Perceptive Reach

Integrated Reach Database System

(IRDS)

IRDS Sandbox Specification



**Department of Veterans Affairs**

Version 2.0

September 2015

Revision History

| Date | Version | Description | Author |
| --- | --- | --- | --- |
| 9/14/2015 | 2.0 | Final Review | Monica Mohler/Paul Bradley |
| 9/9/2015 | 1.9 | Peer Review | Jesse Rogers/Matt Robinson/Kaitlin Reskovac/Radina Ivanova |
| 9/8/2015 | 1.8 | Technical Updates | Robert Snelling |
| 8/21/2015 | 1.7 | Updated | Kristopher Hoover |
| 2/7/2015 | 1.6 | Final Review | Monica Mohler/Paul Bradley |
| 2/6/2015 | 1.5 | Peer Review | Radina Ivanova |
| 2/4/15 | 1.4 | Updated per Statement of Work | Shea Kelly |
| 1/12/15 | 1.3 | Final Review | Monica Mohler/Paul Bradley |
| 1/10/15 | 1.2 | Peer Review | Radina Ivanova |
| 1/9/15 | 1.1 | Update per client comments | Shea Kelly |
| 12/12/14 | 1.0 | Final Review | Monica Mohler/Paul Bradley |
| 12/11/14 | .02 | Peer Review | Radina Ivanova |
| 12/10/14 | .01 | Initial Version | Shea Kelly |

**Table of Contents**

[1. About this Document 4](#_Toc430034276)

[2. Introduction 4](#_Toc430034277)

[2.1. Purpose 4](#_Toc430034278)

[2.2. References 4](#_Toc430034279)

[3. System Description 5](#_Toc430034280)

[3.1. System 5](#_Toc430034281)

[3.2. Security 6](#_Toc430034282)

[3.3. Privacy 7](#_Toc430034283)

[3.4. Authority to Operate 7](#_Toc430034284)

[4. Environment 8](#_Toc430034285)

[4.1. Hardware 8](#_Toc430034286)

[4.2. Software 8](#_Toc430034287)

[4.3. Interfaces Overview 9](#_Toc430034288)

[4.3.1. External Interfaces 10](#_Toc430034289)

[4.3.2. Data Sources 12](#_Toc430034290)

[5. Analytics Workflow 14](#_Toc430034291)

[5.1. Model Development / Applications 14](#_Toc430034292)

[5.2. System Data Flow 15](#_Toc430034293)

[5.2.1. Import Process 15](#_Toc430034294)

[5.2.2. Normalization/Standardization 15](#_Toc430034295)

[5.2.3. Database Tables 15](#_Toc430034296)

[5.2.4. Linking/Common Identifiers 15](#_Toc430034297)

[5.2.5. Risk Model Execution (SQL Server) 15](#_Toc430034298)

[6. Appendix A: Acronyms 17](#_Toc430034299)

# About this Document

The Data Analytics Sandbox Specification Document describes the testing environment for the Perceptive Reach (PR) Integrated Reach Database System (IRDS). This document will be prepared as a one-time deliverable 90 days after contract award and will be updated by future sustainment teams as enhancements to the IRDS are made.

# Introduction

This document provides the general purpose of and specifications for the Integrated Reach Database System (IRDS) Sandbox. The Sandbox environment will be used to develop and test the IRDS solution, which includes data analytics functions, such as data importing, aggregation, and analysis from the Corporate Data Warehouse (CDW), Suicide Data Repository (SDR), Medical SAS (MedSAS), and existing Statistical Analysis System (SAS) datasets[[1]](#footnote-2) to produce Veteran suicide completion risk scores.

## Purpose

The purpose of this document is to describe the specifications needed to setup, configure, and house the Sandbox. The document’s intended audience is for IT Systems Administrators.

