CHAPTER

1

WELCOME TO CAINTEGRATOR2 ONLINE HELP

The following topics introduce you to caIntegrator2 and describe concepts you need to know to use caIntegrator2:

- *Getting Started with caIntegrator2*
- *Creating a New Study*
- *Using the caIntegrator 2 Workspace*
- *Searching a caIntegrator2 Study*
- *Viewing Search Results*
- *Analyzing Studies*
- *Glossary*

See also *Using Online Help*.

CHAPTER

1

GETTING STARTED WITH caINTEGRATOR2

This chaptersection introduces general caIntegrator2 procedures and how to obtain help to use caIntegrator2.

Topics in this chaptersection include:

- *About caIntegrator2* on this page
- *Using Online Help* on page 4
- *Logging In* on page 5
- *Logging In* on page 5
- *Using the caIntegrator2 Workspace* on page 5
- *Logging Out* on page 6
- *Application Support* on page 7

About caIntegrator2

caIntegrator2 is a web based tool that allows you to integrate clinical, genomic, and imaging data. To create studies integrating multiple study data in caIntegrator2, you can upload clinical data of various types in a predefined flat file format or use data in existing caBIG® applications such as caArray and NCIA . You can update an existing caIntegrator2 study to add new attributes or to add or modify data. In caIntegrator2, you can initiate searches or perform analyses on selected studies.

Using Online Help

Logging In

Using the caIntegrator2 Workspace

Logging Out

Application Support

Using Online Help

The online help explains how to use all of the features of caIntegrator2.

To access online help, use any of the following methods:

- Click the **help** menu option on the top menu bar to open the complete online help project.

  Note: You can access online help without being logged in.

- Click the help icon at the top of each page to open a context-sensitive topic. Context-sensitive help displays information that corresponds to the page from which help was opened.

Once you are in online help, several options help you locate topics of interest.

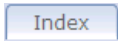
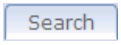
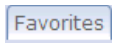
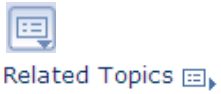
Item	***Description***
	Opens the complete online help table of contents and locates your current topic in the table. When you display the table of contents the following buttons are available:
	Locates and highlights your current topic in the table.
	Select a topic from the complete online help table of contents.
	Select a topic from the online help index.
	Perform word searches of Help by entering query text in the search text box.
	Create a list of your frequently-accessed topics.
	Opens other closely related topics.
	Prints the current topic.
Topic Name > Topic Name	The breadcrumb trail shows the relative location of the current help topic relative to neighboring topics. Click a breadcrumb link to display that help topic.
Back Forward	Navigates through previously viewed topics.

Table 1.1 Online help tips

Logging In

Logging In

Logging Out

Application Support

Logging In

1. On the login page, enter your **username** and **password**.
2. Click the **Login** button. If your login is successful, the Welcome to Browse/ Study page appears *(Figure 1.1)*.).

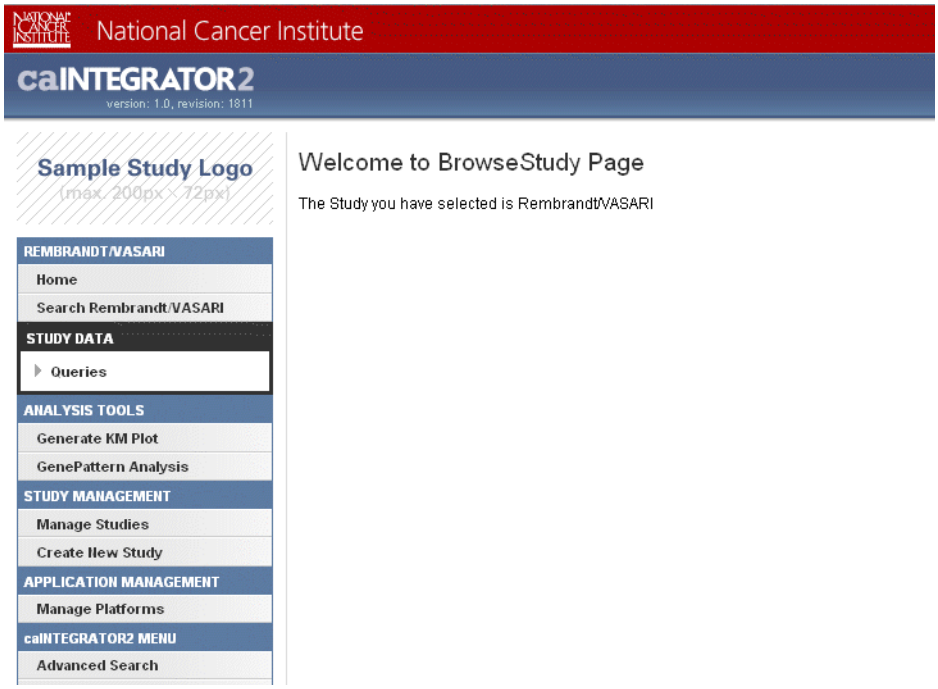


Figure 1.1 The caIntegrator2 workspace

Logging In

Using the caIntegrator2 Workspace

Using the caIntegrator2 Workspace

The caIntegrator2 workspace enables quick access to all caIntegrator2 functions and information. To access caIntegrator2 functions, use the options listed on the left sidebar of the workspace.

See also:

- *caIntegrator2 Functions* on page 5
- *Logging Out* on page 6

caIntegrator2 Functions

When you log into caIntegrator2, the workspace opens to one study, whose name is listed at the top of the left sidebar.

You can access all caIntegrator2 functions from the sidebar on the left side of the page.

*Table 1.2*The following table describes each caIntegrator2 option in the workspace (*Figure 1.1*).

Sidebar Option	***Function***
[Study Name]	When you log in, this study displays in the left sidebar by default. Any study that you select in the My Studies drop-down list in the upper right of the page replaces this default selection.
Home	
Search [study name]	Click this option to open the Search [study name] page from which you can launch queries into your selected study. For more information, see.....
Study Data	Click to open the list of previous queries you saved. Click any item in the list to open the saved query, which displays on the Criteria, Columns and Sorting tabs. From those tabs, you can modify criteria and/or launch the query again. For more information, see *Chapter 3 Searching a caIntegrator2 Study*.
Analysis Tools	Click either of the listed options listed to open a page where you can launch an analysis of the data in the selected study. • Generate KM Plot Xref • GenePattern Analysis Xref
Study Management	Click either of the listed options to manage the selected study through editing or deleting it or by creating a new study. • Manage Studies Xref • Create a New Study Xref

Table 1.2 caIntegrator2 tabs

In the **My Studies** drop-down list in the upper right of the page, select the study you want to use for your current session. (The list includes all studies to which you are subscribed.) As you do so, the left sidebar contents change to reflect options relevant to your study selection.

- the logo for the selected study (if it exists)
- the name for the selected study
- the list of saved queries for that study

Roles??

Logging Out

To log out of caIntegrator2, click the **logout** link (*Figure 1.2*) in the upper right-hand corner.

About caIntegrator2

Using Online Help

Using the caIntegrator2 Workspace

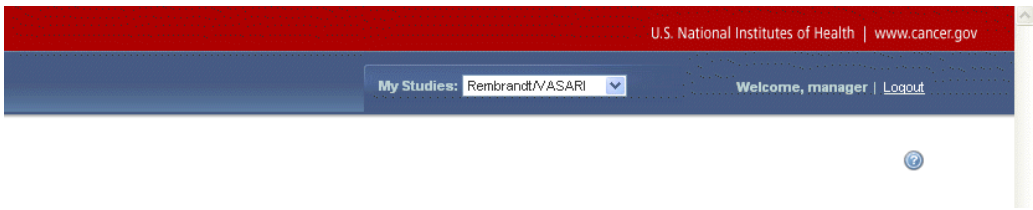


Figure 1.2 Logout link

Application Support

For any general information about the application, application support or to report a bug, contact NCICB Application Support.

Email: ncicb@pop.nci.nih.gov	When submitting support requests via email, please include: • Your contact information, including your telephone number. • The name of the application/tool you are using • The URL if it is a Web-based application • A description of the problem and steps to recreate it. • The text of any error messages you have received
Application Support URL	http://ncicb.nci.nih.gov/NCICB/support
Telephone: 301-451-4384 Toll free: 888-478-4423	Telephone support is available: Monday to Friday, 8 am – 8 pm Eastern Time, excluding government holidays.

About caIntegrator2

Using Online Help

Using the caIntegrator2 Workspace

CHAPTER

2

CREATING A NEW STUDY

This chaptersection describes the processes for creating and managing studies in caInegrator2.

Topics in this chaptersection include:

- *Creating a Study – Overview* on this page
- *Managing a Study* on page 24

Creating a Study – Overview

You can create a caIntegrator2 study by importing clinical study data, genomics data and imaging data, using a combination of spreadsheet/files and existing caGrid applications as source data. Each instance of caIntegrator 2 can support multiple studies. As the manager creating a study, it is important that you understand the study well and that the data you wish to aggregate has been submitted to the applications whose data can be integrated in caIntegrator2.

- The clinical data should be available in CSV files, with a unique patient identifier in one column, one patient per row. Other relevant data can be supplied in other columns, to be identified as annotations in the file from within caIntegrator2. You, as the study creator, must have access to the clinical data file, as the file does not come from a caBIG® repository.
- In order to use caIntegrator 2 to integrate array data, the data should be imported into caArray, either locally or the CBIIT installation, using that system's data file import functionality. You must also have a mapping file in CSV format. The mapping file in CSV format indicates connections between array files and the clinical subjects in the clinical data files. A mapping file consists of two columns: one with the patient ID, and one with the sample ID.
- Imaging data should have been submitted to NCIA grid node as public data, either locally or as part of the CBIIT installation. Image annotations, which includes information about images provided by radiologists or other researchers

can include such information as tumor size, tumor location, etc. must be in CSV format, with unique image series IDs in one column and annotation IDs in the second column. You must also have an image mapping file in CSV format. The mapping file in CSV format indicates connections between images in NBIA and clinical subjects in the clinical data files. A mapping file consists of two columns: one with the patient ID, and one with the NCIA image series ID in the other column.

As you create the study, you define its structure in the process, identifying the data sources and mapping the data between different source data. caIntegrator2 has KM-plot as well as genePattern integration for performing analyses.

Configuring and Deploying Study

Note: Only a user with a Manager role can create a study.

When you create a study, you must specify different data-types (clinical, array, image, tissue etc), data sources (caGrid applications – caArray and NCIA) and map the data, (patient to sample, image series, etc.).

To create a new study, follow these steps:

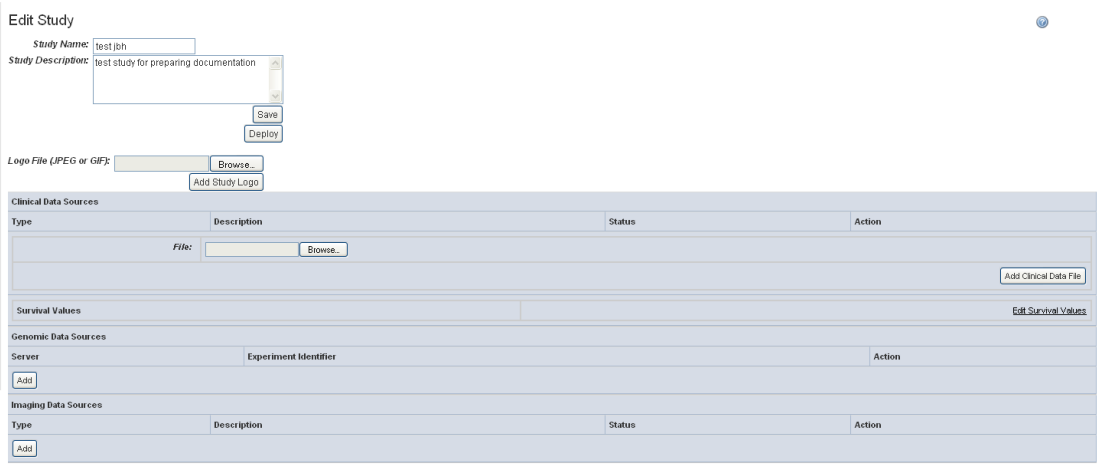
1. In the Study Management section of the left sidebar, click **Create New Study**.
2. In the Create New Study dialog box that opens, provide a name and description for the study you are creating.
3. Click **Save**.

This opens an Edit Study page where you can add identify data files for your study. See *Creating/Editing a Study*.

Creating/Editing a Study

The Edit Study page displays the Name and Description that you entered for a new study, or for an existing study that you are editing.

To continue creating a study or to modify a study, on the Edit Study page complete

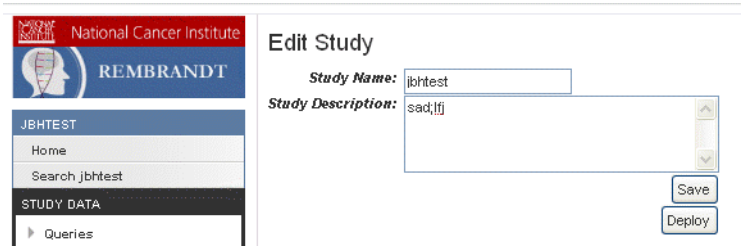


these steps:

1. Change the name and or description, if you so choose. Click **Save**.

   Note: You can save the study at any point in the process of creating it. You can resume the definition and deployment process later.

2. If you choose to add a logo for the study, click the **Browse** button corresponding to **Logo File**. Navigate for the file, then click **Add Study Logo**. Once you save the study (or its edit), the logo displays above the left sidebar.



To continue, you can add clinical data sources, genomic data or imaging data.

Adding Clinical Data

The Edit Study page opens after you save a new study or click to edit an existing study.

