REPORT

# of the

RECORD CONSOLIDATION COMMITTEE

to the

BOARD OF DIRECTORS

of

NAACCR

Submitted March 1, 2000

EXECUTIVE SUMMARY

Based on the Committee's earlier report of a comparative test of record consolidation by states, the charge to the committee was enlarged to evaluate the feasibility of creating a larger more diverse test data sets for two purposes: 1) to measure consistency and accuracy of determining the number of patients and tumors, and 2) to measure the impact of different consolidation methods on incidence rates. The current report explores many theoretical and practical issues involved in meeting these charges and presents the committee's recommendations.

Regarding Charge 1, the committee concludes that the charge is feasible, and recommends that a test data set be created. The committee further recommends that the test be carried out to test consistency of counting tumors according to SEER multiple primary rules, but not to test patient linkage and consolidation at this time. The committee strongly recommends that NAACCR undertake this project, but feels there are significant barriers to be addressed and a substantial need for additional resources. A number of specific recommendations for carrying out the charge are presented.

Regarding Charge 2, the committee concludes that further study of its feasibility is needed, and the additional study would require significantly more time and additional technical expertise. Charge 2 should be reconsidered after the test described in Charge 1 is carried out.

The members of the Record Consolidation Committee are:

Jane Braun, MS,CTR Committee Co-Chair
Minnesota Cancer Surveillance System

Frances Ross, CTR, Committee Co-Chair
Kentucky Cancer Registry

Georgia Armenta Yee, BSW, CTR
Arizona Cancer Registry

Diane Kirsch, RHIA
Pennsylvania Cancer Registry

Deborah Bringman, MPH
University of California/Irvine

Betsy Kohler, MPH, CTR
New Jersey State Cancer Registry

Mary Jane King, CTR
Massachusetts Cancer Registry

Kathleen McKeen
State Health Registry of Iowa

Jennifer Seiffert, MLIS, CTR
TRW/CISSS Project and Registry Plus Software

Kathleen McDavid, PhD, MPH
Centers for Disease Control and Prevention

March 1, 2000

Record Consolidation Test: Feasibility Assessment

INTRODUCTION

Record consolidation is the process of combining data from two or more linked records for the same patient and tumor to produce a single "best" value for each patient and tumor variable. This can be automated, manual, or a combination. Record consolidation involves a series of processes that include editing, patient linkage, tumor linkage, determining the best information, and final editing for a consolidated record. Some central registries may first choose to edit source records in order to verify that the data in the source records are accurate and consistent. Some central registries may choose patient linkage as their first step in record consolidation. Patient linkage is the method to identify and group multiple records belonging to the same individual. Procedures to link data about the same primary tumor within the same patient are termed tumor linkage. The tumor-related information contained in all records belonging to an individual can then be summarized or consolidated based on rules developed by the registry. In determining the number of primary tumors (cancers), registries should adhere to national or international standards if registry data are to be comparable. To determine the final data values for the consolidated record, the registry must determine the level of specificity that is required to provide useful data for the purposes for which the registry exists.

The first report of the Record Consolidation Committee (July, 1999) demonstrated that there are a wide variety of approaches currently being used in central cancer registries to accomplish record consolidation. It concluded that some level of automation is considered necessary and cost effective for all central registries. However, the balance of manual versus automated processing must be weighed individually by each registry. The design of record consolidation processes must be adapted to fit within the context of the entire registry's operations. There is not one set of consolidation rules or procedures that will work for all central registries. Differences in data quality from reporting sources, level of sophistication of automated systems, availability of staff for manual review, and resources for registrar training will determine the best approach for each registry.

The results of a comparative study conducted by seven registries that have representatives on the committee showed that patient and tumor counts were consistent across all registries that use the SEER[*] rules for multiple primaries. However, this was a very limited test data set, and even though the counts of patients and tumors were consistent, many of the multiple primary rules were not rigorously tested. Also, there were many variations in the final data values selected for the consolidated records, even in this very limited test data set. The committee recommended a more thorough study of record consolidation procedures using a larger, more diverse test data set.

Thus, the new charges to this committee became:

[*] The Surveillance, Epidemiology and End Results Program Code Manual, Third Edition, National Cancer Institute, National Institutes of Health, Public Health Service, US Department of Health and Human Services.

