TIES Quality Assurance Plan

- Introduction
  - Purpose and Scope
  - Team
  - Process
- Phase 1
  - Entry Requirements
    - Sample Dataset Characteristics
  - Installation QA Checklist
  - How to determine if a report is correctly processed?
- Phase 2
  - Entry Requirements
  - Import QA Checklist
- Phase 3
  - Entry Requirements
  - Test Case Set Generation
  - De-identification QA
    - Determine validation requirements for your institution
    - Determine number of reports to be validated
    - Validation of de-Identification in TIES
    - How to identify de-identification errors
    - Re-configuring DeID
    - Sample De-identification QA Checklist
  - Section Detection QA
    - Sample Section QA Checklist
  - Query QA
    - Precision Tests
    - Common Queries
    - Institution Specific Queries

Introduction

This Quality Assurance Plan(QAP) sets forth the process, methods, and procedures that can be used by an institution to perform Quality Assurance function over a newly installed TIES node.

Purpose and Scope

This QAP provides a foundation for managing an institution's quality assurance activities to ensure that the system is:

- installed and configured properly to meet the institution's specific data characteristics.
- loaded with all the reports that the institution intended.
- returning accurate search results for known queries.

Team

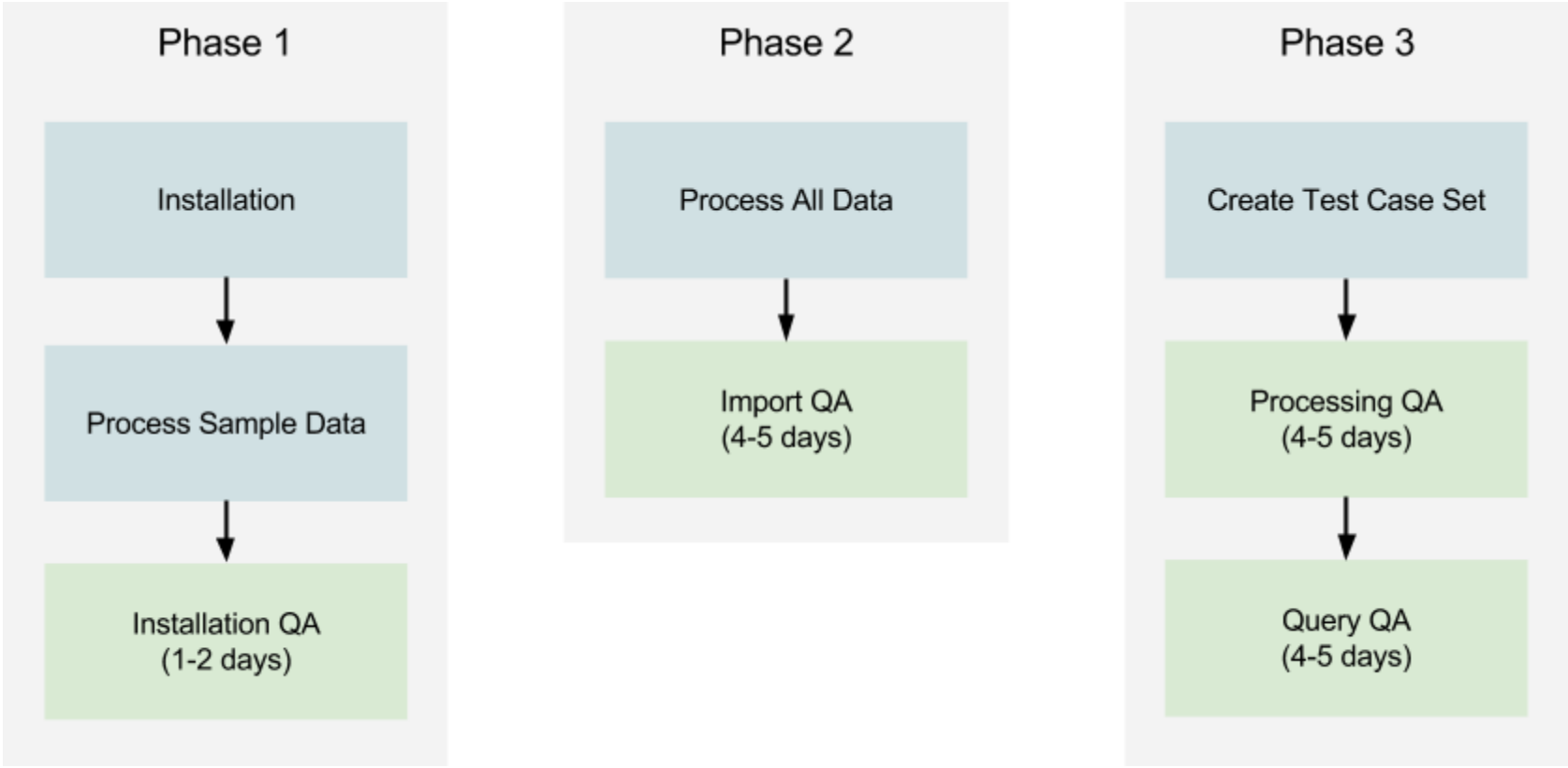
The QA team typically needs the following members to perform all the tasks listed in this plan.