Note: To edit information for an existing study, follow the same basic directions in this topicsection. Instead of entering new information, you will modify existing information.

To add or edit clinical metadata in this page, follow these steps:

1. To add or edit clinical data, on the Edit Study page click the **Browse** button. Navigate to locate the file. Files must be in CSV file format.

   In the Clinical Data Sources section, if a file has already been selected, its information displays in the varying fields.

2. Click **Add Clinical Data Source**. This opens the Define Fields for Clinical Data page.

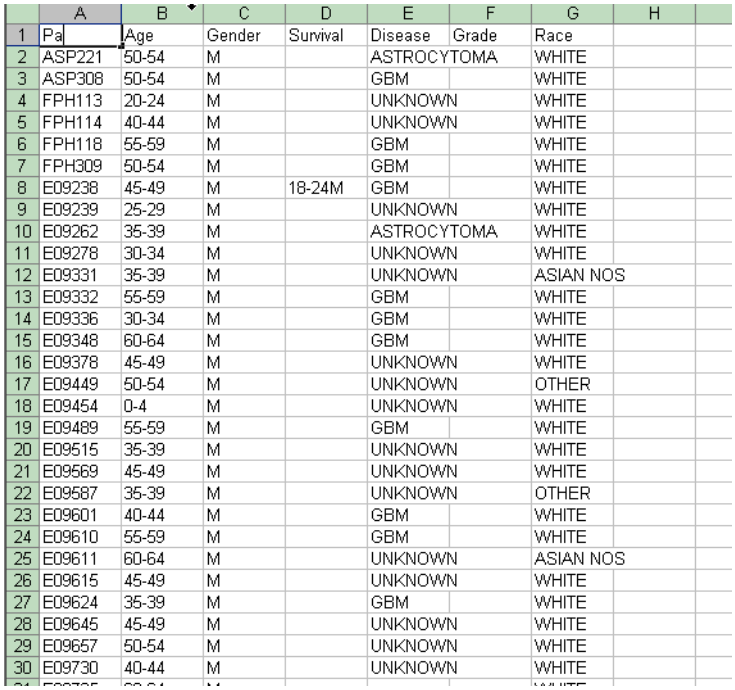
Define Fields for Clinical Data

Study Name: Rembrandt with NCRI Study

Field Definition	Field Header from File
Identifier Change Assignment	Subject
Age Change Assignment	Age
GENDER Change Assignment	Gender
Survival Change Assignment	Survival
Disease Change Assignment	Disease
Grade Change Assignment	Grade
RACE Change Assignment	Race

Clinical data is always added in spreadsheet format (CSV file). The page displays column headers and data values in the file you have designated. You

must map each column name to an existing column name in the caIntegrator2 database or, if it doesn't yet exist, create a custom column name.



	A	B	C	D	E	F	G	H
1	Pa	Age	Gender	Survival	Disease	Grade	Race	
2	ASP221	50-54	M		ASTROCYTOMA		WHITE	
3	ASP308	50-54	M		GBM		WHITE	
4	FPH113	20-24	M		UNKNOWN		WHITE	
5	FPH114	40-44	M		UNKNOWN		WHITE	
6	FPH118	55-59	M		GBM		WHITE	
7	FPH309	50-54	M		GBM		WHITE	
8	E09238	45-49	M	18-24M	GBM		WHITE	
9	E09239	25-29	M		UNKNOWN		WHITE	
10	E09262	35-39	M		ASTROCYTOMA		WHITE	
11	E09278	30-34	M		UNKNOWN		WHITE	
12	E09331	35-39	M		UNKNOWN		ASIAN NOS	
13	E09332	55-59	M		GBM		WHITE	
14	E09336	30-34	M		GBM		WHITE	
15	E09348	60-64	M		GBM		WHITE	
16	E09378	45-49	M		UNKNOWN		WHITE	
17	E09449	50-54	M		UNKNOWN		OTHER	
18	E09454	0-4	M		UNKNOWN		WHITE	
19	E09489	55-59	M		GBM		WHITE	
20	E09515	35-39	M		UNKNOWN		WHITE	
21	E09569	45-49	M		UNKNOWN		WHITE	
22	E09587	35-39	M		UNKNOWN		OTHER	
23	E09601	40-44	M		GBM		WHITE	
24	E09610	55-59	M		GBM		WHITE	
25	E09611	60-64	M		UNKNOWN		ASIAN NOS	
26	E09615	45-49	M		UNKNOWN		WHITE	
27	E09624	35-39	M		GBM		WHITE	
28	E09645	45-49	M		UNKNOWN		WHITE	
29	E09657	50-54	M		UNKNOWN		WHITE	
30	E09730	40-44	M		UNKNOWN		WHITE	

The Field Header from File column on the Define Fields... page displays column headers taken from the source *.csv file.

The MOST important steps in creating a new study in caIntegrator2:

- You MUST designate one column in the file as a unique "identifier" column type.
- You MUST review and define column annotation definitions for each column header in the file.

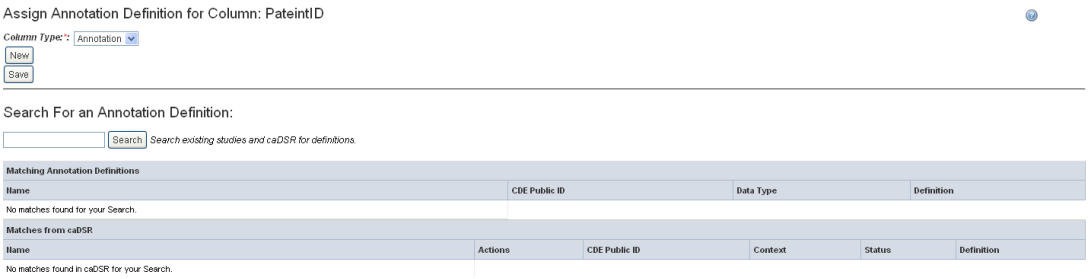
If caIntegrator "recognizes" the same column header in other files already in the system, a term such as "age" or "survival" which is the current definition appears in the **Field Definition** column above the blue **Change Assignment** link. If the area above the blue **Change Assignment** link is blank, no correlating term exists in the database; you must specify the field type, and then the term will populate the space.

3. To indicate the unique identifier of choice, on the row showing the column header (PatientID in the figure, but other examples are subject identifier, sample identifier, etc), click **Change Assignment** in the **Field Header from File** column.

Assigning An Identifier or Annotation

When you click **Change Assignment** on the Define Fields... page, the Assign Annotation Definition for Column dialog box opens. On this page you can change the column type and the field definition for the specific data field you selected.

Note: When you change an assignment, you must make sure the data types match--numeric, etc.



1. For the column (PatientID) that you choose to be the one and only Identifier column, in the **Column Type** drop-down list, select **Identifier**.

   When the definition becomes an identifier, the rest of the page disappears, showing only the Identifier definition.

   Note: If you select **Annotation** after you have already selected **Identifier**, the rest of the page reappears.

2. Click **Save** to save the identifier. This returns you to the Define Fields for Clinical Data page.

3. After you have defined which field is the Identifier, you must ensure that ALL other fields also have a field definition assignment. For those fields without a Field Definition assignment or for those whose Field Definition you want to review, click **Change Assignment**.

4. In the Assign Annotation Definition for Column: [column header] dialog box, select **Annotation** in the drop-down list.

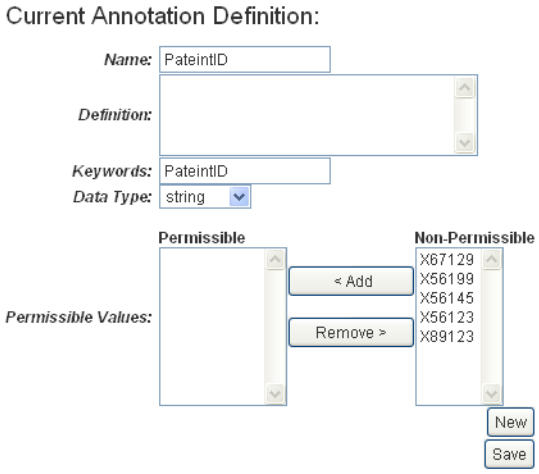
As you select the column type, you can work with column headers in one of four ways in this dialog box.

- You can accept existing default definitions (those that are inherent in the data file you selected).
- You can create your own definitions manually.
- You can search for and use definitions in other caIntegrator2 studies.
- You can search for and use definitions in caDSR.

5. If there is anything you want to change about the annotation definition of the field such as its name, or if you want to view or edit its definition, click the **New** button under the **Column Type** field. The page expands, now including a Current Annotation Definition section.

   Note: If the column header you are working with already has a designated Field Definition, the Current Annotation Definition section of the Assign

Annotation Definition for Column page is already visible when you open this dialog box.



6. You can enter a new name annotation, or any other information about the annotation definition:

Annotation Field	***Field Description***
Description	Enter any information not covered by other fields.
Keywords	Insert keyword(s) that can be used to find the study in a search, separated by commas.
Data Type	Enter a string (default), numeric, or date

Table 2.1

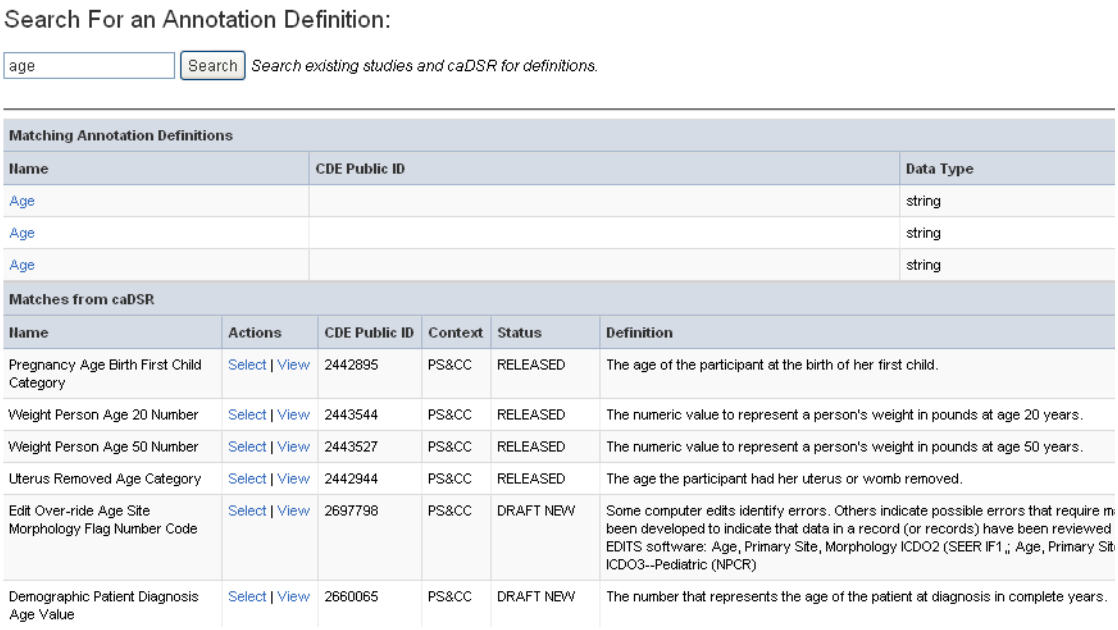
Annotation Field	*Field Description*
Permissible/Non-permissible Values	**Note:** The first time you load a file, before you assign annotation definitions (step 3 on page 12), these panels may be blank. If the column header for the data is already "recognizable" by caIntegrator2, the system makes a "guess" about the data type and assigns the values to the data type in the newly uploaded file. They will display in the Non-permissible values sections initially. Use the **Add** and **Remove** buttons to move the values shown from one list to the other, as appropriate. When you select or change annotation definitions by selecting matching definitions (described in *Searching for Annotation Definitions* on page 15), this may add (or change) the list of non-permissible values in this section. If you leave all values for a field in the Non-permissible panel, then when you do a study search, you can enter free text in the query criteria for this field. If there are items in the Permissible values list, then the values for this annotation are restricted to only those values. When you perform a study search, you will select from a list of these values when querying this field. If there are no items in the permissible values list then the field is considered free to contain any value. To edit a field's permissible values, you must change the annotation definition. You can do this even after a study has been deployed.

Table 2.1

Searching for Annotation Definitions

1. An alternative to creating a new definition is to search for annotation definitions already present in caIntegrator2 studies or in caDSR. Enter search keyword(s)

in the **Search** text box. Click **Search**. After a few moments, the search results display on the page--new screen shot.



2. To view the definitions corresponding to any of the “Matching Annotation Definitions”, which are those currently found in other caIntegrator2 studies, click the [term], such as “age”, hypertext link. The definition then appears in the Current Annotation Definition segment of the page just above. Verify this for new screen shot. You can modify an portion of the definition, as described in step 6 on page 14.

3. The matches from caDSR display some of the details of the search results. To view more details of a match, such as permissible values, click **View**, which opens caDSR to the term. If you click **Select**, the caDSR definition automatically replaces the annotation definition for this field with which you are working.

   Caution: Take care before you add a caDSR definition that it says exactly what you want. caDSR definitions can have minor nuances that require specific and limited applications of their use.

4. Once you have settled on an appropriate field definition for the annotation, click **Save**. This returns you to the Define Field for Clinical Data page.

   Note: If you have not clicked **Select** for alternate definitions in this dialog box, then click **Save** to return to the Define Field...dialog box without making any definition changes.

