDO6 FIELDO7		FIELD03	3	003	201762301	Queried	> Daniel Kowalski	~	Open	Sent at Nov 9, 2017, 4:10:19 PM	LB_URIN_047	The
			4	003	201762301	Queried	Y Daniel Kowalski	~	Open	Sent at May 3, 2018, 10:36:15	LB_CHEM_041	The
			5	003	826006003	Queried	Y Daniel Kowalski	~	Open	Sent at May 2, 2018, 10:09:53	BIOP_SCN_042	The
		FIELD08	6	003	201762301	Queried	✓ Daniel Kowalski	~	Open	Sent at May 2, 2018, 10:09:53	LB_URIN_047	The
AREGEDKEY Page Repeat Number DIMDA_02120176230112345671V2dadis_dtfl2 PR 20175pecify number of tablets taken		Query Field Name	□ 7	0002	250001009	Queried	Michael Farnum	~	Open	Sent at May 8, 2018, 12:41:07	LB_URIN_047	The
		DADISPDT	8	0002	999907	Auto Closed	✓ Michael Farnum	~			LB_URIN_047	The -
nce last dispensation=Not Present			9	0002	250001009	Closed	Y Michael Farnum	V CRA already querie			LB_URIN2_048	The -
iery text	Record Position	SITEID	10	0002	250001009	Queried	Y Michael Farnum	~		Sent at May 8, 2018, 12:46:21	LB_CHEM_041	The
he actual dispense visit date at the prior 1 isit is provided, however the number of ablest taken since last dispensation is		003	11	0002	250001009	Auto Closed	Michael Farnum	~			LB_URIN_047	The -
			12	0002	TEST123	Queried	Michael Farnum	~	Cancelled		LB_HEME_046	The
issing, Please reconcile.	-		13	0002	250001009	Open	Michael Farnum	~			LB_URIN2_048	The
otal time Queried time 99 6		Answered time	14	0002	250001009	Queried	Y Michael Farnum	~		Sent at Mar 27, 2018, 7:17:01	LB_HEME_046	The
			15	0002	250001009	Queried	Michael Farnum	~	Open		DA_021	The
	VIMENTS QUERIES EMAILS	Queried		COAN	ICE. Discr	epancy Status Ma	Tagement Covan	any 952 DIM te No: 000000555555 V	PS			 Daniel Ko
May 15, 2018, 12:07:27 PM Dis		Queried Open				epancy Status Ma	Tagement Covan	ery 952 DM er No: cococossssss V SUMMARY V DISCREPANCI	ES		Lest deta k	Daniel Ko
May 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM Nev ID: 46 Key: 123410 The actual dispense visit date at t	icrepancy Status Change Lynette Thomas w Query Lynette Thomas the prior visit is provided, however the number	Open of tablets taken since last dispensa	Tota	al Summary		epancy Status Ma	Tagement Covan	te No: 000000555555		Dre @Cosed @Auto Dosed	Last data ic	Daniel K
May 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM Nev ID: 46 Key: 123410 The actual dispense visit date at t	krepancy Status Change Lynette Thomas w Query Lynette Thomas	Open	Tota	al Summary	Open Queried C	Dn Hold Unresolvable	Coverd Auto Coverd	SUMMARY of DISCREPANCI		tre @Closed @Auto Closed	Last data ic	Daniel Ko