Role	**Responsibility**	**Skills**
QA Manager	Responsible for the entire QA process. Coordinates the activities of the team and generates deliverables.	Can identify correct personnel for QA team, and supervise their activities
Technical Expert	Responsible for QA activities that test installation and configuration.	Comfortable working with the TIES configuration files, and databases.
Domain Expert	Responsible for QA activities that review accuracy of search results	Qualified to determine whether a report has been correctly coded and retrieved in response to a search result

Process

The QA activities are grouped into three phases that occur in sequence. The next phase does not begin until the previous phase's QA is completed and accepted.

The QA process is iterative in nature. If any major issues are found, they should be fixed and the QA process repeated until satisfactory performance is achieved.



Phase 1

This phase tests whether the TIES software has been installed and configured properly by testing its performance over a sample data file. This is to ensure that any configuration problems are caught early on before we process the entire dataset.

Phase 1 QA is generally performed by the Technical Expert.

Entry Requirements

1. TIES node is installed and configured.
2. A sample data file is imported and processed by all TIES pipelines.

Sample Dataset Characteristics

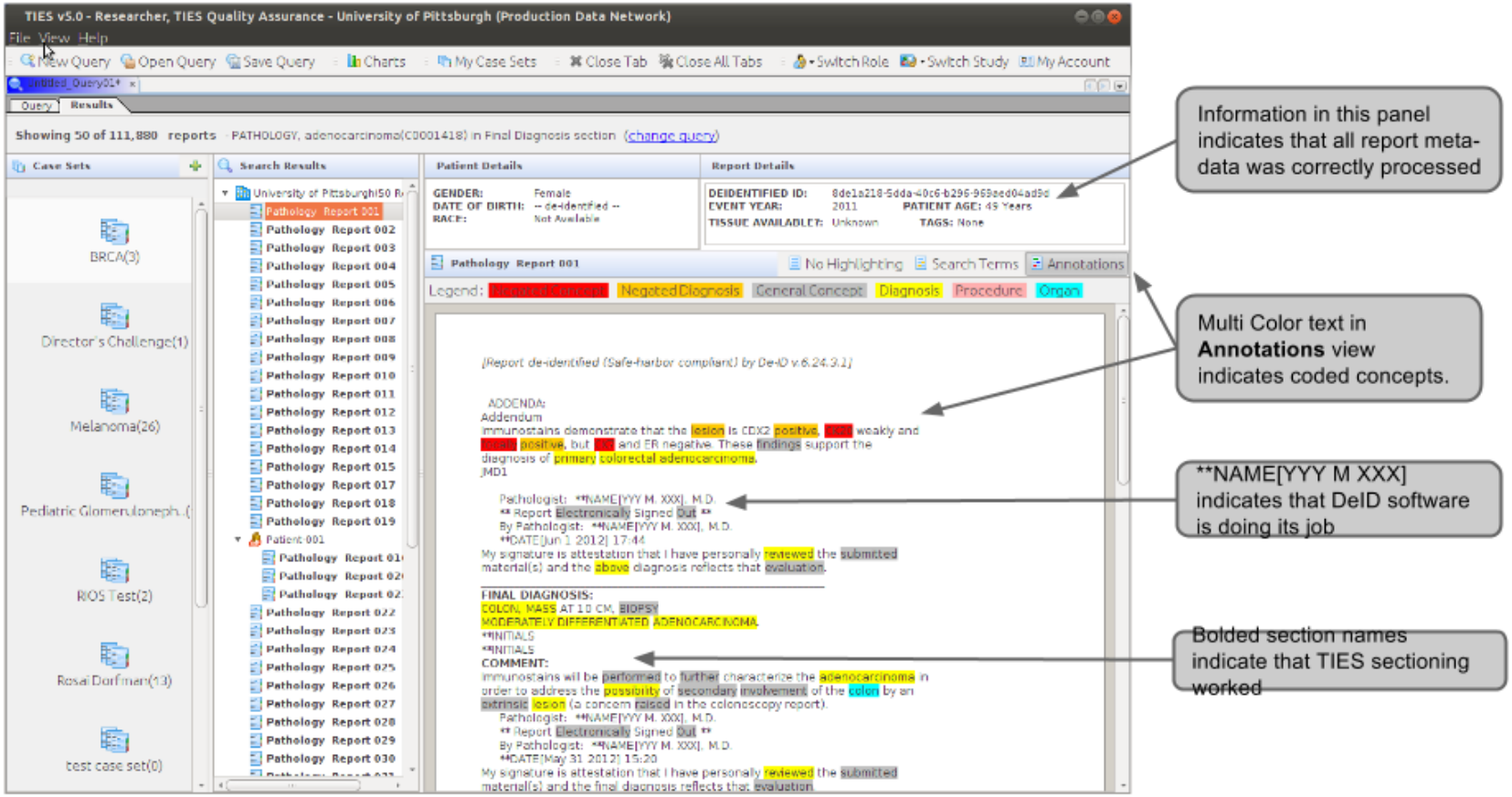
1. Should consist of the real data and not test data..
2. Should contain at least 100 reports and at least 50 patients.
3. Should ideally be a true sample of the entire dataset. i.e. it should be spread out over multiple years and contain all major types of reports.