1) to evaluate the feasibility of creating a large, diverse test data set to be used by central registries to measure the consistency and accuracy among registries in determining the number of unique patients and tumors, and

2) to evaluate the feasibility of creating a representative sample of records from a defined population in order to measure the impact of different consolidation methods on incidence rates.

I. ISSUES WITH CHARGE 1: Measure consistency and accuracy of counts

A. Theoretical Concerns

The initial discussions of the committee focused on interpreting and clarifying the charges outlined above. There are both theoretical and practical issues to be considered when evaluating the feasibility of these projects. The first step was to determine exactly what specific questions would be answered by conducting the consolidation test in central registries. The stated goal of the first charge is "to measure the consistency and accuracy in patient counts and tumor counts" among central registries using different consolidation procedures but using the same large, diverse test data set.

Measuring consistency of counts among central registries seems to be feasible, but measuring accuracy presumes that there will be correct counts for patients and tumors with which the results from different registries will be compared. How will the correct results be determined--for patient counts and for tumor counts? For patient counts, the correct number must be determined by those who construct the test data set. Since no real patient identifiers will be used, the test data will contain fictitious names, social security numbers, birth dates and addresses--the very items by which registries determine if records represent the same person or different persons. The test constructors can create exact matches, conditional matches, and non-matches, but there are no rules or real people to refer to in order to determine the actual number of different people. How would registries resolve the conditional matches? There could be additional information in text files, or maiden name or alias fields. Do we want to test whether central registries read and use these potential sources of additional information? Are existing patient linkage programs already sufficiently documented and tested, such that this aspect of record consolidation need not be part of this test? What about determining the correct number of tumors? For this, the SEER multiple rules are a standard that can be applied to the information contained in the test records to determine the number of primary tumors. However, the multiple primary rules are necessarily complex and occasionally ambiguous, so that an expert panel may be needed to work with the test constructors to determine the "correct" number of tumors in the test data set. Carrying out this aspect of the charge may be the most thorough and valuable study conducted so far by NAACCR to evaluate what the real differences are among central registries in counting the number of tumors by type of cancer.

Additional problems occur however, when determining tumor counts by specific variables. As mentioned earlier, there are no standards to determine the accuracy of the individual data values (i.e. race, sex) in the absence of real people; however, consistency of counts among different registries could be measured. Determining counts by specific tumor variables (i.e., behavior code, stage, primary site, etc.) would again require the reliance on an expert panel for correct answers.

Committee Recommendations: The committee recommends creating a large, diverse test data set to measure consistency among central registries in counting number of tumors by primary site. It is recommended that patient linkage not be part of this test at this time. The test should focus on testing each of the SEER multiple primary rules with a number of different sites - enough to adequately test each rule and its exceptions. The SEER EDITS routines may also be useful in designing the test data set. It is anticipated that the test file will contain 500-1000 unconsolidated source records.

B. SELECTION OF TEST CASES AND CREATION OF TEST DATA SET

The recommendation has been made to construct the test data set to test the SEER Multiple Primary Rules. The SEER Rules offer a universally accepted set of codified resolutions to multiple primary questions with which to test the consistency of tumor record consolidation.

1. Issues

<u>Selection of cancers to be included in the test set.</u> The number of scenarios (primary site and histology variations) that will be required to adequately test each rule needs to be determined. The number of permutations necessary to test each rule probably differs, based on the complexity of the rule and/or the prevalence of the neoplasm or primary site. For example, determining multiple primaries in paired organs where the only issue is laterality is a very common dilemma but not very complex and it should be decided how many cases of this type are necessary in the set. On the other hand, all multiple primary questions that involve different histologic types are more complex and it will have to be determined whether it is useful to test every rule and exception for every possible primary site and every possible histological type.

<u>Inclusion of text in individual records.</u> A decision to include text in the records of the test set depends on the purpose of the test. Text will allow registries to test linkage resolution in a simulated real life environment where text can be necessary to address multiple primary and laterality issues.

The exclusion of text would require registrars to assume that the coded values were correct. This probably would have the effect of making outcomes more predictable (meaning the count of tumors would be consistent) because decisions would have to be based on rules alone. Text is sometimes absolutely necessary to determine multiple primaries in certain situations.

If any text is included, a further decision must be made whether to omit text in some records to make the test set more realistic. In all instances where text is lacking, a determination will have to be made as to whether a specific rule will be provided and what that rule will be.