## References

This document uses the following resources and repositories to support its contents and claims:

* OneVA EA ETA, August 2012

<http://www.ea.oit.va.gov/EAOIT/docs/ETAComplianceFINAL_12_8_28.pdf>

* VA Directive 6221 Accessible Electronic and Information Technology, Directive/Handbook

<http://www.section508.va.gov/docs/Directive_6221.pdf>

* VA Handbook 6500 – Information Security Program  
  <http://vaww1.va.gov/vapubs/viewPublication.asp?Pub_ID=638&FType=2>
* VA Software Document Library (VDL)

<http://www.va.gov/vdl/>

* VA’s Strategic Plan Refresh FY2011–2015

<http://www.va.gov/op3/docs/strategicplanning/strategic_addendum_fy2011-2015.pdf>

* Perceptive Reach Wiki Page

<http://vacloud.us/groups/558/>

* Perceptive Reach GitHub Repository

<https://github.com/VHAINNOVATIONS/PerceptiveReach>

* Perceptive Reach Jira Repository

<https://opensourceehr.atlassian.net/secure/Dashboard.jspa?selectPageId=10600>

# System Description

## System

The Sandbox is a development environment which will house and enable the analysis of multiple integrated datasets using data analytics, modeling techniques, and visualizations to identify individual veterans and populations with higher risk of suicide completion and provide proactive and secure results notifications to Veteran support services.

As shown in **Figure 1**, the Sandbox will be a development environment with the following components:

* **Reach Database.** A Structured Query Language (SQL) database used to aggregate new data sources and relevant SDR data.
* **Data Analytics Platform and Dashboard.** An integrated collection of analytics and visualization tools, including a surveillance dashboard aimed at identifying at-risk individuals and populations
* **Direct Messaging.** A method to construct and transmit a secure message to authorized outreach and intervention service providers.

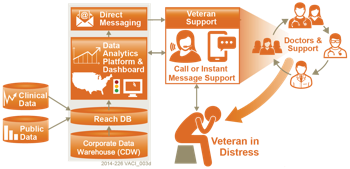
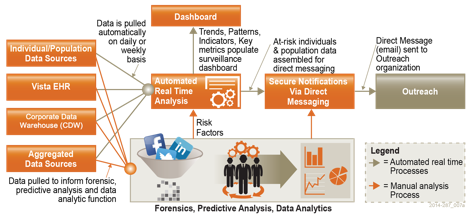


Figure 1: Conceptual Program Design

The underpinning technology and data analytics platform will provide methods by which at-risk populations and individuals can be identified. Specifically, we propose a programmable and configurable solution that can be tailored and enhanced over time as more data sources become available and as clinical research identifies new risk factors. As depicted in the bottom-center of  **Figure 2** a significant component in this effort is the identification of Veteran-specific risk factors, a precursor to the design of an automated reporting model.



**Figure 2: Conceptual System Design**

This effort includes data analysis and predictive modelling, done in collaboration with VA stakeholders and clinical Subject Matter Experts (SME), while at the same time leveraging research data from sources such as Department of Defense (DoD) (e.g. Army Study to Assess Risk and Resilience in Service members [Army STARRS] and the Military Suicide Research Consortium) and VA (e.g. Center of Excellence for Suicide Prevention). Harnessing the automated reporting model, information will be presented in a customizable national surveillance dashboard and transmitted to authorize officials and Veteran support services organizations via Direct Messaging.

## Security

The IRDS solution, including the functions within the Sandbox, will comply with specific security mechanisms per the following regulations and directives:

* Health Insurance Portability and Accountability Act (HIPAA) of 1996
* U.S.C. § 3541, “Federal Information Security Management Act (FISMA) of 2002”
* U.S.C. § 552a, as amended, “The Privacy Act of 1974”
* VA Directive 6500, “Managing Information Security Risk: VA Information Security Program,” September 20, 2012
* VA Handbook 6500, “Risk Management Framework for VA Information Systems – Tier 3: VA Information Security Program,” September 20, 2012
* VA Handbook 6500.1, “Electronic Media Sanitization,” March 22, 2010
* VA Handbook 6500.2, “Management of Data Breaches Involving Sensitive Personal Information (SPI)”, January 6, 2012
* VA Handbook 6500.3, “Assessment, Authorization, And Continuous Monitoring Of VA Information Systems,” February 3, 2014
* VA Handbook, 6500.5, “Incorporating Security and Privacy in System Development Lifecycle” March 22, 2010
* VA Handbook 6500.6, “Contract Security,” March 12, 2010
* VA Directive 6508, VA Privacy Impact Assessment, October 3, 2008
* VA Directive 6300, Records and Information Management, February 26, 2009
* VA Handbook, 6300.1, Records Management Procedures, March 24, 2010
* OMB Memorandum M-07-16, Safeguarding Against and Responding to the Breach of Personally Identifiable Information (PII), May 22, 2007
* NIST SP 800-63-2, Electronic Authentication Guideline, August 2013

