SEARCH 4 MOP
Section 7 - Case Registration
Table of Contents

7. Case Registration and Un-registration

7.1. OVERVIEW

Aim 1 of the SEARCH 4 Registry Study is to obtain estimates of the incidence of diabetes in subjects less than 20 years of age in defined populations. In order to achieve this aim, SEARCH centers will try to identify and obtain information about all persons who meet eligibility criteria for specific geographic or health plan based populations. A summary of the full eligibility criteria to be counted in the numerator for SEARCH include:

- *Diagnosed with any type of diabetes mellitus except gestational*
  - Between 1/1/2016 and 12/31/2020 for incident cases
- *Less than 20 years of age*
  - Age less than 20 years on December 31 of the onset year for incident cases
- *Belonging to the center-specific defined population anytime during the index year (the index year for incident cases is the year of diagnosis)*
  - Resident of defined geographic region for geographic based centers
  - Is a member of the defined eligible health plan for membership-based center

Ineligible if:

- *Active duty military*
  - During the year of diagnosis for incident cases
- *Institutionalized (defined by the Census)*
  - During the year of diagnosis for incident cases
- *Diabetes type is gestational diabetes (only)*

7.2. CASE REGISTRATION

7.2.1. *Case Registration Windows*

Sites will use a 30 month ascertainment window from December 31st of each incident year for incident years 2013- 2017. Beginning with incident year 2018, registration will be closed 20 months after the end of the incident year.

Similarly, all data entry for each incident cohort will close as of the end of the designated registration window - and there will be no further new data entry (duplicates, un-registration, and late registered cases) after the close of the window.

Based on the window following December 31st of the index year, the following current dates for in-window ascertainment are:

Incident Year	Close of Window
2013	June 30, 2016
2014	June 30, 2017
2015	June 30, 2018
2016	June 30, 2019
2017	June 30, 2020
2018	August 31, 2020
2019	August 31, 2021*
2020	August 31, 2022*

*Pending additional funding

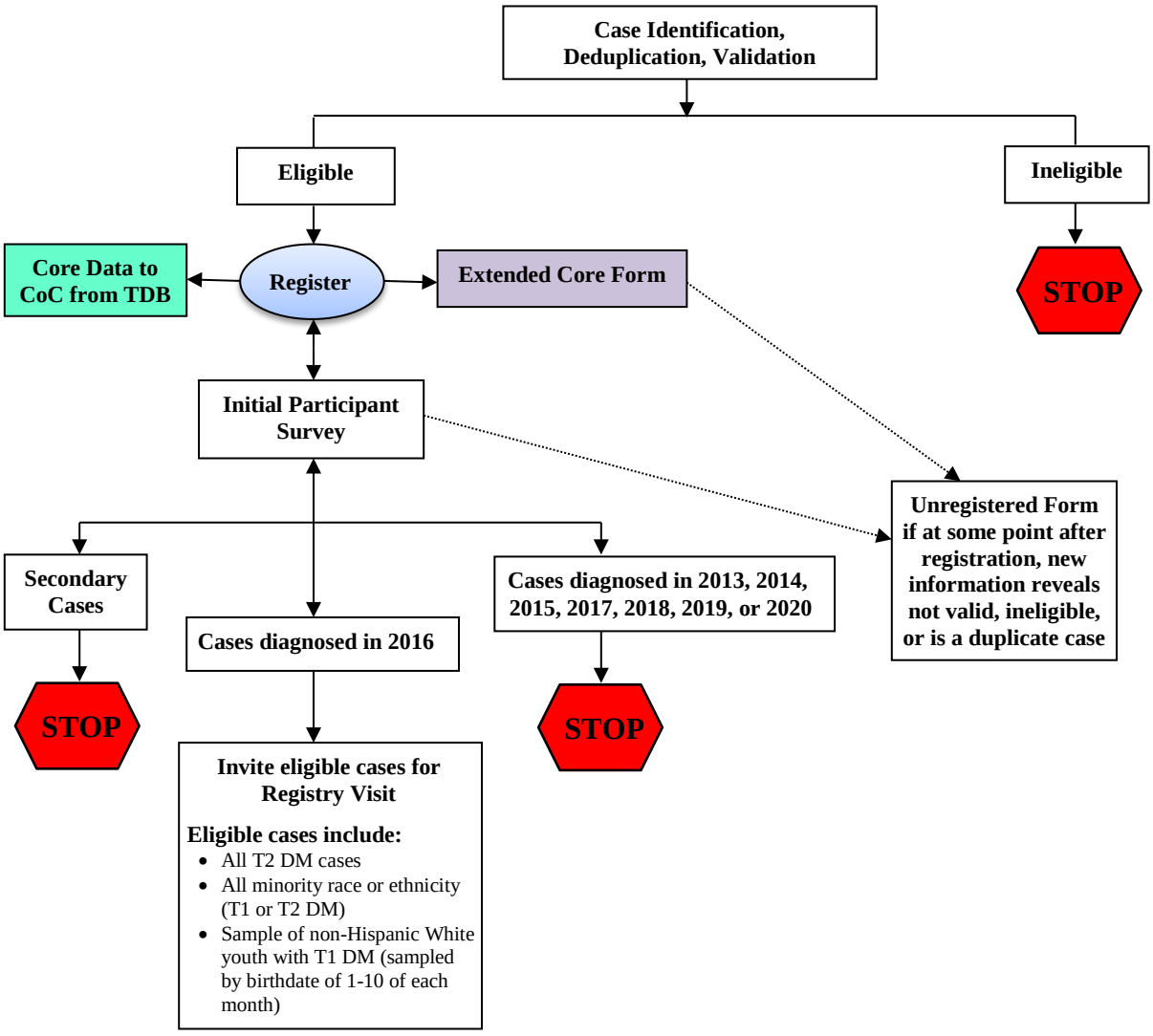
NO LATE CASE REGISTRATION WILL OCCUR AFTER THE CLOSE OF EACH INCIDENT YEAR'S WINDOW.