Installation QA Checklist

No.	Validation Test	Yes	No	N/A	Comment
1.1	TIES tomcatPub and tomcatPvt have been installed and are running under specific username that has .globus/ folder under its home directory	☐	☐	☐	
1.2	TIES MySQL database is up and has ties_public, ties_private and ties_ctrm (as well as their _test equivalents) schemas	☐	☐	☐	
1.3	SectionHeaderConfig.txt used by HL7 Pipeline and Deid pipeline is configured properly with section information. GenderConfig.txt, RaceConfig.txt and EthnicityConfig.txt, used by HL7 Pipeline, is configured properly.	☐	☐	☐	
1.4	Login with admin account and change password. Grant researcher, preliminary user and honest broker roles to admin user.	☐	☐	☐	
1.5	Create new user account and new research study. Grant researcher and honest broker roles to user and assign as honest broker to a new research study.	☐	☐	☐	
1.6	Login as honest broker with new user account, run a query and get results. Both de-identified reports and identified reports are visible.	☐	☐	☐	
1.7	Verify that the report is correctly processed. Section headers are **bolded** and annotations are present in the de-identified report. (See image below)	☐	☐	☐	

REVIEWER	DATE

How to determine if a report is correctly processed?



Phase 2

This phase tests if the data has been imported correctly.

Phase 2 QA is generally performed by the Technical Expert.