May 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM Nev ID: 46 Keyr 123410 The actual dispense visit date at t Apr 16, 2018, 349:16 PM Qui 25 Keyr 123254	icrepancy Status Change Lynette Thomas w Query Lynette Thomas the prior visit is provided, however the number	Open of tablets taken since last dispensa	Tota	al Summary	Open Queried C	Dn Hold Unresolvable	Coverd Auto Coverd	SUMMARY of DISCREPANCI		See Cosed @Aus Cosed	Last data le	Daniel Ko
Mey 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM No. Dis: 46 46 Keys: 12:2410 The actual dispense visit date at tr Apr 16, 2018, 3:46:16 PM Qui Dis: 23 Keys: 12:22:54 Mar 1, 2018, 33:137 PM Dis	vropenny Status Change Lynette Thomas w Query Lynette Thomas the prior visit is provided, however the number ery Status Change Daniel Kowalski	Open of tablets taken since last dispensa Cancelled	Tota	al Summary	Open Queried C	Dn Hold Unresolvable	Coverd Auto Coverd	SUMMARY of DISCREPANCI		Se Cosed Dius Cosed	Last data le	Daniel Ko
May 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM Nen May 16, 2018, 33:61:61 PM Qui Dio 25 Keyr 12:32:54 Mar 1, 2018, 33:137 PM Dis IDi 25 Keyr 12:32:54	v Query Lynette Thomas w Query Lynette Thomas the prior visit is provided, however the number ery Status Change Daniel Kowalski crepancy Status Change Daniel Kowalski	Open Open of tablets taken since last dispense Cancelles Querret Open	Tota dion i To	al Summary	7 <mark>Open Queried C</mark> 10 57	Dn Hold Unresolvable	Coverd Auto Coverd	SUMMARY of DISCREPANCI		tre Cosed @AutoCosed	Last data le	Daniel Ko
May 15, 2018, 12:07:27 PM Dis May 15, 2018, 12:07:27 PM Nen The actual dispense visit date at t Arr 16, 2018, 3:46:16 PM Qui Dio 25 Kign 12:22:54 Mar 1, 2018, 3:31:37 PM Dio 25 Kign 12:22:54 Nen The actual dispense visit date at t 12:25:4	korepancy Status Change Lynette Thomas w Query Lynette Thomas the prior visit is provided, however the number ery Status Change Daniel Kowalski korepancy Status Change Daniel Kowalski w Query Daniel Kowalski	Open Open of tablets taken since last dispense Cancelles Querret Open	Tota dion i To	al Summary ttal Heev 86 5	7 <mark>Open Queried C</mark> 10 57	in fadd Uswaschabfe 4 1	Coverd Auto Coverd	a in concessors 4			en ©Queried	Daniel Ko