## Privacy

The IRDS solution will take a defense in depth approach to protecting PII/Protected Health Information (PHI) data to include the following protection mechanisms:

* IRDS is in compliance with privacy controls per NIST 800-53
* Only accessible from The Department of Veterans Affairs (VA) network
* Application Program Interface (API) protected by a policy enforcement/policy decision point
* VA hosts in the Austin Information Technology Center (AITC) are protected by VA boundary protections at the hosting facility and only administrators have access to the machines
* Data encryption in transit using SQL Server Always Encrypt on any network traffic beyond the local enclave

IRDS underwent a Privacy Threshold Analysis (PTA) and Privacy Impact Assessment (PIA). The official copies of these documents are stored in the VA’s Governance Risk and Compliance (GRC) tool, RiskVision.

## Authority to Operate

The PwC team collaborated with VA stakeholders and SMEs to obtain an Authority to Operate (ATO) in accordance with the "Accreditation Requirements Guide: Standard Operating Procedures" document from VA's Office of Information Security dated June 2014. The ATO was granted on 9/8/2015.

ATO documentation included the following:

* Security Management Plan
* System Security Plan
* Privacy Impact Assessment
* Configuration Management Plan
* Disaster Recovery Plan
* Risk Assessment and Security Certification Checklist
* Proof of VA Handbook 6500 Requirements Fulfillment (as needed)
* Fortify Scan Results and Resolutions (as needed)

# Environment

## Hardware

For testing the SQL implementation of the risk model, the following hardware is required.

Table : Hardware Requirements

| **Hardware Component** | **Requirement** |
| --- | --- |
| Processor | Intel Xeon E5-2600 Family (2670 or 2690), 2.6 Gigahertz (GHZ) or better |
| Random Access Memory (RAM)/Memory | Minimum of 32 Gigabyte (GB) |
| Storage/Hard Disk Drive (HDD) | Minimum of 500GB |

For analyzing and refitting the risk model, the VA Informatics and Computing Infrastructure (VINCI) Workspace is used.

## Software

For testing the SQL implementation of the risk model, the following software is required.

Table : Software Requirements

| **Software Component** | **Requirement** | **Description** |
| --- | --- | --- |
| Operating System (OS) | Microsoft Windows 64 bit OS | Operating system with 64 bit processing power |
| Database | Microsoft SQL Server 2012 | SQL Server is a relational database management system |
| Database | SSIS | SSIS provides data migration task capabilities for SQL Server |
| Analytics Tool | R version 3.1.2 | R is an open source software package for statistical analysis, data mining, and visualization |
| Analytics Tool | RStudio version 0.98.1091 | RStudio is an open source, user interface for the R software package |
| Javascript | AngularJS version 1.3.6 | AngularJS is an open source web application framework |
| Other | HTML5 | HTML5 is an open source web technology markup language |
| Other | CSS3 | CSS3 is a style sheet language used for HTML5 |
| Other | JQuery v 2.1.1 | JQuery is an open source cross-platform for writing JavaScript and HTML |
| Testing | Cucumber JVM 1.2.0 | Cucumber JVM is an open source computational programming testing tool |
| Testing | Jasmine 2.0 | Jasmine is an open source testing framework for JavaScript |

For analyzing and refitting the risk model, R Studio Version 0.98.1091 and Statistical Computing (R) Version 3.1.2 within the VINCI Workspace is used.