Unknown values. A decision will have to be made whether to allow unknown values in fields. Unknown values for which there are no agreed defaults will make tumor consolidation decisions more difficult. Unknown values are particularly problematic in date fields. Different software packages will define date fields differently.

File format. The format for the file containing the test data must be determined. Various central registries will conduct the test. A consideration in the choice of format is what type would be usable by the greatest number of registries. It might be helpful to first determine which registries are going to conduct the test and which NAACCR version most of them are using, for example, NAACCR 6.0.

2. Barriers

Challenges to the creation of a test set include providing sufficient and appropriate records, deciding on the content of these records and the format of the data file.

Sources. Source records must be found for the test data set. Many central registries have not retained source records in their original form. They may have been edited or otherwise modified during consolidation so that the original discrepancies in data values are no longer apparent.

A decision will have to be made as to whether the records will be real source records from central registries or whether they will be fabricated or composite records. If the decision is made to use real records, certain difficulties must be overcome.

Choice of specific records for the test set. The choice of specific records for the test set will depend on the committee's decision as to the number and types of tumor consolidation problems to be tested. For example, records must either be found or created to test specific multiple primary rules. In terms of time, it might be more expedient to create test records rather than attempt to find original cases, especially for rules involving less common problems or less common sites

Text handling. Many registries have not received or retained text in electronic form. It may be difficult or impossible to re-generate a facsimile of the original source record with accompanying text in NAACCR format.

The text may be too long to fit into the allotted space in NAACCR format. The NAACCR format limits the amount of text space that is available. There is a question whether the allowed space is too restrictive when extensive documentation is required. Reporting facilities have solved this problem in some cases by taking advantage of extra text fields available in some commercial cancer registry database programs. Unfortunately, since this extra text is not part of the NAACCR record, it is not available to central registries. Additional text fields in the test would require altering NAACCR format. This would make the file less straightforward to use. It should be determined whether the NAACCR text fields would be sufficient for documentation of multiple primary rules since other types of documentation (e.g. smoking, social history, family history) would not be necessary. Furthermore, the rigorous use of abbreviations might allow even very complicated documentation to be fitted into the allowed space.

Desensitization of data. All genuine original records would have to be desensitized, meaning that all personal identifiers would have to be removed or rendered fictitious. Depending on local conditions, some central registries may have restrictions on sharing even desensitized cancer data. This may inhibit some central registries from providing data for the test.

Standardization of control codes. The test data set file may come from several original sources. It will need to be standardized as to whether it contains carriage return/line feed or other characters, "newline" or other indicators, and will need to be consistent throughout. These specifications will have to be documented and, if necessary, provided to registries testing the set.

3. Possible Resources

Cancer information. Tumor records may be available from the following sources: The Iowa Cancer Registry test set from first round of consolidation testing, the Minnesota Cancer Surveillance System record set, and the Massachusetts Cancer Registry electronic submissions from reporting facilities (stored on diskette in their original, unmodified form).

Supporting text. Records containing text documentation may be available in some cases. The Iowa test records all contain text and some of the records in the Minnesota database have text. Massachusetts's records also contain some text.

Mechanism for editing file. The Kentucky Cancer Registry has a program available to read NAACCR 6.0 formatted records and edit individual fields.

4. Committee Recommendations

Source records. Records selected from a pool of real source records will serve for the majority of test cases. The Committee recommends that they pass a specified group of NAACCR edits but are not modified by consolidation. Some additional cases may have to be fabricated to present rare scenarios.

Text. The majority of records should include text. Some records should contain no text. No space beyond the standard NAACCR layout will be allotted for text.

Simulated followback. Additional information should be created and made available on the NAACCR web site to simulate follow back to reporting sources.

Desensitization of data. Records used in the test data set will be stripped of personal identifiers. The same ID numbers will be assigned to records belonging to the same person. Records for the same person will also be assigned the same fictitious name, social security number and birth date.

Unknown values. There will be no blanks in required fields. Unknown values should be "9." There should be a mixture of unknown and known values. Each registry will be asked to follow its own procedure concerning unknown values when conducting the test.

Format. The test data set will be available in the most widely used NAACCR format(s) for which edits exit. The file should be clearly described.

Resources. The Committee recommends that NAACCR pursue the necessary resources for record creation, text, and file editing listed above in section B3. This may include money, contracted work, or staff support.