Figure 4. Representative screenshots of the Discrepancy Manager.

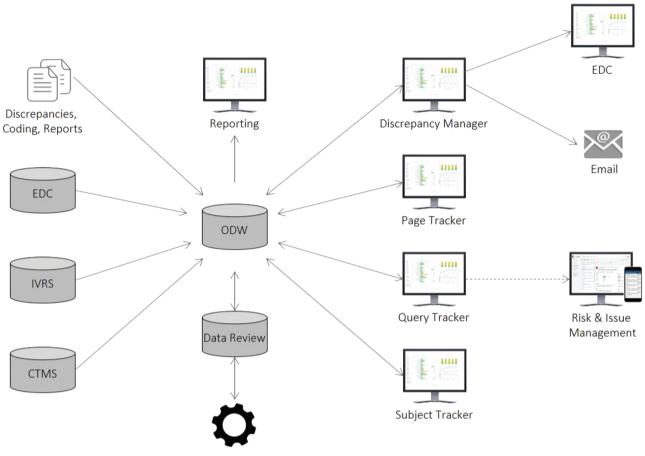
ecker 1		~			II. PATIEN	SUMMARY Q DATA MC	INITOR				
Assigned to me											COLUMNS -
Subject	Site	Investigator	Country	✓ Reviewer	✓ Review Date	Notes and Issues	Cleaning Batch	Patient Status	Death	Projected Er	
390045002	39004	Fname29234 Lname29234	israel	Michael Farnum	✓ 15/Nov/2017	Waiting for Central Lab transf	Priority 1: Death Subjects	End of Treatment	Y	09/Sep/201	Excel Export
390045001	39004	Fname29234 Lname29234	Israel	MaryJo Gore	✓ 15/Nov/2017		Priority 1: Death Subjects	End of Treatment	Y	06/Apr/201	CSV Export
390045003	39004	Fname29234 Lname29234	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 3: Close Queries	Screen Failed	N		
560045001	56004	Fname61119 Lname61119	Korea, South	Lynette Thomas	~	No answered queries	Priority 3: Close Queries	Discontinued Treatment	N	30/Oct/201	RIM
560045002	56004	Fname61119 Lname61119	Korea, South	Lynette Thomas	~	No answered queries	Priority 3: Close Queries	End of Treatment	N	14/Nov/201	7
390015001	39001	Fname40581 Lname40581	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 1: Death Subjects	Screen Failed	N		
390045004	39004	Fname29234 Lname29234	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 3: Close Queries	End of Treatment	N	05/Feb/2018	3
390035001	39003	Fname158674 Lname158674	Israel	MaryJo Gore	✓ 15/Nov/2017		Priority 1: Death Subjects	Discontinued Treatment	Y	30/Dec/201	7
560055001	56005	Fname21630 Lname21630	Korea, South	Select reviewer	✓ 28/Nov/2017	No answered queries	Priority 3: Close Queries	End of Treatment	N	04/Feb/2018	3
560015001	56001	Fname35902 Lname35902	Korea, South	Select reviewer	~	No answered queries	Priority 3: Close Queries	End of Treatment	N	18/Feb/2018	3
260035001	26003	Fname12240 Lname12240	Canada	Mathangi Ashok	~	No answered queries	Priority 3: Close Queries	Screen Failed	N		
560025001	56002	Fname39311 Lname39311	Korea, South	Select reviewer	~	No answered queries	Priority 3: Close Queries	Discontinued Treatment	N	16/Mar/201	8
140125001	14012	Fname28721 Lname28721	United States	Marie Tavani	~	No answered queries	Priority 3: Close Queries	Screen Failed	N		
390065001	39006	Fname37682 Lname37682	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 3: Close Queries	Discontinued Treatment	N	20/Mar/201	8
440015001	44001	Fname29296 Lname29296	Austria	Select reviewer	~	No answered queries	Priority 3: Close Queries	End of Treatment	N	15/Apr/2018	3
140065001	14006	Fname12245 Lname12245	United States	Marie Tavani	~	No answered queries	Priority 3: Close Queries	End of Treatment	N	23/Apr/2018	3
140185001	14018	Fname33913 Lname33913	United States	Marie Tavani	Ŷ	No answered queries	Priority 3: Close Queries	Screen Failed	N		
390125001	39012	Fname36512 Lname36512	Israel	MaryJo Gore	~		Priority 1: Death Subjects	End of Treatment	Y	22/Apr/2018	3
390085001	39008	Fname27509 Lname27509	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 3: Close Queries	Screen Failed	N		
390125002	39012	Fname36512 Lname36512	Israel	Danilo Branco	✓ 15/Nov/2017	No answered queries	Priority 1: Death Subjects	Discontinued Treatment	Y	24/May/201	8
120025001	12002	Fname20343 Lname20343	United Kingdom	Yuyang Sun	~	No answered queries	Priority 3: Close Queries	Screen Failed	N		
240145001	24014	Fname49619 Lname49619	Spain	Select reviewer	~	No answered queries	Priority 1: Death Subjects	End of Treatment	Y	05/May/201	8
390065002	39006	Fname37682 Lname37682	Israel	MaryJo Gore	✓ 15/Nov/2017	No answered queries	Priority 1: Death Subjects	End of Treatment	Y	19/May/201	8
240045001	24004	Fname65851 Lname65851	Spain	Select reviewer	~	No answered queries	Priority 4: SF Cleaning	Screen Failed	N		
140245001	14024	Fname66006 Lname66006	United States	Marie Tavani	~	No answered queries	Priority 3: Close Queries	End of Treatment	N	19/May/201	8

Rows per page: 50 ♀ 1 - 50 of 196 rows 1 of 4 pages | < < > >|

Figure 5. Representative screenshot of the Subject Tracker.

on study-specific settings and rules configured through the application UIs during study setup. The results are stored in the data review database (DRDB) and fed back into ODW

for reporting purposes and are immediately accessible and actionable by central monitors through the application UIs. Direct integrations with the EDC and RIM systems allow