## Interfaces Overview

The primary ingress interfaces will use SQL Server Integration Services (SSIS) to retrieve data from the designated data sources such as the SDR. The SSIS package will be configured to pull data from the designated data source at a specified interval (daily/weekly/monthly/annually).

The primary egress interface will be the Direct Messaging interface used to notify VA Suicide Prevention Coordinators (SPC) and other VA or non-VA Outreach and Intervention coordinators and clinicians. The IRDS will leverage the Virtual Lifetime Electronic Record (VLER) Direct messaging service utilizing a RESTful interface to transmit message content. Another interface utilized by VA Staff and SPCs will be the IRDS Surveillance Dashboard which will provide browser based visualizations of critical data to identify Veterans at high risk for suicide.



Figure 3: IRDS Interface Overview

### External Interfaces

The external interfaces for the IRDS solution include the following:

1. IRDS Surveillance Dashboard – The dashboard will provide a browser based user interface featuring representations of the analyzed consolidated data sources. These are processed and organized into visualizations, which will assist SPCs and other VA personnel in their duties.
   * This dashboard will reflect data updated on a daily basis and may include charts, tables, maps, animations, other graphics and visual technology.
   * The dashboard will have configurable features based on user roles as defined by the stakeholders. This will provide for distinct views to meet each role’s business needs.



**Figure 3: Surveillance Dashboard**

1. VLER Direct Messaging - The Direct Messaging component will integrate into the existing VLER Direct Messaging architecture. The Direct Messaging will create unique messages based on new and existing data annotated in the IRDS.
   * To send messages, the Perceptive Reach Application will utilize the VLER Direct Messaging API using authentication and conforming to a valid web service call based on the Direct Secure Messaging Interface Control Document (ICD).
   * The Direct Messaging will utilize public and private keys in creating the authorization header on each web service request to the VLER Direct Message (DM) API.
   * The Perceptive Reach Application has registered to use Direct Secure Messaging and the authentication credentials have been provided. The Perceptive Reach application has the ability to send a request to use Direct as a Service web services.
   * The Perceptive Reach Application will authenticate with the VLER DM API using hash-based message authentication code (HMAC).
   * Recipient of Direct Messages
     1. Individuals can receive Direct Messages using the Direct Secure Messaging Webmail Portal, which requires a Personal Identity Verification (PIV) card for authentication using certificates to login to the system. Users will have the default user role and will be able to send and receive messages.
     2. An external organization can request access to VLER Direct as a Service (DaaS), which allows edge applications to receive and send Direct messages. The information is encrypted and securely sent to the External Partner, as a Direct Message using the API.
2. SDR SSIS Interface - SSIS will be the primary tool for importing data from SDR into the Reach database.
3. CDW SSIS Interface – CDW has given the IRDS development team a SQL Server project database (VACI\_IRDS), located on a CDW server. The database will be used:
4. To import Veterans Health Information Systems and Technology Architecture (VistA) data from CDW work tables into new tables in the VACI\_IRDS, for the purpose of validating the VA Risk Model and then enhancing it.
5. As work space for the automated import process, built in SSIS, to filter and transform the CDW data before pulling it over the network onto the IRDS server and into the Reach database.
6. CDW SAS Data – The IRDS Analytics team will have access to SAS data files located on the CDW server as part of their validation of the VA Risk model. These files will be accessed through R software located on an authorized VA machine. For the initial run of the risk model in the IRDS production environment, the VA risk model output files will be used to access the Suicide Completion indicator field as part of the calculations. The VA has uploaded these files, in SAS format to the SAS Grid. The files will be imported into SQL tables in the VACI\_IRDS database located on the CDW server VHACDWA01.
7. (FUTURE) Enterprise Data Warehouse (EDW) SSIS Interface – An SSIS packages will be developed to create a SQL connection to the Veteran Benefit Administration (VBA) EDW Oracle database. The details for that interface are currently being worked out between the IRDS management team and their VBA points of contact.
8. (FUTURE) VistA Remote Procedure Call (RPC) Interface (tentative) - The VA uses the VistA system, for managing Veterans health data. Data may be imported into the IRDS system directly from VistA using RPC calls when immediate access to recently updated clinical records is retired. As VistA is modernized additional interface protocols may be used in place of RPCs. This is to be determined.
9. Active Directory – The VA uses Active Directory, to manage user accounts within the VA Network. The IRDS application will make HTTP requests to the Active Directory instance located within the VA network for user authentication and account information.