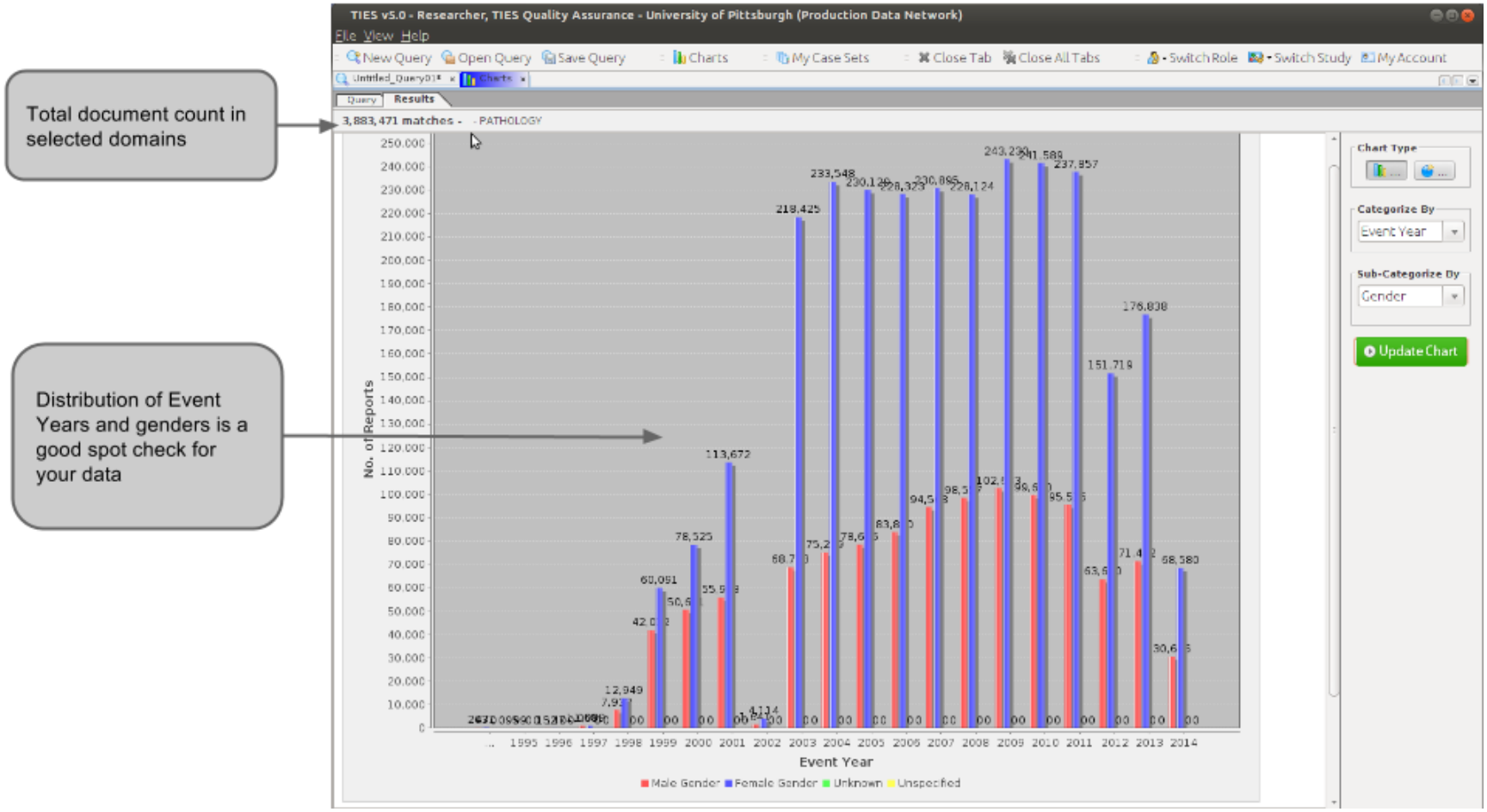
Entry Requirements

1. Phase 1 is completed with all tests passed.
2. All data is imported and processed.
3. The following information is available
    a. Total no. of reports that should have been imported.
    b. Years that the reports span.

Import QA Checklist

No.	Validation	Yes	No	N/A	Comment
2.1	Login as admin user and select preliminary user role. Select a document type, select bar chart and bin by collection year. A bar chart displaying no. of reports by year will be displayed. Confirm that the total count is accurate and the no. of reports in each year is accurate. (See example chart below)	☐	☐	☐	
2.2	Switch the bin to Gender and update chart. A chart displaying no. of reports by gender will be displayed. Confirm that the chart is accurate.	☐	☐	☐	

REVIEWER	DATE



Example chart showing binning by Event Year

Phase 3

This phase tests performance of the TIES system on your entire dataset.

Phase 3 QA is generally performed by the Domain Expert.

Entry Requirements

1. Phase 2 is completed with all tests passed.
2. All data is imported and processed.
3. A test case set is generated.

Test Case Set Generation

The test case has the following characteristics

1. Should consist of at least 10 reports from every year of data imported.
2. Should ideally be a true sample of the entire dataset. i.e. it should be spread out over multiple years and contain all major types of reports.