Calculation Service

Figure 6. Xcellerate Data Review architecture and data flow.

users to raise queries and generate issues and actions as they review the results. The Query Tracker integrates with RIM through a software development kit that utilizes RIM's Representational State Transfer (REST) APIs to allow data managers to create issues and actions that become immediately available in the RIM (10) and CRA Dashboard (7) UIs to enable communication, collaboration and rapid follow-up. The Discrepancy Manager integrates with any EDC, including Rave (11) and InForm, (4) using available web services APIs in a loosely coupled, platform-agnostic manner, which allows users to raise queries in the EDC in real time. Responses received in the EDC are loaded onto ODW using the EDC ODM API and from there onto the DRDB using the ETL process mentioned above. For vendors who do not provide query management and API integration capabilities, the system allows for email creation and ingestion through a standard email server. This enables the system to send and receive emails with attachments containing the selected discrepancies and capture the entire sequence of every incoming and outgoing email communication in DRDB.

The physical architecture consists of multiple independent services, implemented using a standard three-tier architecture with shared data stores. The ODW and DRDB are implemented on SQL Server 2014 (12). Data integration is implemented through a combination of Informatica ETL (13) and a custom .Net file loader. The application layer, including the RESTful APIs and calculation service, are also implemented in .Net. The application layer uses Entity Framework (14), Dapper MicroORM (15) and ADO.Net (16) to access the data layer and the open-source Hangfire framework (17) to provide background job processing for the calculation engine. The presentation layer consists of browser-based, single-page web applications based on the Angular 2 JavaScript framework (18) along with additional third-party UI control libraries. Visualizations in all applications were implemented using the Xcellerate JavaScript visualization library that was originally developed for the Xcellerate Medical Review application (5).

Results

The problem of missing data in clinical trials has been widely discussed (19). Although there has been extensive research on the use of imputation, weighting, slope estimation and other methods to reduce the impact of missing data (minimize bias and preserve the statistical power of the study), very little has been published on its prevention (20). As stated in EMA's Guideline on Missing Data in Confirmatory Clinical Trials: 'it is extremely important to avoid the presence of unobserved measurements as much as possible, by favouring designs that minimize this problem, as well as strengthening data collection regardless of the patient's adherence to the protocol' (21). Efficient review and management of data queries triggered by SDV, programmed edit checks, header reconciliation and database lock activities can help minimize data quality issues, cost overruns and delays. However, to improve the prevention of these issues, one needs to identify their root cause. A comprehensive data management approach should include ongoing analysis and mitigation of the most frequent sources of queries and missing data, both from an operational (which sites/ subjects) and a quality-by-design (which forms/procedures) perspective.

The system presented herein includes a combined database of subject visits from study protocol and IXRS, pages and queries sourced from EDC, discrepancies from data checks programmed in SAS or other external tools and protocol deviations from CTMS or other issue management systems. A purpose-built graphical UI allows data managers to identify sites, EDC forms or study procedures responsible for higher query rates, discrepancies and protocol deviations. The interface also flags sites that fall behind on data entry and query resolution. Further, integration with the ODW and EDC allows visual and interactive review of identified data discrepancies with an integrated capability to generate data queries to sites and any resulting findings are communicated to central and site monitoring staff for action assignment via our RIM system.

Our solution helps reduce avoidable data errors and loss by providing timely and integrated access to all clinical trial data, eliminating unnecessary and duplicative work, enabling communication, collaboration and oversight, facilitating reporting and trending and enhancing the overall user experience. Furthermore, users can easily determine which issues are most prevalent and have the greatest impact on study quality and performance. By adopting adequate issue management measures to identified problems, risk-based corrective and preventive actions can be undertaken. This approach is fully in line with recent ICH E6 R2 guidance that states: 'The sponsor should develop a systematic, prioritized, riskbased approach to monitoring clinical trials. [...] Review, that may include statistical analyses, of accumulating data from centralized monitoring can be used to: (a) identify missing data, inconsistent data, data outliers, unexpected lack of variability and protocol deviations; (b) examine data trends such as the range, consistency and variability of data within and across sites; (c) evaluate for systematic or significant errors in data collection and reporting at a site or across sites, or potential data manipulation or data integrity problems; (d) analyze site characteristics and performance metrics; (e) select sites and/or processes for targeted on-site monitoring' (22).