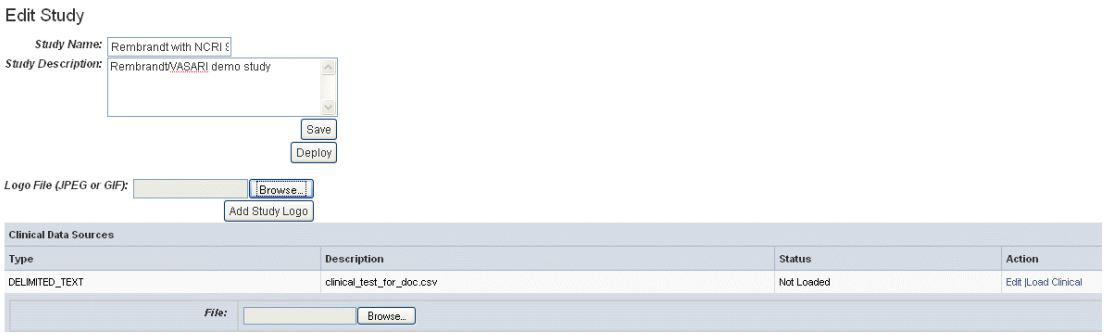
5. From the Define Fileds for Clinical Data screen, be sure and designate the annotations for each field in the file. Click **Save** on each page to save your entries or click **New** to clear the fields and start again. You will not be able to proceed until every Field Definition entry on the Fields for Clinical Data screen has a unique entry.

On the Define Fields for Clinical Data page, the Field Definition column displays the term “Identifier” corresponding to the Column Header you designated as such. The Data From File columns display the column header *values* of the first three rows you designated as “annotations”.

6. Check the boxes how designated?? corresponding to each column header/data set whose data you want to display in your browser.

7. Once you have assigned the identifier and assigned all annotations, click **Save**. This saves the study by name and description, but does not deploy the study, which is described in *Deploying the Study* on page 24.

   Saving the study returns you to the Edit Study page where a “Not Loaded” status now appears for the file whose annotations (column headers) you have defined.



8. Click the **Load Clinical** link in the Actions section to load the data file you configured. At this point, the Status changes to “Loaded”.

   Note: You can add as many files as are necessary for a study. Patients 1-20 in first file, 21-40 in second file, or many patients in first file and annotations in second file, etc. As long as IDs defined correctly, it works.

9. Once you have assigned data types to every column header in the data file and have loaded the clinical data, click **Save and Deploy**. At that point, caIntegrator2 loads data from the file to the caIntegrator2 database.

   Note: You can change assignments even after the study is deployed, using the Edit feature. For more information, see *Editing a Study* on page 18.

   When study is deployed, that status is indicated on the Manage Studies page. For more information, see *Managing a Study* on page 24.

   The Manage Studies page opens when the study is deployed. The **Deployed** status is indicated on the Manage Studies page. For more information, see *Managing a Study* on page 24.

   See also *Deploying the Study* on page 24.

Note: You can repeatedly upload additional or updated subject annotations, samples, image data, array data to the study at later intervals. These later imports do not remove any existing data; they instead insert any new subjects or update annotations for existing subjects.

Mapping Genomic Data to Clinical DataJill, add sentence of introduction...

Uploading Control Samples

Deploying the Study

Creating/Editing a Study

Adding Clinical Data

Editing Survival Values

Adding Genomic Data

Adding Imaging Data

Managing Platforms

Editing a Study

Editing Survival Values

If you plan to analyze your data in caIntegrator2 creating a Kaplan-Meier (KM) Plot, then during the Annotation Definition process described above in *Assigning An Identifier or Annotation*, you must make sure that you have defined at least three fields set to the "date" Data Type. These will be matched to the following three properties during Survival Value Definition.

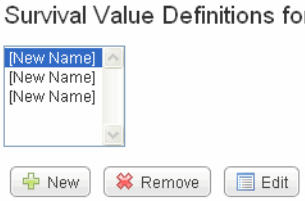
- Survival Start Date:
- Death Date:
- Last Followup Date:

Note: Setting survival values is optional, if you do not plan to use the KM plot analysis feature or even if you do not have this kind of data (survival values) in the file.

For some applications, such as REMBRANDT and I-SPY, survival values are pre-defined in the databases when you load the data is loaded. In caIntegrator2, however, you can review and modify survival value ranges in a data set you are uploading to create a study. To be able to do so, you need to understand what kind of data needs to make up the survival values.

To work with survival values, follow these steps:

1. On the Edit Study page, click **Edit Survival Values**.



2. Click **New** to enter new survival value definitions.

- or -

3. Click **Edit** to edit existing survival value definitions.
4. The dialog box extends, now displaying three drop-down lists that show column headers for date metadata in the spreadsheet you have uploaded.

In the drop-down lists, select the appropriate survival value definitions for each field listed. You might want to refer to the column headers in the data file itself. Dates covered by the definitions are already in the data set. You cannot enter a specific date.

- **Name** – Enter a unique name that adequately describes the survival values you are defining here. *Example*: Survival from Enrollment Date or Survival from Treatment Start. The name you enter displays later when you are selecting survivals to create the KM plot.
- **Survival Start Date** – Select the column header for this data
- **Death Date** – Select the column header for this data
- **Last Followup Date** – Select the column header for this data

Note: Dates corresponding to each field in the drop-down lists are already in the data set columns. You cannot enter a specific date.

Adding Genomic Data

Once you have loaded clinical data and identified patient IDs, you can add either array genomic sample data from caArray, which caIntegrator2 maps by sample IDs to the patient IDs in the clinical data, covered in this section, or you can load imaging files from NCIA, also mapped by IDs to the patient data, covered in *Adding Imaging Data*the following section. Genomic sample data and imaging data are independent of each other, so neither is required before loading the other.

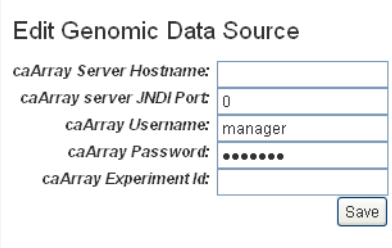
It is essential that you are well acquainted with the data you are working with--the clinical data, and the corresponding array data in caArray.

caIntegrator2 supports a limited number of array platforms. For more information, see *Managing Platforms* on page 25.

To add genomic data to your caIntegrator2 study, follow these steps:

1. On the Edit Study page where you have selected and added the clinical data, click the **Add** button under Genomic Data Sources. You can upload genomic data only from caArray.

   This opens the Edit Genomic Data Source dialog box. Enter the appropriate information in the fields.



   - **caArray Host Name** – Enter the hostname for your local installation or for the CBIIT installation of caArray, array.nci.nih.gov. If you misspell it, you will receive an error message.
   - **caArray JNDI Port** – Enter the appropriate server port. See your administrator for more information. *Example*: For the CBIIT installation of caArray, enter ***8080***.
   - **caArray Username** and **caArray Password** – If the data is private, you must enter your caArray account user name and password; you must have been given permissions in caArray for the experiment. If the data is public, you can leave these fields blank.
   - **caArray Experiment ID** – Enter the caArray Experiment ID which you know corresponds with the clinical data you uploaded. *Example*: Public experiiment “beer-00196” on the CBIIT installation of caArray (array.nci.nih.gov). If you misspell it, you will receive an error message.
   - Click **Save**.

   caIntegrator2 goes to caArray, validates the information you have entered here, finds the experiment and retrieves all the sample IDs in the experiment. Once this finishes, the experiment information displays on the Edit Study page under the Genomic Data Sources section.

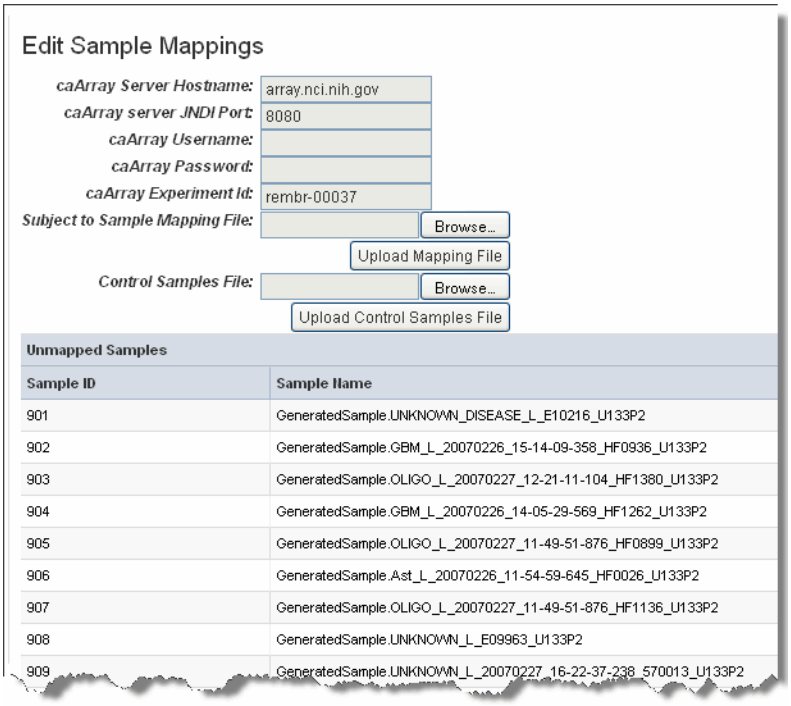
2. If you want to redefine the caArray experiment information, you can edit it. Click the **Edit** link corresponding to the Experiment ID. The Edit Genomic Data Source dialog box reopens, allowing you to edit the information.

Mapping Genomic Data to Clinical Data

Jill, add sentence of introduction...

To map the samples from the caArray experiment to the patients in the clinical data you uploaded, follow these steps:

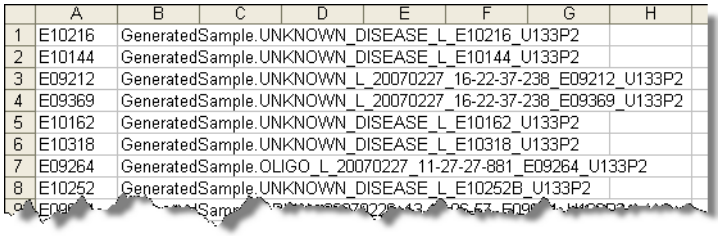
1. On the Edit Study page, click the **Map Samples** link. This opens the Edit Sample Mappings page.



When you first open this page, all of the samples in the caArray experiment you selected are listed as unmapped, because caIntegrator2 does not know how these sample names correlate to the patient data in the clinical file until you upload the mapping file.

2. Browse for the .csv file that identifies the mapping information. Click the **Upload Mapping File** button.

   The mapping file has only two columns (typically without headers)–one that shows the subject ID (designated in caIntegrator2 as the “Identifier) and one that has “Sample name” field from the linked caArray experiment, with one subject per row. This provides caIntegrator2 with the information for mapping patients to caArray samples.



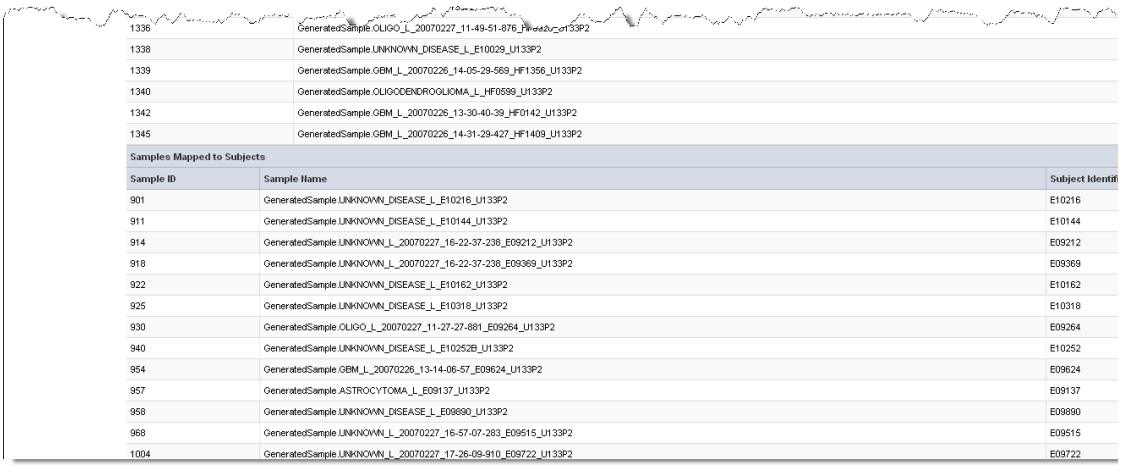
	A	B
1	E10216	GeneratedSample.UNKNOWN_DISEASE_L_E10216_U133P2
2	E10144	GeneratedSample.UNKNOWN_DISEASE_L_E10144_U133P2
3	E09212	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09212_U133P2
4	E09369	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09369_U133P2
5	E10162	GeneratedSample.UNKNOWN_DISEASE_L_E10162_U133P2
6	E10318	GeneratedSample.UNKNOWN_DISEASE_L_E10318_U133P2
7	E09264	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2
8	E10252	GeneratedSample.UNKNOWN_DISEASE_L_E10252B_U133P2

Figure 2.1 Example sample mapping file, in CSV format

Note: When you open the mapping file, make sure that the patient ID is used for mapping.

Unmapped samples continue to show at the top of the caIntegrator2 page. They were loaded from caArray, but they are not in the mapping file. These are not used for integration.