### Data Sources

The data sources to be imported into the Reach database of the IRDS have been identified as:

1. VA Suicide Data Repository (SDR) - The SDR contains VA and DoD suicide and mortality data. The data store is Microsoft SQL Server. Periodically data is imported into the system from

* Mortality search results from the National Death Index (NDI)
* State Death Certificate Data (SDCD)
* Veterans Crisis Line (VCL)
* Suicide Prevention Applications Network (SPAN)
* VA Beneficiary Identification Records Locator Subsystem (BIRLS) Death File

1. Veterans Health Information Systems and Technology Architecture (VistA) - It is assumed that some Veterans Health Administration (VHA) data will be retrieved directly from the VA VistA system. VistA uses an M data platform. VistA data will be imported into the Reach database using custom and/or currently existing RPC calls.
2. Corporate Data Warehouse (CDW) - CDW warehouses VHA data (mostly VistA data) in SQL Server format. VistA data stored in CDW will be imported into the Reach database. The automated data import solution (in the form of SSIS packages) will be deployed on an Extract, Transform, Load (ETL) server provided by CDW. The SSIS packages will:
3. Make SQL connections to both the CDW server that contains the Patient data, and the IRDS production server
4. Pull data from the CDW Work database on the CDW server into the VACI\_IRDS database on that server
5. Perform the appropriate data transformations in the VACI\_IRDS database
6. Push the post transformed data from the VACI\_IRDS database into the Reach database on the IRDS server



**Figure 5: CDW/IRDS databases**

1. (FUTURE) Enterprise Data Warehouse (EDW) – The Veterans Benefit Administration (VBA) store Veteran data in SQL format in an Oracle database. Some data elements from VBA (financial status, marital status) may be pulled into the Reach database and linked to VHA records for those individuals by matching on elements such as Social Security Number (SSN).
2. HealthIndicators.gov - The web site is a source for statistical information. The IRDS system imports data pertaining to suicide death rates for specified demographics via a call to a web service. The data is returned in XML format. That data is stripped from the XML tags and stored in a table in the Reach database, which is queried by the dashboard against Patient data that it is integrated with.
3. (FUTURE) Other Non-VA data sources - The VA has expressed an interest in importing and analyzing data that is external to VA data sources. Some of the sources being considered are LexisNexis and Public Access to Court Electronic Records (PACER). The requirements team is currently reviewing possibilities with the VA.
4. MedSAS- VistA data is collected in SAS data sets in the MedSAS system. The risk model applied to the IRDS system was developed by the VA against the MedSAS data set. When the IRDS analytics team ran the model against the same data in the CDW, the model was no longer applicable. The IRDS analytics team developed a hybrid risk model that uses VistA data from CDW, but uses MedSAS data for demographics of patients (DOB, Race, Marital Status, Urban/Rural, Service Connected, Region).The Reach database will contain a table that contains patient data with the date elements listed above and their Scrambled SSN (ScrSSN) as a unique identifier. When patients are imported from CDW to the Reach database in the IRDS system, that data is linked to the MedSAS table via ScrSSN. If demographic data for a patient is available in MedSAS the MedSAS values overwrite the CDW values for the patient in the Patient and PatientRiskFactors tables.

# Analytics Workflow

## Model Development / Applications

The data is manipulated using SQL queries to develop the modeling dataset which will be exported from SQL and imported into R Studio on the VINCI Workspace. The data imported into R Studio will be used to refit the risk model. The coefficients fitted during the model development will be stored as a SQL table in the CDW.

On an ongoing basis and using SQL, these stored coefficients will be applied to updated data to produce suicide completion risk scores for VHA users. The risk scores will be translated to risk classifications which will be used in connection with the IRDS dashboard and direct messaging purposes.