Akin to personal finance software that helps consumers manage all their accounts from a single site, our solution creates a layer of abstraction that allows seamless connectivity to the source systems, provides comprehensive and up-to-date views of data cleanliness and completeness, enables bulk management of queries, offers traceable database persistence and documentation of every activity, supports multiple modes of communication and enables continuous data cleansing and reconciliation that reduces site and study risk and speeds up database lock.

By utilizing data integrated in ODW, complex, crossvendor rules may be created to define conditions for missing data and simplify reconciliation across multiple data sources. Automatic detection of new and updated records minimizes redundant work and helps users focus on the latest issues. Queries can be managed individually or in bulk in a system- and vendor-agnostic way, either through direct API integration with the EDC or through email for vendors who do not support API endpoints. The entire history of every query and every discrepancy is tracked consistently regardless of source and is readily available for inspection during an audit. A comprehensive, up-to-date view of the status and quality of the data is readily available through the master subject tracker to help assess readiness for database lock. Finally, the calculations can be run on a preset schedule or at any time upon request throughout the life of the trial and alerts and notifications drive immediate action when predefined thresholds on inactivity or backlog are exceeded.

Discussion

Owing to the heavily regulated nature of the pharmaceutical industry and the protracted length of clinical trials, introduction of new clinical systems requires time and careful planning, and legacy tools tend to linger on long after their functionality has been outdated by newer technologies (23). The Discrepancy Manager, Page Tracker and Subject Tracker were released to production in November 2017 and the Query Tracker in July 2018. As of this writing, the tools are being used in 16 active studies (still relatively early in their course) and are being configured for 15 additional ones. While the technology is too early on its adoption curve and quantitative evidence of impact is still lacking, the feedback received from the pilot teams has been very encouraging, suggesting significant improvements in usability, productivity, quality and compliance, as well as in employee engagement and retention in an area where turnover has been historically high, partly due to the lack of modern tools to reduce repetitive, manual work. Performance metrics along all of these dimensions will be provided in a subsequent publication.

It is important to make a distinction between 'data review' and 'medical review'. The tools presented here address operational aspects of clinical data quality, which are distinct from, and complementary to, the review performed by study physicians. Data review allows data managers to identify and correct missing, erroneous and inconsistent data and ensure that the data sets submitted for statistical analysis are complete and error free. Medical review allows physicians to assess patient safety, protocol deviations and other clinical issues that may degrade trial integrity and involves periodic assessment of patient demographics, adverse events, medical history, concomitant medications, laboratory results, clinical endpoints, protocol deviations, disposition events and many other types of patient data collected during the course of study. Our tool for medical review will be presented in a subsequent publication (24).

Conclusion

We have presented an integrated solution for clinical data review that helps data managers identify missing, erroneous and inconsistent data, manage queries and organize their workload in a unified, system-agnostic and efficient way. The system centralizes all data review and cleansing activities under a common UI, enabled by an operational data warehouse that integrates in near real time all relevant clinical data collected during the course of the trial, communicates seamlessly with the source systems and maintains a complete audit trail of every change in the underlying data. Planned enhancements include a new integrated environment that will greatly simplify the development, maintenance and reuse of complex edit checks, programmable deviations, reconciliations and other data validation and quality checks and additional automation to facilitate database lock and the generation of submissionready Standard Data Tabulation Model (SDTM) data sets.

Acknowledgements

We would like to thank Lynette Thomas, Jatinder Hunjan and Michelle Jones for testing earlier prototypes and providing valuable feedback during the development and deployment of the software.

Conflict of interest. The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. No study sponsor has been involved in the design of the system, the collection, analysis and interpretation of the data presented herein, in the writing of the manuscript or the decision to submit the manuscript for publication.