3. Scroll down the page to see samples that are mapped to the patients in the clinical data.



1336	GeneratedSample.OLIGO_L_20070227_11-49-51-876_HF1432_U133P2
1338	GeneratedSample.UNKNOWN_DISEASE_L_E10029_U133P2
1339	GeneratedSample.GBM_L_20070226_14-05-29-569_HF1356_U133P2
1340	GeneratedSample.OLIGODENDROGLIOMA_L_HF0599_U133P2
1342	GeneratedSample.GBM_L_20070226_13-30-40-39_HF0142_U133P2
1345	GeneratedSample.GBM_L_20070226_14-31-29-427_HF1409_U133P2

Samples Mapped to Subjects

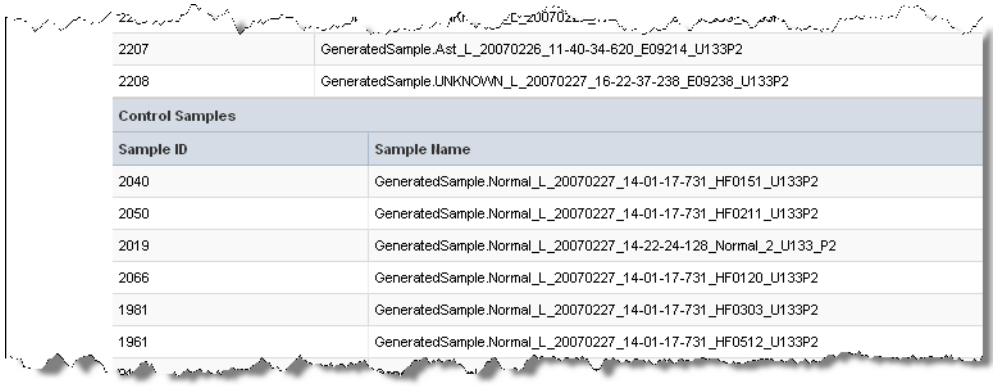
Sample ID	Sample Name	Subject Identifi
901	GeneratedSample.UNKNOWN_DISEASE_L_E10216_U133P2	E10216
911	GeneratedSample.UNKNOWN_DISEASE_L_E10144_U133P2	E10144
914	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09212_U133P2	E09212
918	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09369_U133P2	E09369
922	GeneratedSample.UNKNOWN_DISEASE_L_E10162_U133P2	E10162
925	GeneratedSample.UNKNOWN_DISEASE_L_E10318_U133P2	E10318
930	GeneratedSample.OLIGO_L_20070227_11-27-27-881_E09264_U133P2	E09264
940	GeneratedSample.UNKNOWN_DISEASE_L_E10252B_U133P2	E10252
954	GeneratedSample.GBM_L_20070226_13-14-06-57_E09624_U133P2	E09624
957	GeneratedSample.ASTROCYTOMA_L_E09137_U133P2	E09137
958	GeneratedSample.UNKNOWN_DISEASE_L_E09890_U133P2	E09890
968	GeneratedSample.UNKNOWN_L_20070227_16-57-07-283_E09515_U133P2	E09515
1004	GeneratedSample.UNKNOWN_L_20070227_17-26-09-910_E09722_U133P2	E09722

Uploading Control Samples

The Control Samples file is used to calculate fold change data, which compares “tumor” sample gene expression in the caArray experiment to the control samples to identify those that exhibit up or down regulation. Control samples can be the “normal” samples, but that is not necessarily the case.

To upload the control samples, follow these steps:

1. On the Edit Sample Mappings page, click the Map Samples link.
2. Click **Browse** to navigate for the control samples file, and click the **Upload Control Samples** File button. Scroll down the page to view the list of control samples that have been added.



2207	GeneratedSample.Ast_L_20070226_11-40-34-620_E09214_U133P2
2208	GeneratedSample.UNKNOWN_L_20070227_16-22-37-238_E09238_U133P2

Control Samples

Sample ID	Sample Name
2040	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0151_U133P2
2050	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0211_U133P2
2019	GeneratedSample.Normal_L_20070227_14-22-24-128_Normal_2_U133_P2
2066	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0120_U133P2
1981	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0303_U133P2
1961	GeneratedSample.Normal_L_20070227_14-01-17-731_HF0512_U133P2

The control samples now display toward the bottom of the page.

3. This information will be used when performing other tasks in caIntegrator2, to be described in other sections.

Adding Clinical Data

Adding Imaging Data

Managing Platforms

Adding Imaging Data

Once you have loaded clinical data and identified patient IDs, you can add either array genomic sample data from caArray which caIntegrator2 maps by sample IDs to the patient IDs in the clinical data, or you can load imaging files from NCIA, also mapped by IDs to the patient data, covered in this section. Genomic sample data and imaging data are independent of each other, so neither is required before loading the other.

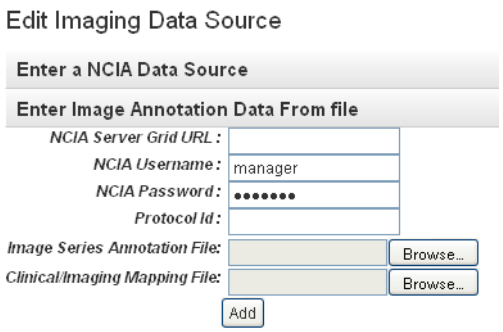
It is essential that you are well acquainted with the data you are working with--the clinical data, and the corresponding imaging data in NCIA.

Any data in NCIA can be uploaded to caIntegrator2. Imaging data consist of images and or annotations for images.

To add imaging data to the study you are creating or are editing, follow these steps:

1. On the Edit Study page, click the **Add** button under Imaging Data Sources section. Imaging data can be NCIA images or image annotations, which are uploaded in spreadsheet format.

   This opens the Edit Imaging Data Source dialog box. Enter the appropriate information in the fields..



   - **NCIA Server Grid URL** – Enter the URL for the grid connection to NCIA
   - **NCIA Username and NCIA Password**. This information is not required, as currently all data in the NCIA grid is Public data.
   - **Protocol ID** – Enter the protocol ID assigned to the NCIA file.
   - **Image Series Annotation File** – Click **Browse** to navigate to the appropriate image annotation file in NCIA. This file, which must be in CSV format, is similar to a clinical data file. It contains features and /or descriptions of images and their corresponding series IDs. If you select this option, you must also upload a mapping file, described in the next point.
   - **Clinical/Imaging Mapping File** – Click **Browse** to navigate to the to appropriate clinical/imaging file. It is similar to the genomic data sources mapping file in that the spreadsheet shows only two columns–one that shows that subject ID and one that displays the corresponding image series ID (assigned by NCIA), providing caIntegrator2 with the information for mapping patients to images or annotations.

2. Click **Add** to upload the data to caIntegrator.

3. Once the data is loaded, you must assign identifiers and annotations to the data in the same way you did with the clinical data. For more information, see *Assigning An Identifier or Annotation* on page 12 and *Searching for Annotation Definitions* on page 15.

4. To deploy the study, see *Deploying the Study*.

Adding Clinical Data

Adding Genomic Data

Managing a Study

Managing Platforms

Deploying the Study

When you are ready to deploy the study, click the **Save and Deploy** button on the Edit Study page. See questions about deployment at the bottom of page 16. caIntegrator2 retrieves the selected data from the data service(s) you defined and makes the study available for use to a study manager or to anyone else who may want to analyze the study's data. The study is then listed by name under Manage Studies on the left sidebar.You can then configure and share data queries and data lists to be available for all investigators who access the study.

When you have deployed the study, you have either uploaded data (clinical and genomic sources) or or told caIntegrator2 where all the data you are interested in resides (NCIA), and you have uploaded files that provide the mapping between the systems.

See also *Managing a Study*.

Managing a Study

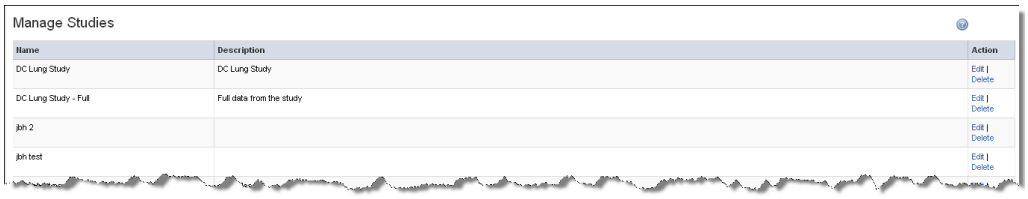
Note: A user without management privileges has no access to this section of caIntegrator2.

Once you have started to create a study or have deployed it, you can update an existing study in the following ways:

- Add new attributes (annotations) and upload relevant data to an existing study. Delete a study
- Reload subset of study data and re-deploy the study and perform new analyses
- Re-deploy the entire study with new set of data and mappings.

To update, edit or delete a study, follow these steps:

1. On the left sidebar, click **Manage Studies**. The Manage Studies page appears.



All of the in process or completed studies display on this page.

2. Click the **Edit** link corresponding to your study of choice to open the Edit Studies page. On this page you can edit any details such as adding or deleting files, survival values, and so forth. For information about working in the Edit Study, see *Creating/Editing a Study* on page 10.

3. Click the **Delete** link to delete the corresponding study.

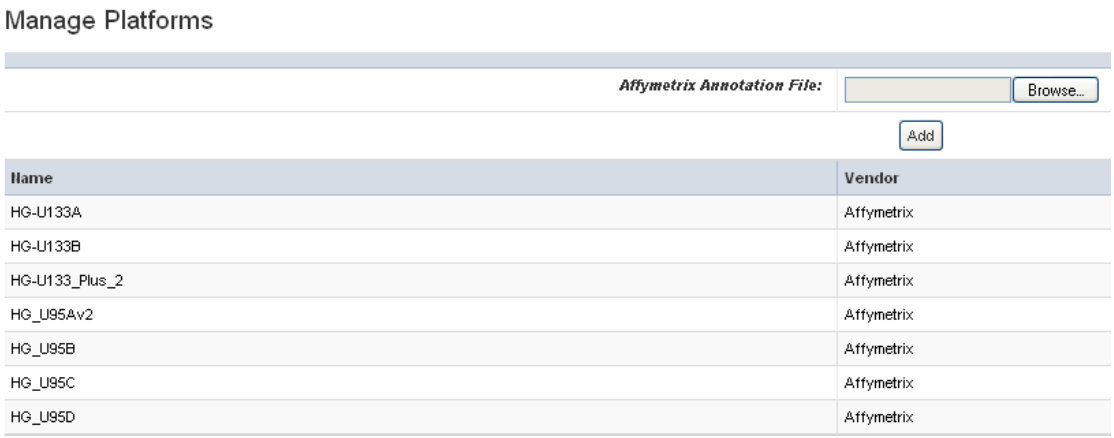
Managing Platforms

caIntegrator2 supports a limited number of array platforms, all of which originate from Affymetrix. While they do not represent all of the platforms supported by caArray, caIntegrator2 must have array definitions loaded for the platforms it supports, and be able properly load the data from caArray and parse it.

On the Manage Platforms page, you can identify, add, edit or remove supported platforms.

To manage platforms in caIntegrator2, follow these steps:

1. Click **Manage Platforms** on the left sidebar. The Manage Platforms page that opens lists the platforms caIntegrator2 currently supports, those that the system can pull from caArray.



2. You can create a study without genomic data, but you cannot add genomic data to a caIntegrator2 study without a corresponding supported array platform. Supported array platform are listed on this page. If you try to do so, an error message informs you of that fact. To do so, click **Browse** to navigate for the Affymetrix Annotation File you want to add.

3. Genomic data must be parsed and stored in caArray to be able to use it in caIntegrator2.

4. Click the **Add** button.

3/24/09 “Ongoing:”. Expect caArray to parse data. With Agilent, will take data from caArray, cai2 will parse it. Too early to tell implementation

CHAPTER

3

SEARCHING A CAINTEGRATOR2 STUDY

This chaptersection describes the processes for searching studies within caIntegrator2 using the search and browse tools.

Topics in this chaptersection include:

- *Search Overview* on this page
- *Launching a Search* on page 28
- *Managing Queries* on page 33

Search Overview

The search and browse functions in caIntegrator2 allow you to search for clinical data, genomic or imaging data that were uploaded into the application as part of a study. When gene expression and imaging data are uploaded into a caIntegrator2 study, mapping files that correlate the data in those files to patient IDs in the clinical data file must also be uploaded. When you launch a search, caIntegrato2 finds and integrates the clinical, genomic and imaging data based on the mapping files and the criteria that you define in the search query.

In a search query, you can specify criteria for just one of the data types, or configure complex search criteria that join two or three data types. The available criteria for the query were defined when the study was deployed.

The basic workflow for a study search follows these steps:

1. Select the study to be searched.
2. Select one data type:
    - **Clinical**: searches one or more uploaded CSV files for data identifiers or annotations (column headers) specified when the study was created

- **Genomic**: searches caArray experiments samples uploaded in the study for gene expression data by gene name or reporter ID.
- **Image Series**: searches NCIA files uploaded in the study for image annotations or links to images, identified by subject identifiers or image series IDs

3. Define criteria for the search in the selected data type and run the search.
4. For a more complex search, select multiple criteria from more than one data type.
5. Specify whether you want clinical/imaging annotations to display or genomic data to display.
6. Review search results.
7. Configure results column and sorting display settings. You can do this before or after you run a search. If you choose to do it after, you must re-run the search.
8. Download annotation search results as a CSV file. The CSV file contains only the data you specified in the annotation and display configurations.
9. Follows links to NCIA in the search results to view or download images located in the search.

Launching a Search

Managing Queries

Launching a Search

To initiate a search of all annotations and/or other data in a study, follow these steps:

1. In caIntegrator2, in the upper right hand corner, select the study you want to browse or perform a simple search.
2. On the left sidebar, under the first section that displays the [study name], click **Search [Study Name]**. This opens a simple search query page with five tabs.

3. On the Criteria tab, in the drop-down list, select the type of data you want to search. The listed options reflect the type of data that have been uploaded to the study.

Note: You can perform a search using one or more criteria you set in one of the data types, or you can define criteria in more than one data type per query, creating a more complex search.

- **Clinical**
- **Gene Expression**
- **Image Series**

Note: NCIA submissions are organized in the following hierarchy, which illustrates the relationship of an image series to its parent study and patient, as well as to the images in the series.:

Clinical trial > Patient (Subject) > Study > Series > Images

4. Click the **Add** button to define annotation elements for the search.

Clinical and **Image Series**

- If you select Clinical or Image Series data types, an additional drop-down list displays data elements that are annotation definitions specified when the data was uploaded into the study. Select a search criterion from among the options. You can make only one selection at a time.