Create the test case set by running a TIES query that fetches all reports for a specific year. Add 10 random reports from the result to a case set. Repeat this query for all year's that you have imported, and add 10 reports from each year to the case set.

De-identification QA

This QA involves Domain Experts reviewing all the documents in the test case set and errors that may occur in the de-identification of the documents.
You may also want to read the De-identification Standard Operation Procedure for the TIES Cancer Research Network for more information.

Determine validation requirements for your institution

1. Engage the stakeholders at your Institution to determine institution-specific requirements for validation and documentation. It is expected that individual institutions may set institution-specific thresholds for assuring sufficient deidentification, intervals for QA, or other requirements.
2. Determine any thresholds for de identification (e.g. frequency of PHI elements, etc) for initial and ongoing QA.

3. Determine the frequency of QA checks for ongoing QA at your institution
4. Determine any requirements for randomization in report selection for QA checks for initial and ongoing QA.
5. Create a document outlining data contained in the reports loaded into TIES that may be viewed as PHI or could be used to identify a patient. In this document delineate whether each data item is considered PHI [RC3] and whether the system is expected to remove or tag it. This document will be used by the QA managers and their team to validate the de-identification process.

Determine number of reports to be validated

1. Enumerate the number of reports being loaded into TIES that require QA checking
2. If your IRB or Legal Counsel requires that a specific methodology be used (e.g. sample size calculation) you will define the steps for determining number accordingly
3. For example, for sample size determination, determine the number of reports that need to be manually checked for appropriate de-identification by entering the following criteria into the sample size calculator found at: https://www.mccallum-layton.co.uk/tools/statistic-calculators/sample-size-calculator/
    a. Margin of error that is acceptable
    b. Population size = total number of reports in the system that require QA
    c. Select a confidence level of 95% or above
    d. Then, use the calculated value to set your criteria for the number of reports that will require manual validation of de-identification.

Validation of de-Identification in TIES

1. The QA manager logs in to TIES as an Honest Broker and runs a query that will result in the return of all the reports which require validation (e.g. reports from a specific year). Check to assure number of reports returned is at least the number that you are required to validate.
2. Re-run the query checking the '*Randomize'* checkbox on the right hand side of the query interface. Also indicate how many reports you would like returned – this number needs to be larger than the number of reports required to be validated (calculated above). Using the built in randomize function will return a randomized subset of reports from the total set required for validation.
3. Create a case set folder appropriately labeling as QA set with a description of the reports being checked (e.g.QA

CY2002pathology reports).

4. Choose reports that will be checked by clicking and dragging them into a case set. [RC6]
5. Create an excel output of all cases to be checked and use that excel document to record the following information:
    a. Report de-Identification Pass/Fail/Over scrubbing
    b. Failure Detail –
        i. Data item
        ii. Issue
    c. Notes – use this field to add any comments or notes regarding the findings

6. Proceed with the de-identification checking by reviewing all chosen reports in the case set. Using the TIES PHI listing of allowable data elements document, review each report contained in the case set. Record a Pass if no PHI or PHI like data remained in the pathology report and Fail for those reports where a data item that should have been removed was not. For failed reports, please record the data item that was missed and what you think was the reason or issue if it can be assessed.
7. Once all reports have been validated and results have been recorded, enumerate and generate a summary count of the results.
8. Review results with institutional stakeholders. Determine the level of risk that each failure generates and whether it is necessary to address the failure in the system.

How to identify de-identification errors

1. **De-identifier Over scrubbing**
   The de-identifier may incorrectly classify a piece of text as a HIPAA identifier and then remove it from the text. This typically happens when the de-identifier mistakes a disease name for a person name. This error is marked only when significant clinical information is lost due to over scrubbing.

- **De-identifier Under scrubbing**
  The de-identifier may miss some HIPAA identifiers in the text if they are not present in its dictionary. This error is marked by noting the under scrubbed identifier under the appropriate column.