C. CREATION OF "CORRECT" RESULTS

1. Issues

Critical data items. The data items for which a "best answer" is needed in order to evaluate the effects of varying consolidation procedures on frequency counts must be determined. Items such as race, sex, date of birth/age at diagnosis, and zip/postal code are not going to be considered at this time. They will be addressed in Section III regarding Charge 2. Critical for calculating frequency by type of cancer are the final values for primary site, histology, laterality, behavior code and date of diagnosis.

Determination of the correct results. Since the SEER rules are the rules NAACCR has adopted, SEER would presumably set the gold standard for determining the correct number of cases by site of cancer. Expert panel members would need to be selected to represent the correct application of SEER rules.

Differences between SEER and CoC rules for multiple primaries. Central registries running the test may come up with different numbers of tumors for certain patients, based upon whether their systems use the SEER or the American College of Surgeons' Commission on Cancer (CoC) rules for determining multiple primaries.

Registries that use IARC rules. At least one Canadian registry uses the International Agency for Research on Cancer (IARC) rules for record consolidation. As envisioned thus far, the proposed test data set would be designed to measure consistency of applying SEER multiple primary rules. In order to measure consistency of applying IARC rules, a separate test would have to be developed.

Specificity of automated determinations. Best information can be determined through a central registry's automated system, through manual adjudication, or a combination. There are two balance points to consider: accuracy vs. specificity, and automation vs. staff review. [These concepts are discussed in detail in the Committee's report Central Cancer Registry Record Consolidation: Principles and Practice, pp. 8-10.] The issue to be determined is what level of specificity is necessary for the "best information" for the intended level of data use. For charge 1, the expert panel must define what range of values constitutes a correct result for the data fields: primary site, histology, behavior, laterality and date of diagnosis. In charge 2 of the consolidation test, each consolidated record must be specific enough to calculate incidence rates. Therefore, in charge 2, correct results may have to be determined for race, sex, birth date, and residence codes as well as the tumor characteristic variables.

2. Barriers

Subjective or ambiguous "correct" answers. Because there are no existing standards for consolidation, it is difficult to determine what the "correct" answer should be for all consolidation issues. Even when there are standards, such as for the SEER multiple primary rules, interpretation can be difficult, and experts can differ in coding individual records. There may be no authoritative correct answer that everyone can agree on, even in principle. Rules in place in various central registries differ significantly--for example, one feels that the first record received was abstracted closer to the diagnosis date and more likely contains correct information, whereas another feels that the most-recent record is more likely to have been corrected over time, and should be given greater consideration. However, experience with other data quality studies has shown that for most cases, it is likely that a consensus answer can be agreed upon.

Selection of "experts". Since there is ambiguity and subjectivity involved, the experts selected might bias the outcome. It will be important that the experts selected have the authority to represent the organizations that have set the standards being used, specifically SEER. Since the test data set will be larger, the experts selected would have a larger workload than with the first test.

Need for followback. There is no definitive way to determine the correct values for several of the key data items without reviewing cancer registry databases or medical records at reporting facilities. Issues of time, staffing, access to records, etc., often make individual review prohibitive; central registries must decide which data items are so critical that they warrant resources to perform medical record review. Varying resources and priorities can also result in differences in the level of followback within registries from year to year. In addition, it is much more difficult to ensure the consistency of manual adjudication compared to that achieved through automation.

Lack of text. Many records contain no or incomplete text, so the additional information necessary to resolve discrepancies is often not available.

3. Possible Resources

Expert panel. One option might be to use the experts that served for the first test, i.e., Jennifer Seiffert, and Dr. John Young. Another option might be to ask SEER to designate appropriate experts. It would also be helpful to include experts designated by COC and perhaps by the NAACCR Uniform Data Standards Committee.

Standard rules. Standard consolidation rules could be developed to assist registries in choosing the “best information” for data items for which there are no existing rules. These new standards could include a hierarchy for determining which value to choose. For example, records including text could be ranked for probability of correctness based on class of case; if no text is included, the most-common answer could be considered "best"; a choice could be made between equally-common values based upon which was most recently submitted, etc.

These rules would need to be developed by experienced staff knowledgeable about record sources, common errors, and coding policies. This level of expertise is available among committee members, staff from other central registries, staff from standard-setting organizations, or from a contractor.