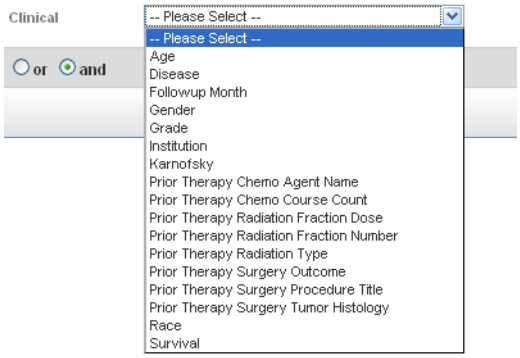


Figure 3.1 Additional clinical search criteria

- Each choice opens other fields relevant to the selection where you can further define your search query.
  - If permissible values were added when the annotation was defined, you must select among the values in a drop-list that displays on the right side of the page.

— If no permissible values were defined as part of the annotation, you have the option to enter descriptive text in a text box on the right side of the page. (X ref).

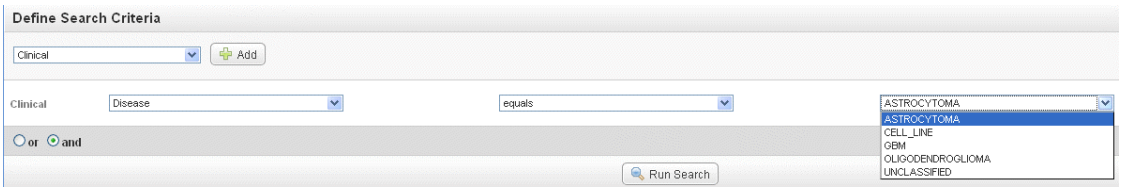


Figure 3.2 You may be able to further define search criteria when you select a specific clinical or imaging annotation element

Note: When working with image data, if only an Imaging Mapping file was uploaded when the study was created and not an Image Series Annotation file, you cannot enter image search criteria. The search results will, however, display a link that allows you to view the associated images in NCIA.

Gene Expression

Searches for gene expressions data will return results which you are permitted to see.

- For the Gene Expression selection, you can select **Gene Name** or **Fold Change**.
  - — Enter as many gene names as you like, separated by commas. Text limit is 100 characters.
  - — Fold change is used to identify genes with expression differences compared to control samples, as defined when the study was deployed in caIntegrator2. You can enter query values in greater/lesser-than-or-equal-to arguments.

Tip: the fold change option appears only if genomic control samples have been uploaded to the study.

5. To add additional criteria for the search, repeat steps 3 and 4, as appropriate. You can set more than one data type or more than one criterion for a data type. The criteria become cumulative, thus refining the search.
6. Once you have configured the query criteria, select the Boolean **Or** or **And** search operator at the bottom of the page.
   - **Or** finds a data subset with at least one of the search criteria
   - **And** finds a data subset with both/or all search criteria.
7. Click the **Remove** button to clear any data elements you have defined.
8. You can launch the search from this tab. Click the **Run Search** button. For information about the search results, see *Chapter 4 Viewing Search Results*. You may want to run the search first to see what kind of results you get before you configure the data display, described in step 9.

– or –

9. On the Columns tab, you can specify the columns you want to display in the search results data. On the Sorting tab, you can specify how the data is to be sorted. By default, only the column showing the [subject] identifier displays. For more information, see *Columns and Sorting Tabs* on page 31.

Note: As long as you are still in the current query session, you can return to the Criteria, Columns and Sorting tab to add, modify or remove data and display criteria and re-run the search. If you configure another query without saving the first, the first query will be lost. If you save the query, your current search criteria are saved.

Saving a Query

Browsing Search Results

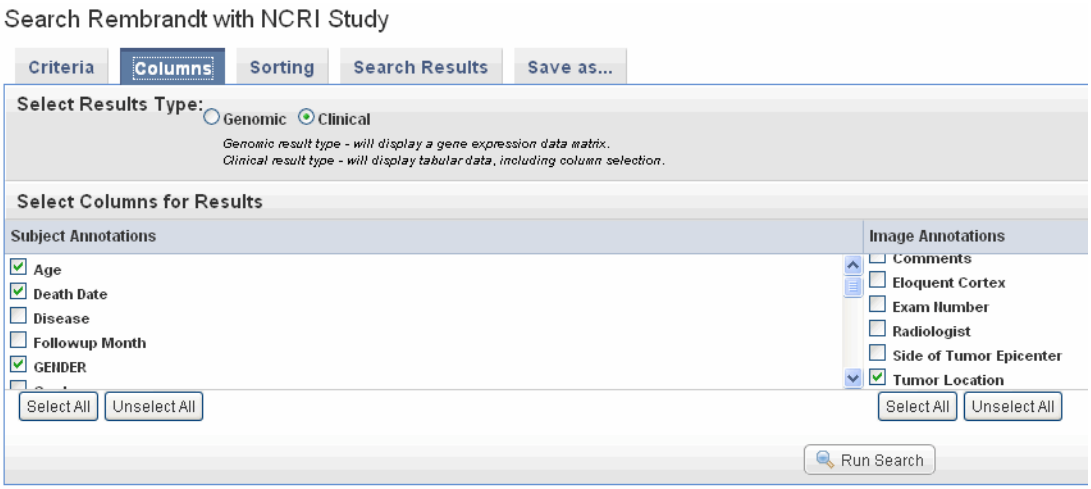
Exporting Data

Columns and Sorting Tabs

You can specify columns and sorting options for the way you want the search results to display either before you run the search, or after you run the search. If you run the search directly from the Criteria tab before setting the columns/sorting features, by default only the Subject Identifiers display on the Search Results tab. You can then come back to the Columns and Sorting tabs to expand the display options and re-run the search, having set the display parameters.

The selection you make on the Columns tab determines whether caIntegrator2 displays search results for clinical or genomic data. It filters the search based on the criteria you set on the Criteria tab, whether it is clinical, gene expression or image series data type(s). In other words, if you select clinical criteria on the Criteria tab, but select Genomic on the Columns tab, the data subset that displays on the Search Results tab is genomic data that is filtered by the clinical criteria you defined on the Criteria tab.

1. On the Columns tab, select the **Clinical** or **Genomic** radio button to search clinical data.



Clinical – Select the annotation elements that you want to display in the search results. All elements listed are column headers in the data uploaded to the study. You can make multiple selections on this list.

Note: For Clinical Annotations, the Patient or Subject Identifier display by default in the search results.

Results display as tabular data.

Genomic – Select the Reporter Type define:

- Gene Name--finds and summarizes at the gene level all reporters that match criteria for the gene you defined on the Criteria tab.
- Reporter ID--finds all reporters that map to the gene you identified on the Criteria tab

Results display in a gene expression data matrix.

Imaging – If imaging annotations have been added to the study, annotation elements also display on the lower right section of this page when you select **Clinical**. All elements listed are column headers in the image annotation data uploaded to the study. You can make multiple selections on this list.

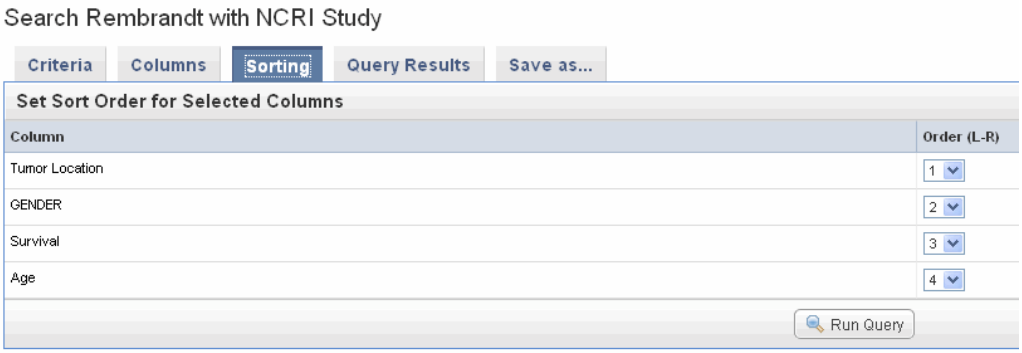
Note: If you select even one Image Annotation, the Image Series IDs display by default in the search results. If you select no Image Annotations, however, even if you have selected image series criteria on the Criteria tab, no image series IDs display in the search results. The fact that images can be located, however, in NCIA is indicated by two image-related buttons at the bottom of the Query Results page. You can open the images in NCIA, but they will be at StudyInstance UID level.

Results display as tabular data.

2. Use the **Select All** or **Unselect All** buttons to aid you in making your selections.

   The column selection is saved as part of the query if you save it. See *Saving a Query* on page 33.

3. Select the Sorting tab and indicate the column order of the Search Results.



Sorting parameters are saved as part of the query if you choose to save it using the Save Query feature. On the Search Results page, you can also sort the results by clicking on a column name.

4. Click **Run Search**. Search results display on the Search Results page. For information about the search results, see *Chapter 4 Viewing Search Results*.

Launching a Search

Saving a Query

Editing a Query

Browsing Search Results

Managing Queries

When you create a search query in caIntegrator2, you can save the query for later use or edit it. For more information, see these topics:

Saving a Query

Editing a Query

Saving a Query

1. Click the **Save As** tab and enter a **Search Name** and **Search Description**, unique to the search. *Example*: ***Batch ID 6 and female***
2. Click **Save**.

Once the query is saved, it is listed by its name under the **Study Data > Queries > My Queries** in the left sidebar, whenever the study to which the query applies is selected in caIntegrator2. Click on the saved query in this list to either edit or re-run the query. Click on the query name to retrieve query results. If you hover over the ... text for the query, a popup displays the query description.

Launching a Search

Columns and Sorting Tabs

Editing a Query

Browsing Search Results

Editing a Query

1. To edit a query, select it in the left sidebar under the **Study Data > Queries > My Queries.**
2. Change the query and display criteria on the Criteria, Columns and Sorting tabs.
3. On the Save As tab, check the appropriate options and click **Save As**. You can use the same name as the original query or modify the name as needed.

Launching a Search

Columns and Sorting Tabs

Saving a Query

Browsing Search Results

Exporting Query Results

After running a search, you can exporet the result set or a subset as a tab-delimited text file. For more information, see *Exporting Data* on page 42.

Launching a Search

Columns and Sorting Tabs

Saving a Query

Browsing Search Results

CHAPTER
4
VIEWING SEARCH RESULTS

This chaptersection describes search results that caIntegrator2 returns after queries.

Topics in this chaptersection include the following:

- *Search Results Overview* on page 35

Search Results Overview

After you launch a search of a caIntegrator2 study, the system automatically opens the Search Results tab showing the results of your search.

If you have not configured column and sort display parameters before launching the search, by default the tab shows only the [subject] identifiers and a column that allows you to select each row of the data subset.

To display and/or sort additional data, you must return to the Columns and/or Sorting tabs to set display parameters, then re-run the search. The new search results will display the additional information, with the columns and data sorted as you specified.

caIntegrator2 paginates search results into pages of configurable size (default 50) with standard paginated navigation controls. Also be sorted ascending or descendin on any displayed field.

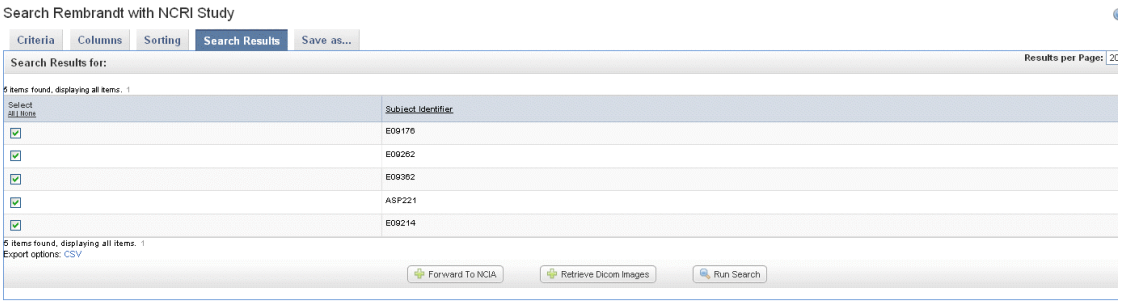
You can download search results as a CSV file. The file contains the annotations, columns and data sort configurations you specified in the search query.

Browsing Search Results

Clinical and Imaging Data

Clinical and/or imaging annotations you specified in the search query displays in the search results in tabular format. If you run the search before configuring column and

sort display parameters, only the [subject] ID that meet the criteria and a column allowing you to select each row appear on the table.



You can add details for a single subject by configuring them on the Columns tab. Annotations listed there are the column headers in the CSV file(s) that were uploaded to the study. For information about using the Columns tab, see *Columns and Sorting Tabs* on page 31.

Genomic Data

Genomic data search results display in a gene expression data matrix. Because the data was downloaded from caArray, the data permissions granted there still apply. In other words, if you have been given access to the data in caArray, you can see it in caIntegrator2.

In the matrix, samples in the experiment form the column headings. If the rows display gene name, the cells display the median gene expression value for each gene. If the rows display Probe Set, the cells display the normalized signal-based value for a given reporter for a given sample.