  The 18 HIPAA identifiers that are to be completely scrubbed are:

1. Names;
2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Phone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social Security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

Re-configuring DeID

After all documents are reviewed the data from the sheet should be aggregated to create the three unique lists (one for each identifier type). The Patient Information and Healthcare Provider name lists can be added to the DeID dictionary and all the reports should be de-identified again. The Healthcare Provider initials are not considered PHI by themselves. However, if their

scrubbing is desired, they can also be added to DeID dictionary.

Sample De-identification QA Checklist

Here is an example sheet showing just three of the 18 HIPAA identifiers. You can add a column for each type of identifier.

		Under Scrubbing		
Report No.	Patient Information	Healthcare Provider Name	Healthcare Provider Initials	Over Scrubbing
1	John Smith	Sunflower Hospital	None	None
2	None	None	DFT:nr	
3	None	Dongfeng	None	

Section Detection QA

This QA involves Domain Experts reviewing all the documents in the test case set and errors that may occur in the section detection configuration of TIES. Since the test case set is the same as the De-identification QA, the domain expert may note these errors when they are reviewing the documents for de-identification errors.

1. Create the test case set by running a TIES query that fetches all reports for a specific year. Add 10 random reports from the result to a case set. Repeat this step for all year's that you have data for.
2. Review each document in the case set and look for the following errors:
    - **Missed Section**
      A section may not be detected as a distinct section, and may be treated as part of the previous section. Sections typically begin with section headers that are in **bold.** If the reviewer finds section headers that are not bolded, that section is to be treated as a missed section.
- **Incorrectly coded section**
      A section may be incorrectly skipped by the concept coder, meaning, it may not have any concept annotations. Alternatively, a section that should not be concept coded (sections with synoptic data) gets coded

3. After all documents are reviewed the data from the checklist should be aggregated to create three lists - missed sections, sections coded incorrectly and sections skipped incorrectly. This information is then to be used to modify the section configuration and all documents need to then be re-coded.

Sample Section QA Checklist

Report No.	Missed Section	Section Coded Error	Section Skipped Error
1	None	SYNOPTIC	FINAL DIAGNOSIS
2	GROSS DESCRIPTION	None	

Query QA

This is an holistic test of the coding, search and retrieval functions of the TIES node. It involves comparing TIES search results with results generated using existing systems and processes.

Precision Tests

Listed below are 30 queries that were part of a scientific evaluation of the TIES system. You may run these queries in TIES and compare the precision numbers you get with the ones listed here. These precisions numbers are listed as a benchmark and you do not have to exactly match them. However, if your precision is significantly less than the precision reported, it may indicate a problem with the system.

1. Run each query in TIES and review each report returned.
2. Determine if the report was a correct match to the query. Use the following table to track this information.

	Query No.	Report No.	True Positive
	1	1	1
	1	2	1
Query 1 TOTAL			2

	2	1	0
	2	2	1
	2	3	1
Query 2 TOTAL			2

3. To calculate the precision for each query, divide the no. of true positive reports by the total no. of reports returned for that query.

	Complexity	Query	Precision
1	Low	Men, 60-80 with prostatic adenocarcinoma on prostatectomy	0.98
2	Low	Women, 30-50 with atypical endometrial hyperplasia	1.00
3	Low	Patients, 20-50 with phaeochromocytoma	0.98
4	Low	Patients with hemangiosarcoma of scalp	1.00
5	Low	Patients 10-30, with cystosarcoma phylloides	0.89
6	Low	Patients with superficial spreading melanoma, metastatic	1.00
7	Low	Patients with medullary carcinoma in thyroid gland	0.96
8	Low	Patients with adenocarcinoma in brain	1.00
9	Low	Men with invasive ductal carcinoma of breast	1.00
10	Low	Patients, >60 with Hodgkins disease	0.68
11	Moderate	Patients with prostatic hypertrophy and PIN on prostate biopsy	1.00