References

- Krishnankutty, B., Bellary, S., Kumar, N.B.R. *et al.* (2012) *Indian J. Pharmacol.*, 44, 168–172.
- Oracle Oracle Health Sciences Data Management Workbench. http://www.oracle.com/us/products/applications/healthsciences/data-management-workbench (5 November 2018, date last accessed).
- Oracle Oracle Life Sciences Data Hub. http://www.oracle.com/ us/products/applications/health-sciences/e-clinical/data-hub (5 November 2018, date last accessed).
- Oracle Oracle Health Sciences InForm. http://www.oracle. com/us/products/applications/health-sciences/e-clinical/inform (5 November 2018, date last accessed).
- Agrafiotis, D.K., Lobanov, V.S., Farnum, A.M. *et al.* (2018) Riskbased monitoring of clinical trials: an integrative approach. *Clin. Ther.*, 40, 1204–1212.
- McCaffrey, R.P., Tevelev, B.M., Yang, E. *et al.* A statistical evaluation of observed risk flux in RBM-managed clinical trials. Submitted.
- Farnum, M.A., Mohanty, L., Ashok, M. *et al.* A dimensional warehouse for integrating operational data from clinical trials. Database, in press.
- Yang, E., Scheff, J.D., Shen, S.C. *et al.* A late binding, distributed, NoSQL warehouse for integrating patient data from clinical trials. Database, in press.
- Manasco,P.K. (2016) Quality remote monitoring: the tools of the game. *Appl Clin Trials*, 25(6), http://www. appliedclinicaltrialsonline.com/quality-remote-monitoringtools-game.
- 10. Ciervo, J., Shen, S.C., Stallcup, K. *et al.* A new risk and issue management system to improve productivity, quality and compliance in clinical trials. JAMIA Open, in press.
- 11. Mediata *Rave Data Capture and Management*. https://www. mdsol.com/en/products/rave (5 November 2018, date last accessed).
- Microsoft SQL Server 2017. Available at: https://www. microsoft.com/en-us/sql-server/sql-server-2017 (5 November 2018, date last accessed).
- Informatica Enterprise Cloud Data Management. https://www. informatica.com (5 November 2018, date last accessed).
- Microsoft ADO.NET Entity Framework. https://docs.microsoft. com/en-us/dotnet/framework/data/adonet/ef (5 November 2018, date last accessed).
- Dapper Dapper ORM. https://dapper-tutorial.net (5 November 2018, date last accessed).
- Microsoft ADO.NET. https://docs.microsoft.com/en-us/dotnet/ framework/data/adonet (5 November 2018, date last accessed).
- Hangfire Hangfire—An Easy Way to Perform Background Processing in .NET and .NET Core Applications. https://www. hangfire.io (5 November 2018, date last accessed).
- Angular Angular—One Framework, Desktop and Mobile. https://angular.io (5 November 2018, date last accessed).

- 19. O'Neill,R.T. and Temple,R. (2012) The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clin. Pharmacol. Ther.*, 91, 550–554.
- Little,R.J., D'Agostino,R., Cohen,M.L. *et al.* (2012) The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.*, 367, 1355–1360.
- 21. European Medicines Agency *Guideline on Missing Data in Confirmatory Clinical Trials.* https://www.ema. europa.eu/documents/scientific-guideline/guideline-missingdata-confirmatory-clinical-trials_en.pdf (5 November 2018, date last accessed).
- 22. International Committee for Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human

Use ICH Harmonized Guideline. Integrated Addendum to ICH E6(R1): Guideline for Good Clinical Practice. E6(R2). Available at: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2_Step_4_2016_1109.pdf. Published June 11, 2015. Accessed April 7, 2018.

- 23. European Medicines Agency Reflection Paper on Risk Based Quality Management in Clinical Trials. https://www.ema. europa.eu/documents/scientific-guideline/reflection-paper-riskbased-quality-management-clinical-trials_en.pdf (5 November 2018, date last accessed).
- 24. Lobanov,V.S., Rovskiy,V., Du,F. *et al.* Improving medical monitoring and safety review through data integration and visual analytics. In preparation.