Simulated followback. It would also be possible to simulate reviewing a hospital medical record or physician's office chart by setting up a repository of text to clarify conflicting information, such as on the NAACCR website. Registries attempting to consolidate the test data set could query a special area of the website to obtain additional information on specific tumors in situations where manual adjudication is necessary and insufficient information exists in the source record. This site could be set up to keep track of the number of inquiries for each tumor, to monitor the number of registries where the automated system is unable to determine a "best" value.

Setting up this source of additional information would require staff familiar with cancer etiology and writing cancer registry text so that registries would be provided with realistic information that would be compatible with their rules for manual adjudication. This level of expertise is available from the same sources as those listed above.

4. Committee Recommendations

Expert panel. We recommend using an expert panel as listed above. The SEER program should either designate the experts or agree to those selected. The consensus of the expert
panel will prevail.

Data items. Determination of "best information" is possible for all tumor related data items. It is feasible and recommended only for those items which affect multiple primary determination (primary site, histology, laterality, behavior, and date of diagnosis).

Standard Rules. SEER rules will prevail in determining the number of primaries. In ambiguous situations, the panel will need to develop written rules and algorithms for manual adjudication of discrepant items. No stipulation will be made to test the consistency of applying the IARC rules. Discrepancies resulting from differences between SEER and CoC rules or their interpretation will need to be described in the study results.

Text/Simulated followback. Additional information, in the form of fabricated text, must be provided to registries participating in the test. This test should either be part of the source records or be available on a website.

D. Running The Test In A Central Registry

1. Issues

Test instructions. Building a test file to run against a wide variety of systems will require assumptions be made, for example that the user can accept all data items in a specific version of the NAACCR format. This in turn will pose challenges and obstacles for those using the test file. Instructions will be needed to assist central registry personnel in deciding if running the test is feasible within their system and in identifying modifications needed to complete the test. The instructions must address, at a minimum:

- Purpose of the test,
- File format including NAACCR version and data item specifications,
- Discussion of potential barriers, and
- Resources for assistance

Preventing registry file corruption. A mechanism enabling participation in the test without corruption of central registry programs and live database(s) must be addressed. Running the test will require the establishment of a parallel database within the system, or the ability to somehow back out of the linkage and consolidation process. An option for unique identification of test records may be considered enabling the records to be removed upon completion of the test. Registry staff must be careful that test records do not get inadvertently incorporated into the live database. For many central registries this is a difficult, if not impossible, task to achieve.

Editing. Editing of records is a preparatory step to consolidation for most registries. Ideally the registry, assuming the test set consists of previously edited records, could import records into consolidation programs and forgo editing. However, the design of some systems may require records to pass through edits before reaching consolidation. Additionally the application of edits may be necessary to ensure specific data items, critical to the registry's automated tumor and data consolidation rules, meet registry definitions. Once records are consolidated edits are often applied to assure the consolidation process did not create irregularities in the data. The file format description must clearly identify data items that may present obstacles. It may be necessary for particular edits to be bypassed and, this too, may be a difficult or time-consuming task.

Text. Additional information may be required to accurately complete consolidation procedures and can be made available in NAACCR text fields. The storage of and access to text, however, varies among registries. While some systems have not yet incorporated text into the central database, providing instead a separate storage and retrieval mechanism, other systems have gone a step beyond the standard text fields incorporating additional supplemental text in the record format. Will the text available in standard NAACCR fields be adequate for decision making? Will there be a need for supplemental text in the record and will registries be able to import and access such f fields?

2. Barriers

Though the diversity of systems in use bar a discussion of all obstacles an individual registry may face, a number of obvious barriers to running the test within a central registry are further addressed in the following section. Each of these issues and the potential barriers they may present must be addressed within the test instructions.

Preventing registry file corruption. Test records must not be allowed to enter the registry's active data files without a pre-established mechanism for their identification and removal at the completion of the test. Ideally an alternate database could be established to serve during the test run. While this point should be made in the test instructions, each registry would need to determine if this is a viable alternative. The wide variety of systems in use foregoes the documentation of a model for this process. Recognizing that the establishment of a secondary database may be reason alone to discourage participation in the test other alternatives should be sought. Perhaps test records could be flagged in such an obvious way to allow for identification and deletion after completion of the test. One suggestion may be to insert 'NAACCR' in the medical record number field of all test records. This mechanism assumes a field to be critical in all systems. Should this approach be chosen the ability to mass delete the cases would be needed for such a large number of records. Although both approaches will serve to assure the registry's live database is not corrupted, ultimately the registry must decide which fits their system best. Since it must be the responsibility of the registry to create the alternate database, determining the data field to flag may also be best left to the registry.