12	Moderate	Patients with either scar or radial scar, and intraductal papilloma on mastectomy or excisional biopsy of breast	1.00
13	Moderate	Patients, 40-60 with tubulovillous adenoma and adenocarcinoma in colon or rectum	0.98
14	Moderate	Patients with lung fibrosis secondary to systemic lupus	1.00
15	Moderate	Patients with adenomyosis on endocervical biopsy or hysterectomy	1.00
16	Moderate	Patients with prostatic adenocarcinoma, and PIN but no perineural invasion	0.97
17	Moderate	Patients with papillary carcinoma of thyroid in the setting of multinodular goiter	1.00
18	Moderate	Patients with osteosarcoma of femur or tibula showing tumor necrosis	0.78
19	Moderate	Patients with lobular carcinoma in situ and microcalcifications undergoing a procedure in which a sentinel lymph node was biopsied	0.96
20	Moderate	Patients 40-60 with cirrhosis or fibrosis and hepatocellular carcinoma on liver biopsy	0.86
21	High	Patients with sclerosing cholangitis on liver biopsy who have also had ulcerative colitis on another procedure	0.00
22	High	Women diagnosed with LCIS who had a subsequent mastectomy within 1 year	0.92
23	High	Patients with dysplastic nevi who were diagnosed with melanoma after an interval of at least 1 year	0.76
24	High	Patients with diagnosis GERD or Barrett’s esophagus who later had esophagectomy showing adenocarcinoma	1.00
25	High	Men with anaplastic astrocytoma who later developed glioblastoma multiforme	1.00
26	High	Patients with both schwannomas and meningiomas	1.00
27	High	Patients with tissue documented Berger’s disease who later underwent kidney transplantation	0.33

28	High	Patients with DFSP and a second procedure for local extension or recurrence within 3 months.	0.89
29	High	Patients with colonic adenocarcinoma who also have had Invasive ductal carcinoma of breast	0.91
30	High	Patients with renal carcinoma in kidney tissue who also have lung tissue with metastatic renal cell carcinoma	0.88

Common Queries

Because each institution that deploys TIES will have its own unique patient population, it can be difficult to use direct comparisons of results between institutions as a validation procedure. However, certain queries can be expected to return results with useful predictable patterns. The following test queries can be used as a face validity check to a fully operational TIES node.

Query	Expected Results
Charts Query for Concept = Neuroblastoma, in Pathology Report Final Diagnosis, Categorize by Age	Results should be almost entirely in the pediatric population, excluding olfactory neuroblastoma in adults
Charts Query for Concept = Endometrial Carcinoma, in Pathology Report Final Diagnosis, Categorize by Gender	Results should be exclusively in females. Note that it is not uncommon to see a small number of Male Genders which may represent erroneous gender coding in the source data. You may wish to check individual reports using an honest broker
Reports Query for Concept = Dermatitis , in Pathology Report Final Diagnosis	Results should include a huge variety of different types of dermatitis. Click on Search Terms to inspect documents. A number of these will not have the query concept dermatitis, but rather a sub type of dermatitis.
Reports Query for Concept = Melanoma, in Pathology Report Final Diagnosis, then use a Diagram Interface and add temporal query with additional pathology report showing melanoma at	Results should show multiple types of melanoma in initial results set. Second query should show a subset of results from query 1.

greater than 2 weeks.	
Reports Query with Diagram Interface for Concept = Breast Carcinoma in Pathology Report Final Diagnosis, add a temporal query with Concept = ovarian carcinoma in Pathology Report Final Diagnosis at any other time	Results should provide a small number of women with multiple neoplasms, presumably BRCA positive
Reports Query with Diagram Interface for Concept = Prostatic Adenocarcinoma and Concept = Perineural Invasion	Results should not include specimens with no evidence of PNI (these should be the majority) showing that the negation detection in TIES is working.

Institution Specific Queries

Finally, you should check the results of the TIES query against results previously generated from other systems at your institution.

1. Identify previously run queries and for which you have a list of valid reports that match that query. Generally try and choose queries that do have a relatively small result set.
2. Run the query in TIES and compare the results. Generate two lists
    a. Reports that are only in the TIES result set
    b. Reports that are only in the other system's result set.
3. For each report in the first list, confirm that the report is in fact an accurate match. If it is not a match, try to determine the reason why this report was returned, and attempt to construct a query that would eliminate this report from the result set.
4. For each report in the second list, confirm whether the report was in fact imported into TIES. If it is present in the TIES database, then confirm whether it was processed by all pipelines. You can do that by viewing the report in TIES and inspecting its annotations. You can fetch reports using their accession nos. or record IDs, from the Honest Broker perspective.