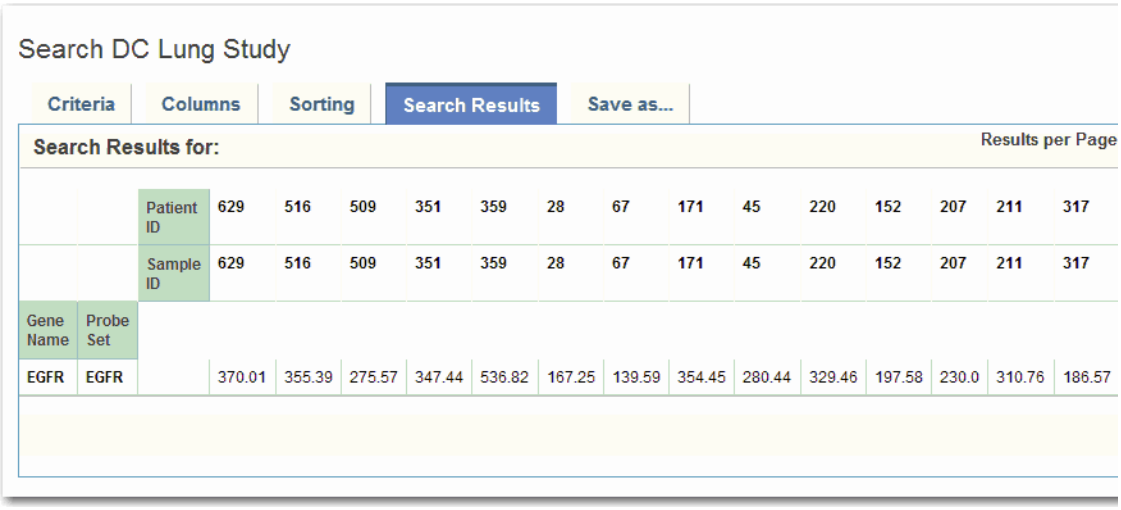
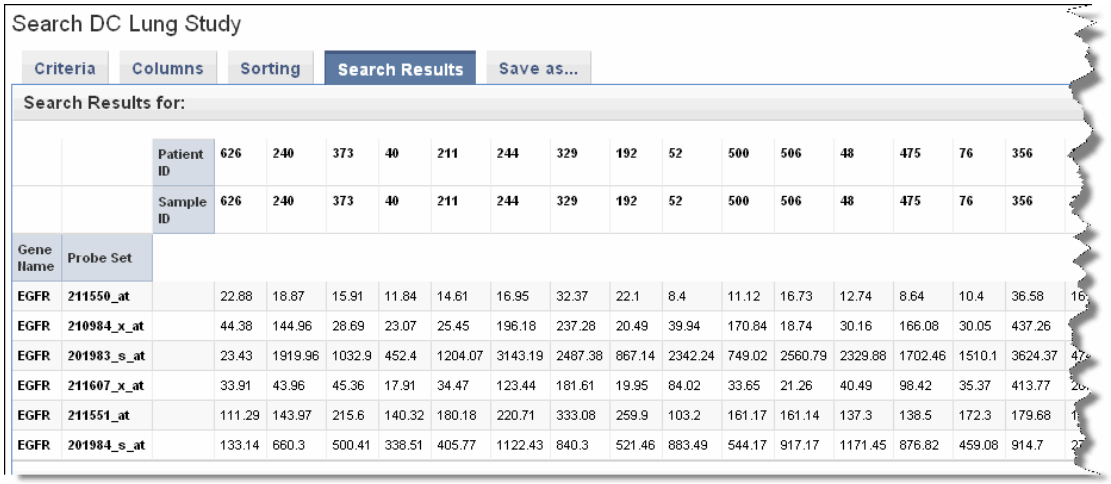


Figure 4.1 Gene Name search EGFR, Reporter Type: Gene



Search DC Lung Study

Criteria | Columns | Sorting | Search Results | Save as...

Search Results for:

		Patient ID	626	240	373	40	211	244	329	192	52	500	506	48	475	76	356
		Sample ID	626	240	373	40	211	244	329	192	52	500	506	48	475	76	356
Gene Name	Probe Set																
EGFR	211550_at		22.88	18.87	15.91	11.84	14.61	16.95	32.37	22.1	8.4	11.12	16.73	12.74	8.64	10.4	36.58
EGFR	210984_x_at		44.38	144.96	28.69	23.07	25.45	196.18	237.28	20.49	39.94	170.84	18.74	30.16	166.08	30.05	437.26
EGFR	201983_s_at		23.43	1919.96	1032.9	452.4	1204.07	3143.19	2487.38	867.14	2342.24	749.02	2560.79	2329.88	1702.46	1510.1	3624.37
EGFR	211607_x_at		33.91	43.96	45.36	17.91	34.47	123.44	181.61	19.95	84.02	33.65	21.26	40.49	98.42	35.37	413.77
EGFR	211551_at		111.29	143.97	215.6	140.32	180.18	220.71	333.08	259.9	103.2	161.17	161.14	137.3	138.5	172.3	179.68
EGFR	201984_s_at		133.14	660.3	500.41	338.51	405.77	1122.43	840.3	521.46	883.49	544.17	917.17	1171.45	876.82	459.08	914.7

Figure 4.2 Gene Name search EGFR, Reporter Type: Reporter ID

- Genomic data does not display in tandem with clinical and imaging data; it only displays when you select the Genomic result type on the Columns tab. Genomic data is however, filtered by clinical and imaging query criteria configured on the Criteria tab.
- Click the Export Options CSV link to download the CSV file whose data displays on the Search Results tab. When you do so, the CSV file opens automatically in MS Excel or similar applications for working with spreadsheets, showing the columns and sorting as you defined them in caIntegrator2 on the appropriate tabs.

Imaging Data Results

Clinical and/or imaging annotations you specified in the search query display in the search results in tabular format. If you run the search before configuring column and sort display parameters, the Image Series IDs that meet the criteria and a column containing one checkbox per row display by default. If your annotation choice on the Columns page identifies annotations such as tumor size or tumor location, the search results display image series subsets that have those annotations. The checkboxes work in conjunction with buttons at the bottom of the results page.

In reviewing search results, it is important to remember the hierarchy of submissions in NCIA. This illustrates the relationship of patient to study to series and lastly to images.

Clinical trial > Patient (Subject) > Study > Series > Images

For example, the Study Instance UID is the set of images resulting from one patient office visit. When you upload a spreadsheet of an image series, the hierarchy of images in an image series might look like this:

Study Instance UID (one office visit):

Brain (image series)

- Brain image 1

- Brain image 2
- Brain image 3

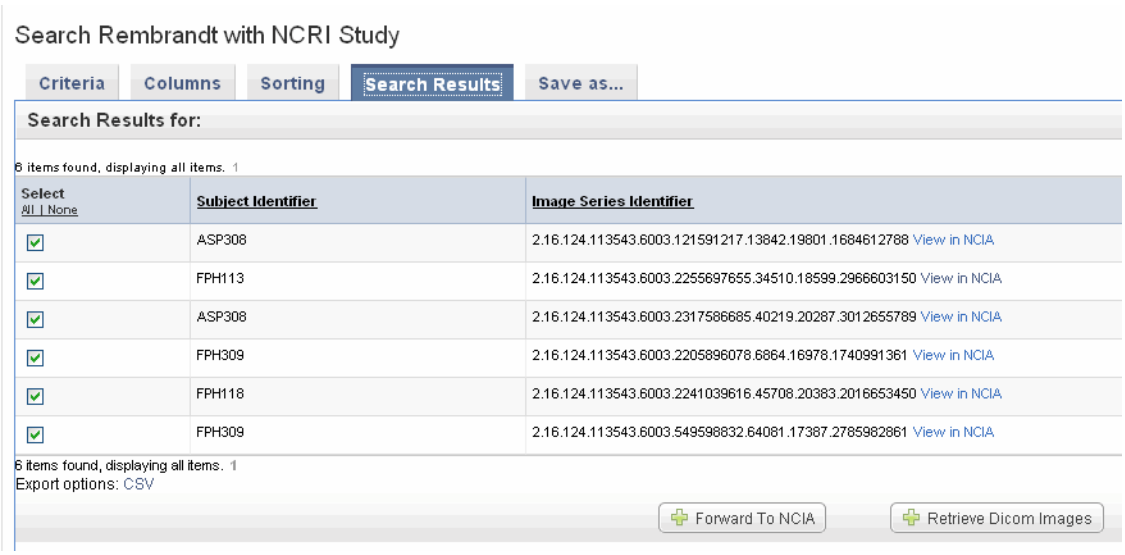
Leg (image series)

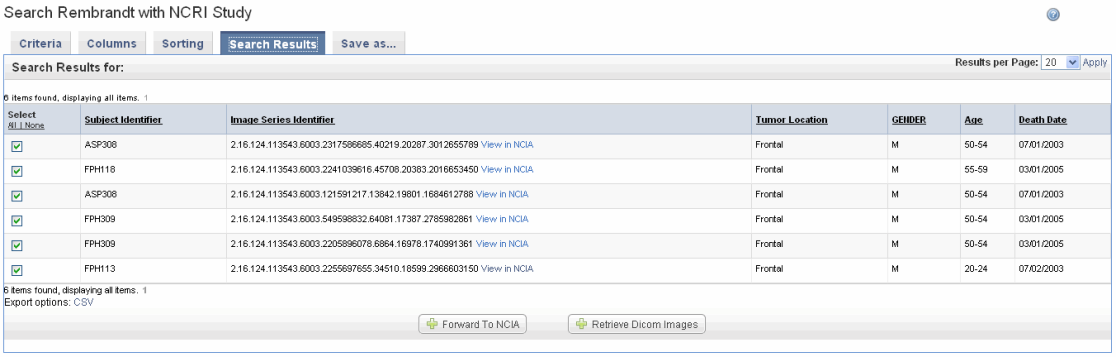
- Leg image 1
- Leg image 2
- Leg image 3

You can add details for images by configuring image annotations on the Columns tab. Annotations listed there are the column headers in the image series CSV file(s) that were uploaded to the study. Examples of image details include the following:

- All image details (name, size, etc.)
- The series that the image belongs to
- Image feature attributes
- The subject ID. Click the subject ID under Clinical Annotations on the Columns tab to display this.

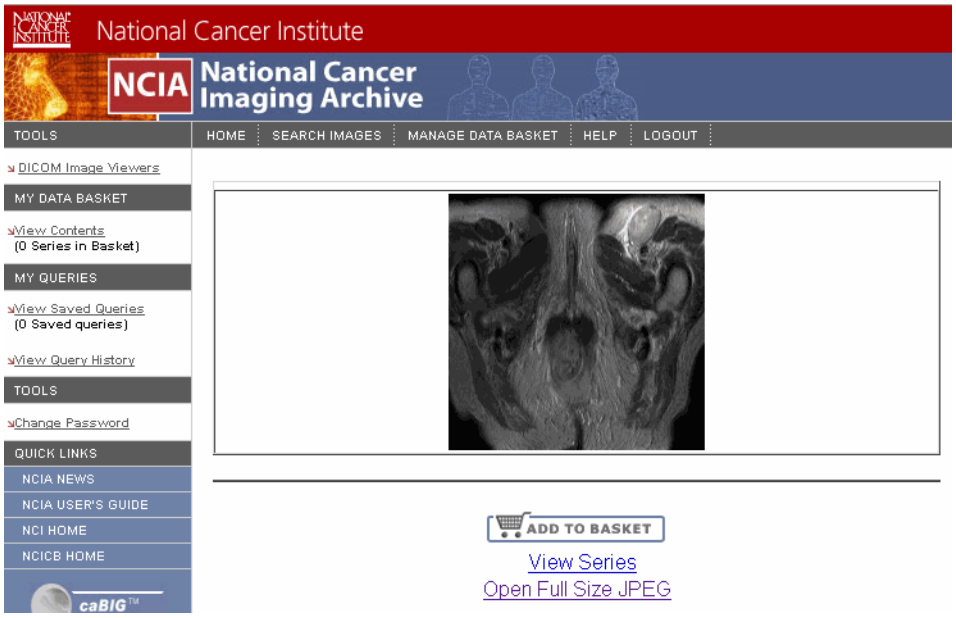
You can set display parameters for the results on the Columns and Sorting tabs. For more information, see *Columns and Sorting Tabs* on page 31. By expanding the display parameters, you can view complete details for an image.





You can click links on the Search Results tab to view or download image data.

- **View in NCIA** – This link corresponds to each Image Series listed in the results table. If you click the link, NCIA brings up the first image in the corresponding Image Series. You must log in to NCIA to see the data. On the NCIA page that opens, you can opt to view the entire series containing this image, or you can display the image as a large JPEG-formatted image. You can also add the image to the NCIA basket. For more information, see the NCIA online help or user's guide accessible from NCIA.



- **Forward to NCIA** – This button is linked to results you have selected by row. Click the button to open NCIA, where the image series you select are loaded in the NCIA image basket. In the event that the caIntegrator2 study was NOT configured with image annotation for an image series, caIntegrator2 sends NCIA a list of Study Instance UIDs, for which NCIA will add all corresponding image series to the basket. In the event that the caIntegrator2 study was configured with annotations for an image series, the system sends NCIA a list of Image Series IDs, for which NCIA adds all corresponding image series to the basket.

- **Retrieve Dicom Images** – This button is linked to results you have selected by row. Click the button to retrieve the corresponding image(s) from NCIA through the grid. NCIA organizes the downoad file by patientID, StudyInstanceUID, and ImageSeriesUID, and compresses it into a zip file. When caIntegrator 2 notifies you that the file is retrieved, the DICOM Retrieval page indicates whether the retrieved files are Study Instance UIDs or Image Series UIDs. For more information, see the note below.

Click the **Download DICOM** link to download and save the file. caIntegrator2 unzips the file and displays the list of images in the file. To open the DICOM images, you must have a DICOM image viewer application installed on your computer. X ref??