File format. Central registries differ in their ability to handle files dependent upon the NAACCR version and the presence of specific characters such as the cr/lf (carriage return/line feed). These factors can affect the ease with which the test can be conducted. In an effort to concentrate on enhancements and system upgrades many registries have chosen to delay conversion to more recent versions of the NAACCR file format. If the test file is to be available in the latest NAACCR format provisions for conversion to earlier versions will increase the number of registries able to participate in the test. Likewise the cr/lf is needed by many systems to recognize the end of individual records and the absence of the character will render the test impossible in these systems.

Editing. Editing both prior to and after consolidation is a crucial step in registry procedures. Many central registries may not be able to handle variances in the test data when similar conditions are not allowed in live data. For example many central registries may require that all or parts of dates be known. Other systems may not permit the processing of records if fields are missing, fields which may not be present in the test data set. Any variance used to create the test data can cause problems. Furthermore the test cases are likely to be non-residents of the area covered by the central registry performing the test. Thus, the test may not be comparable to the central registry's usual routine of processing. Invalid or missing variables such as county or postal code may cause difficulties. In systems with edits hard-coded in consolidation programs it will be a time consuming effort to identify edits which must be temporarily shut off for the test and this very likely will discourage some registries from conducting the test.

Staffing. Even if the registry devises a safe mechanism for processing the test file there may not be staff available to conduct the test. The test will include a large number of records and this number alone could discourage participation in registries already struggling to conduct normal business with limited staff. The degree of automation in a registry's consolidation process may too influence the decision to run the test. The more manual the consolidation process the more time the test will take unless the registry is well staffed.

Text. Central registries differ in procedures for handling text, an essential element for accurate consolidation. Some registries feel that more text than is provided for in the NAACCR record is required for adequate consolidation. There are no standards for handling supplemental text and few, if any, registries will have procedures in place for storing or accessing such text. Analysis of text too presents its problems. To optimize text fields abbreviations are often used and the interpretation of abbreviations may vary by geographic region of the country or even from state to state. When text is inadequate registries contact staff from the reporting facility for additional information. This resource will not be available in the test. Of course there are times when additional information is not available and a condition of the test may be that only text within the record is to be used. However, a file on the NAACCR web site that could be accessed during the test may represent facility follow-up.

3. Possible Resources

Contracted assistance. A contract firm to assist in technical aspects of running the test should be considered. Since the uniqueness of the registry's systems may require familiarity available only with internal staffing, cost/benefits may preclude this option. Committee members who conducted the previous test may be able to provide consultation, however some mechanism for reimbursement of time will be needed as this could become a time-consuming commitment.

Available software. Data editing issues may be addressed by using Genedits software with which edits maybe turned on or off with ease. A metafile with an edit set suited to the test records could be made available for download. During the committee's previous test a program was developed by the Kentucky Cancer Registry to allow the user to read NAACCR records and make changes to an individual field. This program may be made available to assist others in modifying test record fields, such as dates, so the data meets specifications to run through consolidation programs.

Web site. A file on the NAACCR web site to act as facility follow-up could provide supplemental information where text is inadequate. Additionally several committee members have offered the use of standard abbreviation lists already in use in their central registries. Choosing one list or compiling the lists would provide for consistency in text interpretation.

4. Committee Recommendations

Test instructions. Detailed instructions will be needed to assist central registry personnel in deciding to run the test and identifying modifications needed to complete it. These instructions should be accessible separate from the test file so they may be thoroughly reviewed before a decision is made to download the test file. The instructions must address the issues previously discussed in this section of the report.

Text. Text included in the records should be accommodated in the text fields of the existing NAACCR format. If text in the test records is skillfully abstracted and abbreviations are used the most important points can be included. There should be a simulation for contacting reporting facilities for additional information, such as a web site that contains detail on records with questions that cannot be resolved by the data and text alone. As in real situations, some records should have no additional information on this site.

Consultant assistance. Some provision should be made by NAACCR to make available a consultant to help central registries who want to run the test.