**Figure 6** provides a visual overview of the data workflow within the VINCI Workspace and Sandbox environment.



Figure 6: Example of Data Analytic Model Workflow

## System Data Flow

### Import Process

The import process for analytic model development will take place in two phases:

1. SQL Server will query data from CDW and the development and validation containing MedSAS data provided from the Serious Mental Illness Treatment Resource and Evaluation Center (SMITREC) team, and save the data tables as flat files to be accessed by R.
2. R will then import the data from flat files for model development.

### Normalization/Standardization

For current suicide completion risk model recreation and development of its modifications, all data normalization and standardization will be performed in SQL.

### Database Tables

The database tables will be queried from the CDW; these tables will inherit the structure from their native source.

### Linking/Common Identifiers

The field used in the CDW data to identify a unique VHA user will be ScrSSN.

### Risk Model Execution (SQL Server)

The suicide completion risk model supports the Perceptive Reach program by identifying Veterans at high risk for suicide completion. The application of the suicide completion risk model coefficients to produce individual Veteran suicide completion risk scores will be calculated within the SQL Server. More specifically, the coefficients from the predictive suicide completion risk model will be saved to a table within the SQL Server and applied to the ongoing Veteran data in SQL. These risk scores will be imported into the dashboard and used to generate direct messages.

# Appendix A: Acronyms

Table : Acronyms and Abbreviations

| Acronym | Term |
| --- | --- |
| API | Application Program Interface |
| A&A | Assessment and Authorization |
| AITC | Austin Information Technology Center |
| ATO | Authority to Operate |
| BIRLS | Beneficiary Identification Records Locator Subsystem |
| CDW | Corporate Data Warehouse |
| CIO | VA Chief Information Officer |
| CISO | OIS Chief Information Security Officer |
| DaaS | Direct as a Service |
| DM | Direct Message |
| EDW | Enterprise Data Warehouse |
| ETL | Extract, Transform, Load |
| FISMA | Federal Information Security Management Act |
| GB | Gigabyte |
| GHZ | Gigahertz |
| GRC | The Governance, Risk and Compliance |
| HDD | Hard Disk Drive |
| HIPAA | Health Insurance Portability and Accountability Act |
| HMAC | Hash-Based Message Authentication Code |
| ICD | Interface Control Document |
| IRDS | Integrated Reach Database System |
| ISO | Information Security Officer |
| MedSAS | Medical SAS |
| NDI | National Death Index |
| OCS | Office of Cyber Security |
| OIS | Office of Information Security |
| OS | Operating System |
| PACER | Public Access to Court Electronic Records |
| PIA | Privacy Impact Assessment |
| PHI | Protected Health Information |
| PII | Personally Identifiable Information |
| PIV | Personal Identity Verification |
| PR | Perceptive Reach |
| PTA | Privacy Threshold Analysis |
| R | Statistical Computing |
| RAM | Random Access Memory |
| RPC | Remote Procedure Call |
| SAS | Statistical Analysis System |
| ScrSSN | Scrambled Social Security Number |
| SDCD | State Death Certificate Data |
| SDR | Suicide Data Repository |
| SME | Subject Matter Expert |
| SMITREC | Serious Mental Illness Treatment Resource and Evaluation Center |
| SPAN | Suicide Prevention Applications Network |
| SPC | Suicide Prevention Coordinator |
| SPI | Sensitive Personal Information |
| SQL | Structured Query Language |
| SSIS | SQL Server Integration Services |
| SSN | Social Security Number |
| TRM | Technical Reference Model |
| VA | Department of Veterans Affairs |
| VBA | Veteran Benefit Administration |
| VCL | Veterans Crisis Line |
| VHA | Veteran Health Administration |
| VINCI | VA Informatics and Computing Infrastructure |
| VistA | Veterans Health Information Systems and Technology Architecture |
| VLER | Virtual Lifetime Electronic Record |

1. The existing SAS datasets shall be the datasets used to replicate the initial suicide completion risk model. [↑](#footnote-ref-2)