Note:

If you do not set search criteria for image annotations on the Columns tab, even though I've selected image criteria on the Criteria tab, no imaging results display. Images can be accessed in NCIA if see buttons there. see comment from previous chapter

In the search results, not all of the patients in the data subset may be mapped to image series IDs. If you select a mixture of patients that have image annotations as indicated by an image series ID and patients that do not have image annotations (no image series ID), when you click the **Retrieve DICOM Images** button, NCIA retrieves the images for the entire *NCIA study instance UID* that includes the image seriesIDs you checked.

If on the Search Results tab you select only patients that have image annotations as indicated by an image series ID, when you click the **Retrieve DICOM Images** button, NCIA retrieves images for the *NCIA image series* that were matched in the search. If the results are a mixture, but you select one specific row with a valid image annotation, caIntegrator2 aggregates to the *images series*. If results are a mixture and you select multiple rows, caIntegrator2 aggregates to the NCIA study in which multiple image series you have selected in the search results are found.

If query does not have image annotations and all check boxes are selected, results will go up to image series UID and gives all image series in it.

Sometimes when I select multiple rows, DICOM Retrieval page lists several image series IDs rather than the StudyInstance ID. What causes what? Can I predict what will happen? How do I know? Possiblity for multiple studies?

Depends on how the study was created.

If no image series showing in query results, simply mean not mapped. Will to up to Study Instance UID.

2 ways defined:

1. Image series annotations. Pulls up image series IDs.

2. Image links to NCIA. In this case, pulls up Sugdy instance level.

If query does not have image annotations and all check boxes are selected, results will go up to image series UID and gives all image series in it.

User is searching a study that has image data and image annotation for at least one image series.

1. Open a study that has imaging data associated with it, and who's pointing to production NCIA server (small-study doesn't currently).
2. Make a query that will have image series or patients who are associated to Image Studies and select X amount of those patients in the checkbox.
3. Click the "Retrieve Dicom Images" button

   Note that it aggregates to the image study
4. Now go back to Columns tab, select all image annotations and run the query again.
5. Select an image series type column and click Retrieve Dicom Images button.

   caIntegrator2 now aggregates to the Image Series that were selected and not the Image Study.

6. Select a row that doesn't have image series data, and a row that does, and push the button.

   This should aggregate to the study for the rows selected.

7. Click Forward to NCIA. YOu should see the same types of aggregation for these tests.

When the image Study is in the checked boxes (regardless of image series being there or not) aggregated up to the Image Study level.

Exporting Data

You can choose to download tabular search results as a CSV file. Click the Export .csv link at the bottom of the page. You may need to scroll the page to see it.The file contains the annotations, columns and data sort configurations you specified in the search query.

You will not see this option when genomic data displays as query results.

CHAPTER

5

ANALYZING STUDIES

Data Analysis

Once data mapping is complete, user configures the analysis tools from a selection menu. For example – selects the clinical outcome (survival, performance scores etc) for a K-M plot. User can pre-configure Gene-Pattern workflows to be made available to the end user. Link-outs to 3rd party tools either within caBIG or external should be made possible. Simple link-outs with well-structured query strings are preferred.

Data analysis and visualization will be performed using

- Integration with Gene-Pattern
- Integration with GeWorkbench (Web version)
- Existing caIntegrator2 analysis capabilities.

Much about these kinds of analysis in the I-SPY and Rembrandt documentation.

Performing Data Analysis

user analyzed data will never see lists of IDs/samples. Just be able to use the data to continue to work with it here. Allowing integration of clinical and genomic data

Include specifics about different tools

...follow these steps:

1. The system presents the user with a form or wizard interface to name the analysis method configuration and to set the analysis method parameters.
2. The Research Investigator supplies the necessary analysis method parameters.
3. The system saves the supplied analysis method parameters and executes the analysis job.

4. The system informs the user that the analysis method is complete.
5. The user reviews the analysis results.

As a shortcut, the Research Investigator may launch the functionality described in the use case Manage Queries at any point where the analysis parameter selection is requesting a data list. After defining or running the desired query to use to populate the data, the Research Investigator is returned to the analysis method definition with the selected query results used as the input data.

Fold-change is launched during query configuration. Others?

TrWorkbench will support at least the analysis and visualization methods that are currently bundled in caIntegrator2 (specifically those in Rembrandt). These tools include the following: currently not implemented??

- KM Plot – select from a list of clinical annotation fields that have permissible values associated (and 'no value' if there are subjects with no value for that field) with them. They can then select two or more of those values (and each value should have a number next to it showing the number of subjects with that value). USBAT select a measure of survival from a list of one or more of them. USBAT to submit the request and get back a Kaplan-Meier Survival Curve plot.

  KM Plot – Select a study with gene expression data; specify a single gene whose expression value can be used to split the subjects into three categories: high, low and intermediate expression.

  Must define at least one survival values before creating KM Plot.

  After a user clicks 'Create Plot', the ensuing plot should contain the P-Value that was used in the calulation. USBAT see the number of patients in each category on the plot.

Query and KM plot functionality in caIntegrator 2 provide functionality to allow users to compare expression levels for a given gene against expression levels for a set of controls samples designated at study definition. The relative expression level is referred to as “fold change” and the numeric value for a given sample and reporter combination is the ratio of the expression value for that particular reporter for the given sample to a reference value calculated for that reporter across all control samples. The reference value is calculated by taking the mean of the $\log_2$ of the expression values for all control samples for the reporter in question. The $\log_2$ mean value (n) is then converted back to a comparable expression signal by returning 2 to the exponent n.

In queries that include a fold change criterion and that are configured to return genomic data, the raw expression values are replaced with the calculated fold change value.

Creating Kaplan-Meier Plots

The Kaplan Maier method is used for survival analysis of comparative groups of patients or samples. For example, you might identify a group of patients with smoking history and compare survival rates with a group of non-smoking patients. You could compare groups of patients with varying gene expression levels. The key is to first identify subsets of patients that meet criteria you want to establish, and calculate a Kaplan-Maier curves based on their survival probability as a function of time. Survival differences are analyzed by the log-rank test. As an

To create a *Kaplan-Meier* survival plot for gene expression data (K-M Gene Expression), follow these steps:

1. From the Simple Search page, select **Create Kaplan-Meier survival plot for Gene Expression Data**.
2. Enter a gene keyword, for example, enter a *HUGO* gene symbol, such as `EGFR` or `WT1`, to plot a gene expression profile based on the expression of your gene of interest.

   Note: If you do not enter a valid gene symbol, the following message appears: *The gene you entered is either invalid, or not in the database. Please select another.* Close the message window, and enter another gene symbol.
3. From the **Restrict to Sample Group** drop-down list, select a saved sample group.

Click the **Go** button.

Redrawing...

To redraw a K-M Gene Expression data, follow these steps:

Note: If you restricted the search to a group, **Constrained to group** appears at the top.

1. To dynamically modify the fold change thresholds and redraw the plot, adjust the **Up-Regulated** and **Down-Regulated** values.
2. Specify a **Unified** or a **Affymetrix Reporter Type**.
3. To visualize the K-M plot for the unified probeset, click the **Reporters** drop-down list *(Figure 5.1)*.

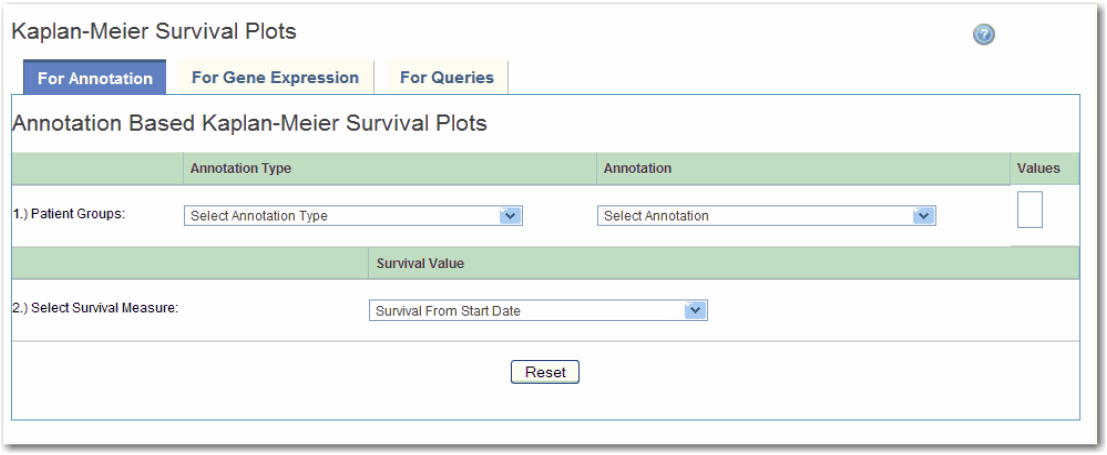
Figure 5.1 Redrawing a Kaplan-Meier Gene Expression data

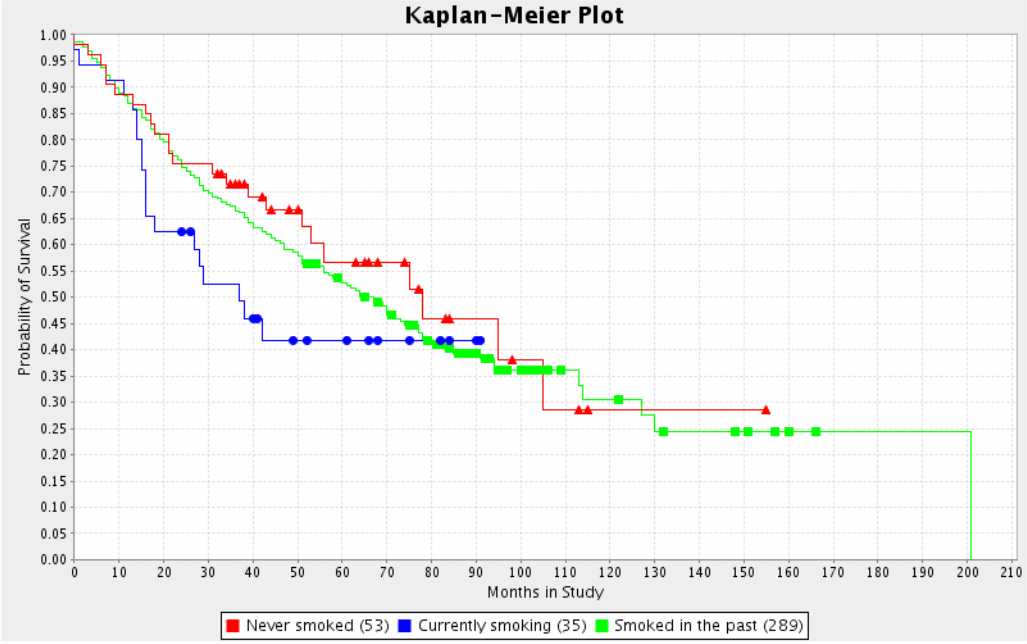
4. Select an individual reporter or one of the following options:
   - **Median** is the median value of all Reporters (default).
   - **Mean** is the mean value of all Reporters.
5. Click the **Redraw Graph** button.

1) Select a study with Gene Expression data. 2) Select Generate KM Plot on the left menu 3) Select "For Gene Expression" tab 4) Input valid Gene and up/down regulated parameters 5) Hit "Create Plot" and notice that there's a KM plot on the tab

KM Plot for Annotations

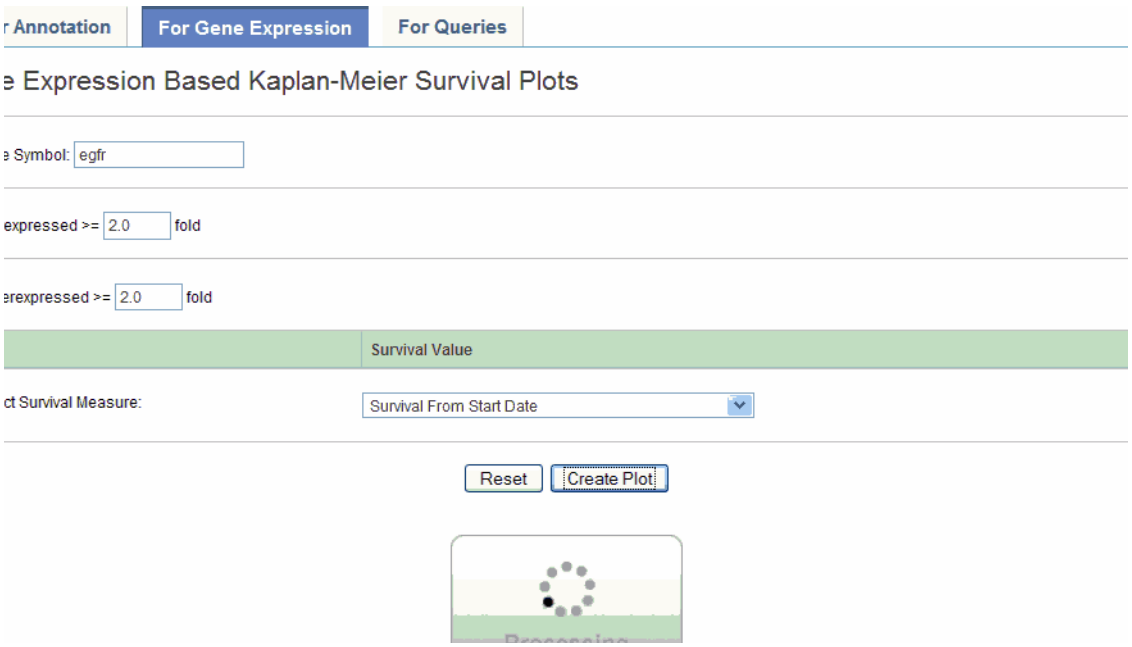
Create a KM Plot graph and validate that the number of subjects for each group appears embedded in the legend of the graph. Also validate that all groups with 0 patients are in the legend and not printed in Red error text anymore.