E. Analyzing the Data and Reporting Results

10 Issues

"Gold Standard" data file for comparison. Before analysis can begin, a data file must first be created that contains the correct number of patients, the correct number of tumors and the "best information" for consolidated data items. The constructors of the initial test data file that contains the source records used in the study must also create the data file that contains the results after linkage and consolidation of the source records. This "gold standard" data file should contain one record per tumor and include the correct number of patients, tumors and consolidated values as determined by an expert panel. Records should be written in the current version of the standard NAACCR record layout and be in ASCII format with trailing carriage return and line feed (CR/LF). Minimum data requirements include those data items in the original test data file excluding text variables. Issues related to the construction of the "gold standard" data file have been discussed previously.

Output File format. Registries that participate in the consolidation study will be asked to process the input test data file containing source records and by using linkage and consolidation procedures that are routine for their registry to create a resulting output data file that contains one consolidated record per tumor. The output data file must adhere to NAACCR standard definitions and codes. It should be written in ASCII format with trailing carriage return and line feed (CR/LF) and comply with the current version of the standard NAACCR record layout.

Reporting format and feedback. To analyze the results of the consolidation study, each participating registry's resulting data file will be compared to the "gold standard" data file and discrepancies noted. Results will be summarized and reported to study investigators and registry participants. The identity of the registries will be excluded from summary data. However, due to the substantial amount of time that is invested in conducting the study, it is important to provide confidential feedback to each participating registry outlining in detail their results. This may require the expertise of an expert panel member to explain the differences found in the comparative analysis.

Assignment of records to patients/tumors. The first objective of the study is to measure the consistency and accuracy among registries in determining the number of unique patients and tumors. It is inadequate to simply tabulate the number of unique patients and tumors in each registry's data file. Two registries may have the same results but arrive at them differently. A registry may correctly link two source records to the same patient, but incorrectly linked them to the same tumor or consider two source records to represent two unique patients when they should represent the same patient. To measure consistency and accuracy, one must also consider the source records associated with each tumor in the resulting data file. To accomplish this, the original test data file must include a unique identification number for each source record. The "gold standard" data file must also include the identification numbers of the source records that linked to the resulting tumor. Space will be allocated in the record layout for inclusions of these numbers. Registries participating in the study must be able to identify the source records associated with each tumor in their resulting data file. This can be done by including the identification numbers of the source records for each tumor in the output file or by producing a second file containing the patient and tumor identifiers and their associated source record identification numbers.

Patient and tumor counts will be tabulated by site groupings and discrepancies between the "gold standard" and the registry data files will be noted. Reports will be produced for individual registries depicting records where false positive or negative patient and/or tumor linkage existed. Comments will be provided by expert panel members that explain the reasons for the differences. For example, if the expert panel considered two source records for an individual patient to represent two unique tumors and the registry considered them to be one tumor, the SEER multiple primary rule that underlies the need for two separate primaries will be delineated in the confidential report.

2. Barriers

Handling input file. Ideally, registries participating in the study should be able to produce an ASCII data file in standard NAACCR format that includes all tumors that resulted from processing the initial test data file. However as discussed previously, there are many barriers that may impede a registry from accomplishing this. Registries may be able to link the test data file to identify unique patients and their corresponding source records, but may be reluctant or incapable of processing the records further for fear of corrupting their live database. Such registries may want to proceed by manually linking and consolidating the resulting tumors. Results would be provided in hardcopy format and would have to be manually entered into a database for analysis. This would be a time consuming task for both the participating registry and the data analysts and may not truly reflect the linkage and consolidation procedures for that registry.

Labeling output file. Registries that are able to process the test data file completely within their computer system and produce an ASCII output file in standard NAACCR format may not be able to include the identification numbers of the source records associated with each tumor in the output file. Additional programming resources may be needed to produce a second file containing patient and tumor identifiers and their associated source record identification numbers. Some registries may find it impossible to retain the source record identification numbers for each tumor.

Some registries may not be able to produce an output file in the most current standard NAACCR record layout. If different record versions are acceptable then a conversion program must be available to convert the data file to the most current record version.

Technical support. Computing resources are necessary for analysis. A database management system should be used to store the data from the data files. A statistical software package should be used to analyze the data. Statistical expertise is needed to provide meaningful interpretation of the resulting data. Data entry personnel are needed for those registries that submit hardcopy results. Computer programming expertise is necessary to produce customized reports for feedback.