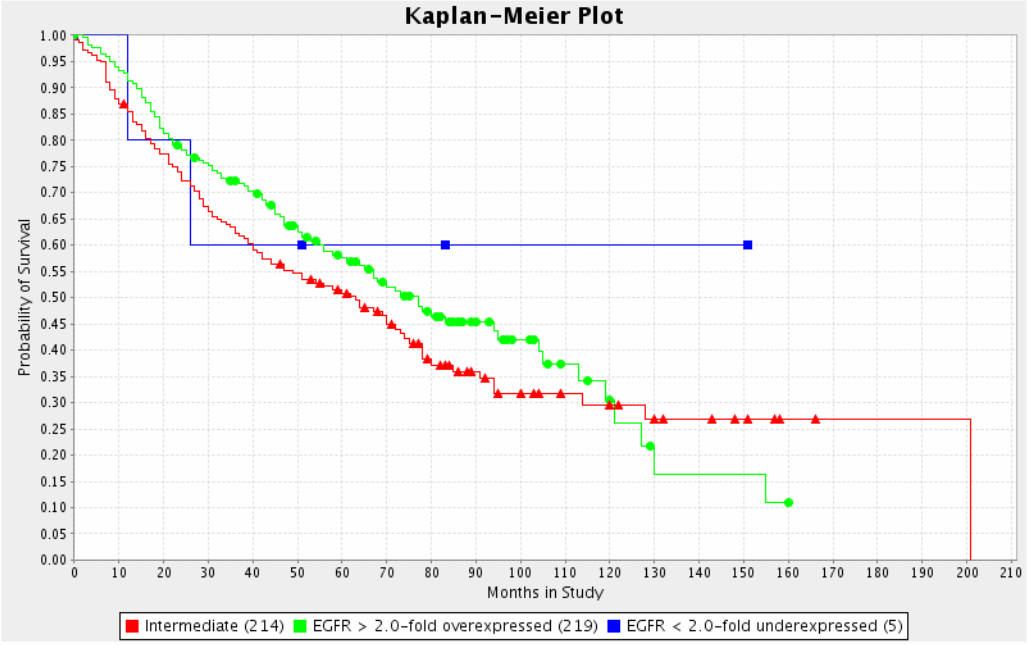




Log-rank P-value for significance of difference in survival between groups
Never smokedvs.Smoked in the past=0.53
Never smokedvs.Currently smoking =0.12
Currently smokingvs.Smoked in the past=0.19

KM Plot for Gene Expression

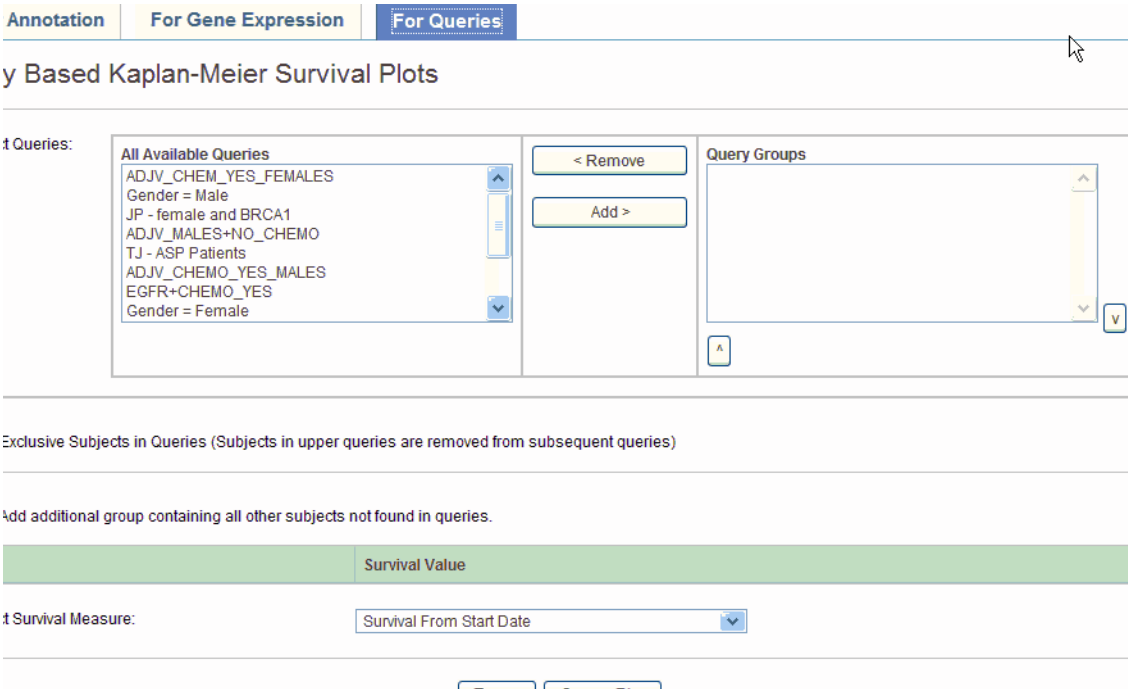




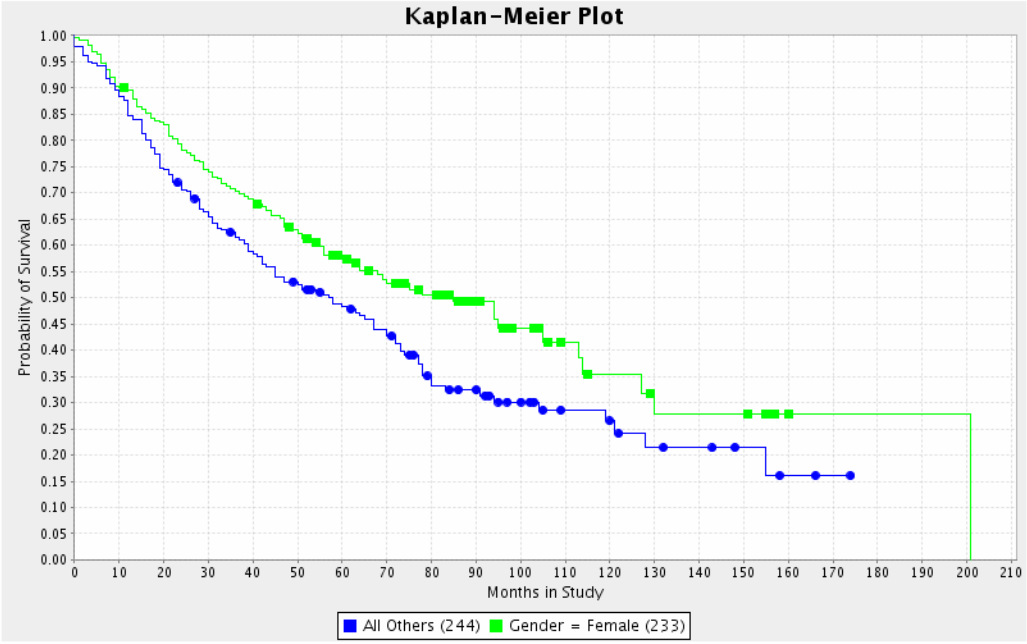
Log-rank P-value for significance of difference in survival between groups

EGFR < 2.0-fold underexpressedvs.Intermediate =0.45
EGFR < 2.0-fold underexpressedvs.EGFR > 2.0-fold overexpressed=0.49
EGFR > 2.0-fold overexpressedvs.Intermediate =0.14

KM Plot for Queries



asdfl;lasdlfk



Analyzing Data with GenePattern

caIntegrator2 provides a link to GenePattern. In the dialog box, specify connection information. GenePattern pulls up analysis modules for gene expression analysis. After you've estabished the criteria, and click ...., the data is sent to GenePattern for

analysis. Want GP to identify a list of userful genes. Work in GenePattern and generate a .csv file.

GLOSSARY

This glossary defines acronyms, abbreviations, and terminology used in this guide.

Term	*Definition*
caBIG	cancer Biomedical Informatics Grid
caBIO	Cancer Bioinformatics Infrastructure Objects
caCORE	cancer Common Ontologic Representation Environment
caDSR	Cancer Data Standards Repository
caMOD	Cancer Models Database
CGH	Comparative Genomic Hybridization
EBI	European Bioinformatics Institute
EVS	Enterprise Vocabulary Services
MAGE 1.1	MAGE 1.1 is a widely-used microarray data standard or guideline
MAGE-ML software format	XML-based standard for representation of microarray data
MIAME 1.1	MIAME1.1. is a standard or guideline for the minimum amount of information required to make a microarray record useful to others.
MGED Ontology	MGED Ontology is a controlled vocabulary standard that concisely defines terms as they relate to Microarrays and caArray as a whole
MGED	Microarray Gene Expression Data Society
MMHCC	Mouse Models of Human Cancers Consortium
NCI	National Cancer Institute
NCICB	National Cancer Institute Center for Bioinformatics
URI	Uniform Resource Identifier
URL	Uniform Resource Locators
XML	Extensible Markup Language (http://www.w3.org/TR/REC-xml/) - XML is a subset of Standard Generalized Markup Language (SGML). Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML has been designed for ease of implementation and for interoperability with both SGML and HTML

APPENDIX

B

DATA IMPORT CONFIGURATIONS

Subject Clinical Data Configuration

The following clinical data configuration information is collected:

- Clinical Data Source (delimited text)
- Protocol Id (of study to import)
- For delimited text see Special Requirement: Delimited-Text Annotation Import. For subject annotation files, one field must be identified as the subject identifier.
- See Special Requirement: Annotation Field Configuration for details on specification of visibility and browse configuration.

Delimited-Text Annotation Import

Delimited-text annotation files must be in standard comma-separated value format. The file must include a header line that specifies the name for each field. Each row of data must contain the same number of values as the header row. The file must include a column that will be designated as the identifier (e.g. subject identifier, sample identifier, etc.) for each row. Optionally a file may include a single column that will be designated as a time-point indicator. Each row must contain a unique combination of identifier and time-point indicator of a unique identifier if no time-point is included. An example of the content of a file including a time-point is shown below.

“patientId”, “timepoint”, “bloodPressure”, “weight”
“1234”, “T1”, “120/80”, “180”
“1234”, “T2”, “125/80”, “190”
“5678”, “T1”, “120/85”, “200”

After upload of the file, the Study Manager must indicate for each field:

- Field type (identifier, timepoint indicator, text, integer, float or boolean)

- After specification of these types, the file will be validated to ensure that the values are valid for the types selected and that the file conforms to the requirements given above.

Annotation Field Configuration

For each annotation field (regardless of the source), the Study Manager must specify the following information:

- Annotation semantics: each annotation field (whether associated with a subject, image series, image or sample) must either:
  - be associated with an existing annotation definition known to the system,
  - be associated to an existing CDE in caDSR or
  - have sufficient semantic metadata recorded so that the field may be submitted for registration as a CDE in caDSR.
- Field authorization: each field must be either declared publicly visible or a restricted to a list of groups. The default will be the visibility settings given at the study level.
- Whether the field is to be included in the results list for a given entity type (i.e. Subject, Sample, ImageSeries or Array Data) when browsing data (See Use Case: Browse Study Data).
- Whether the field is to be included in simple single-input searches when browsing data (See Use Case: Browse Study Data).

Sample Data Configuration

Sample data may be uploaded from either caArray 2 or from delimited-text import. Samples imported from caArray 2 may have annotation updated by use of the delimited-text import functionality if sample annotation is required. Import from caArray 2 requires specification of the following information:

- caArray server hostname
- caArray server JNDI port
- caArray username
- caArray password
- Either the experiment identifier (to import all samples in the experiment) or a file containing a comma-separated list of samples in the format “experiment identifier”, “sample name”.
- Mapping of samples to subjects. This may be specified by a comma-separated list in the format “subject identifier”, “sample identifier” or by a regular-expression based mapping formula.

When samples are imported via delimited-text import, the time-point is associated to the sample itself. This means that each sample may be associated with only one time-point (i.e. multiple time-points for the same sample are invalid).

Genomic Data Configuration

All genomic data (i.e. array data) is imported from caArray 2. First the Study Manager must specify sufficient information to map study samples to caArray 2 samples. If all samples were imported directly from caArray 2 as described in Special Requirement: Sample Data Configuration, no further information is required for this step. If samples were imported via delimited-text, the Study Manager must specify

- caArray server hostname
- caArray server JNDI port
- caArray username
- caArray password
- A mapping of caIntegrator 2 sample identifiers to caArray 2 samples, specified as a comma-separated list in the format “caIntegrator 2 sample identifier”, “caArray 2 experiment identifier”, “caArray 2 sample name”.

The system will enable the Study Manager to navigate easily to the selected caArray 2 instance.

Next, the system will indicate the available platforms and array data types available for the study samples. The Study Manager will indicate which platforms and data types to import and for each platform/data type combination will indicate:

- Whether to import the data
- The visibility of the data; either public or restricted to a set of groups. Low-level genotyping data (raw data and normalized) will always have restricted visibility.

Imaging Data Configuration

The following imaging data configuration information is collected:

- NCIA grid server hostname (defaults to NCICB instance)
- NCIA grid server port (defaults to NCICB instance port)
- Protocol Id
- Mapping of NCIA Patients to subjects imported from clinical data source. This may be specified by a comma-separated list in the format “subject identifier”, “NCIA patient identifier” or by a regular-expression based mapping formula.
- Which annotation fields to import from NCIA.
- The system will enable the Study Manager to navigate easily to the selected caArray 2 instance.

Additional annotation for either images or image series may be imported using the delimited-text import functionality.

INDEX

C

caArray
  glossary 13
Captions 3
Character formats 6
Cross-reference formats 7

F

Figures 3
Figure table 5
Format
  character 6
  cross-reference 7
  lists 4
  table 4

G

Glossary 13

H

Headings 2

L

Lists 4

T

Table
  figure 5
  formats 4
  header 4
  no border 5
  no header 5