3. Possible Resources

Feedback. The expertise of expert panel members is needed to provide registries with meaningful feedback on their results from the comparative analysis.

4. Committee Recommendations

Gold standard data file. The "gold standard" data file should be constructed by a panel of experts with experience in record consolidation.

File format. Registries participating in the study should provide a computerized data file of their results in NAACCR record layout version 6 or higher. Hard copy reports instead of computerized data files will also be accepted.

Confidential feedback. Reports should be produced for individual registries detailing their results from the comparative analysis. Discrepancies between the "gold standard" and the registry data should be noted. Records with false positive or negative tumor linkage should be identified. Meaningful feedback by expert panel members should be provided.

Funding for contractor. Due to the substantial amount of work involved in analyzing the data and reporting results, the Committee recommends that funding be provided to a contractor to facilitate the development of the "gold standard" data file, to provide the means to input data from hard copy reports, to analyze the data and report results, and to develop customized reports for feedback to participating registries.

II. ISSUES WITH CHARGE 2: Measure impact of Consolidation on Incidence Rates

A. THEORETICAL ISSUES

The second charge involves evaluating the feasibility of measuring the impact of different record consolidation procedures on incidence rates. In order to do this, the test records would have to contain all the tumor reports from a defined population, or perhaps a representative sample from a defined population. This set of records may be difficult or impossible to find. If it exists, it would have to be modified in order to conceal the identities of the real people for whom these cancers were reported. Once again, this may involve creating fictitious patient identifiers for variables critical to patient linkage and thereby possibly influencing the calculation of rates by race and sex.

Measuring the impact of consolidation on incidence rates would involve a test of patient linkage procedures. If the patient characteristics by which the rates are to be calculated (i.e., race, sex, county of residence) could be real rather than fictitious and could have a "correct answer," then it would be feasible to measure the consistency of rates for these variables by such data items as total tumor counts or counts by primary site.

B. PRACTICAL ISSUES

1. Issues

Minimum data set. The minimum data requirements include those data items in the NAACCR Call for Data for CINA publication. These include race, sex, age, date of birth, sequence number, date of diagnosis, site, laterality, histology, behavior, grade, diagnostic confirmation, type of reporting source, summary stage, state or province, and county of residence at diagnosis, and Spanish/Hispanic origin.

Case counts by sex, age, race/ethnicity, site, year of diagnosis, state or province, and county of residence and stage will be tabulated. Incidence rates will be computed and the variability in the rates among the participating registries will by analyzed to measure the impact that different consolidation methods have on reporting incidence rates. The appropriate statistical methods that are used for measuring these differences are yet to be identified.

2. Barriers

Statistical validity. Statistical expertise is needed to conduct this test appropriately and to analyze the results adequately.

Demographic consistency. Demographic data must pass standard edits, but would contain fictional patient names and have real patient values on the patient linkage variables.

Critical data items. The data items for which a "best answer" is needed in order to evaluate the effects of varying consolidation procedures on incidence rates must be determined. Items such as race, sex, date of birth/age at diagnosis, and zip/postal code are critical for calculating rates, whereas items such as name and social security number are more important for patient linkage and have no effect on rates.

Missing data items. Some registries may not collect the minimum data set that is required for analysis. Data items such as race, Spanish/Hispanic origin, full valid dates, or summary stage may be incomplete or non-existent. Special consideration would have to be given to those registries lacking minimum data requirements when analyzing the data.

3. Possible Resources:

Demographic information may be available from the following sources:
Iowa Cancer Registry test set.
Demographics from New Jersey State Cancer Registry demonstration data sets could be manipulated to add to Minnesota cases.
The results of charge 1 may assist in conducting this second record consolidation test.

4. Committee recommendations:

More study. A more thorough evaluation of the feasibility of the second charge is needed and would require a significant amount of time and technical expertise. Since it may also depend on the results encountered in carrying out the first charge the second charge should be reconsidered after the first charge is carried out.

Statistical expertise. The committee agrees that some sophisticated statistical expertise would definitely be needed to adequately address the issues involved in creating an appropriate sample of records from a defined population. It would also be needed to appropriately assess the impact of different consolidation procedures on incidence rates.

Alternate approach. The Monte Carlo method of randomly creating changes in real data was suggested as another means of measuring the effect of different consolidation procedures on rates.