7.2.2. *Case Registration Criteria*

Some validated cases will have one or more eligibility criteria unknown, e.g., residence or military status in the index year. Limiting registration to validated cases for which residence in the geographic population of SEARCH can be confirmed, may selectively underestimate incidence. If allowed by the local Institutional Review Board, completion of the Initial Participant Survey (IPS) may be useful in establishing eligibility prior to registration.

To be eligible for registration, a person must have physician-diagnosed diabetes, be age eligible, not ineligible, and not known to be a duplicate. See Figure 1 for Case Registration Flow. See Appendix B for ICD-9 and ICD-10 codes for Case Ascertainment.

Figure 1 - Case Registration Flow



7.2.2.1. Valid case

A case can be considered valid when there is information sufficient to believe the person has been diagnosed with diabetes by a physician. This determination can be made by provider report, self-report, or medical record review.

7.2.2.2. Age eligible

Meets age eligibility criteria if:

- Date of birth (DOB) and case eligibility status (year of diagnosis) are known:

  - DOB from 1/1/1994 to 12/31/2013 for 2013 incident cases (e.g., 1/1/1995 - 12/31/2014 for 2014 incident cases, 1/1/1996 - 12/31/2015 for 2015 incident cases, etc.)

- DOB is known but case eligibility status is unknown:
  - Cases identified in 2013 will be age eligible if the DOB is on or after 1/1/1994 (e.g., identified in 2014 and DOB on or after 1/1/1995). Case eligibility status must be determined, however, before the case can be registered.
  - Age eligibility for subsequent incident years will be based on age eligibility criteria as outlined once the case eligibility status has been established.
- DOB partially known:
  - If the year of birth but not day or month is known, information is sufficient to classify eligibility based on DOB criteria listed above.
  - If only the age in years on a certain date is known, information *may* be sufficient to classify participant as eligible or ineligible.

Examples (2010 Incident Cases)

- If it is known a Participant was 12 years old in 2005,
  - The Participant is then known to be less than 20 years old on 12/31/2010 and therefore ***eligible*** in all circumstances.
- If it is known a Participant attended a camp for 12-16 year olds in 2010,
  - Sufficient information is known to classify them as ***eligible***.
- If it is known a Participant was 19 years old in 2009,
  - Sufficient information is available to classify them as ***ineligible***.

7.2.2.3. Not ineligible

A case is considered not ineligible if no information is available indicating the participant is ineligible based on geography, health plan, institutionalization, military or gestational diabetes eligibility/ineligibility requirement.

Note: There are no requirements to obtain information on institutional status. If information is not readily available from existing databases or the medical record indicating the participant is ineligible, the participant should be coded as eligible.

7.2.2.4. Duplication

Data available for duplicate checking and the procedures and patterns used for duplicate checking are center specific, but may include use of any of the following data: full name, initials, date of birth, zip code of residence, or date of diagnosis. It may be impossible to be certain that the case is NOT a duplicate. Registration should occur when you have concluded that reasonable efforts have been made to identify duplicates.

7.2.2.5. Issues Pertaining to Cases with "Cured" Diabetes

The SEARCH Steering Committee has agreed that for purposes of the SEARCH Study, "once a case, always a case." A participant may report a past history of diabetes but no current diabetes. For example, an incident participant, enrolled in SEARCH in 2012, reported that he had bariatric surgery. During his registry study visit after the surgery the participant reported the diabetes was "cured" (because his blood sugars had returned to "normal"). In this example, the participant definitely had diabetes in 2012 and was registered as a valid, incident case. For SEARCH purposes, this participant will always remain a case and should be invited for a study visit, if they meet eligibility for a study visit according to the protocol (e.g., diagnosis in the incident year for which registry study visits are being completed).

7.2.3. *Case Registration Procedure*

Once an individual has been determined to be eligible for registration, the participant's record in the tracking database system (TDBS) should be flagged as registered. On a regular basis, each center will upload new data to the Coordinating Center. The following is a list of information that, if available, will be uploaded from a center's local TDBS to the Coordinating Center:

- ID
- Age
- DOB
- Gender
- Race/ethnicity
- County of residence
- Zip Code
- Diabetes validated (yes)
- Method of validation (Medical record review/Direct verification by a physician/Clinically verified database/Death certificate/Self-report)

- Presumed Diabetes Type
- Case Status (Incident Year)
- Date of Diagnosis
- Secondary Diabetes [diabetes caused by another source e.g., illness or medication] (Yes/No)
- Residence Eligibility (Eligible/Pending/Not Applicable)
- Health Plan Eligibility (Eligible/Pending/Not Applicable)
- Military Eligibility (Eligible, Unknown)
- Institutional Eligibility (Eligible, Unknown)
- Date registered/unregistered

NOTE: All identifying information remains solely within the local TDB of the five participating centers. Only the above list of minimal information about the participant will be forwarded to the Coordinating Center. This will allow estimation of incidence rates by diabetes type, age, gender, race/ethnicity, and will assist with evaluation of typology. To protect participant health information in accordance with HIPAA guidelines, a Limited Data Use Agreement between each HIPAA-covered center and the Coordinating Center will be in place prior to the uploading of any data to the Coordinating Center. Centers that are not a HIPAA-covered entity may not require a data use agreement.

In the future, if more complete or accurate information becomes available that confirms the participant to be ineligible, the participant will be unregistered (Section 7.3).

The Coordinating Center will work with each center to provide methods for ID generation, registering participants, and procedures for uploading registration information.

7.2.4. *Case Registration Using ACCESS TDBS*

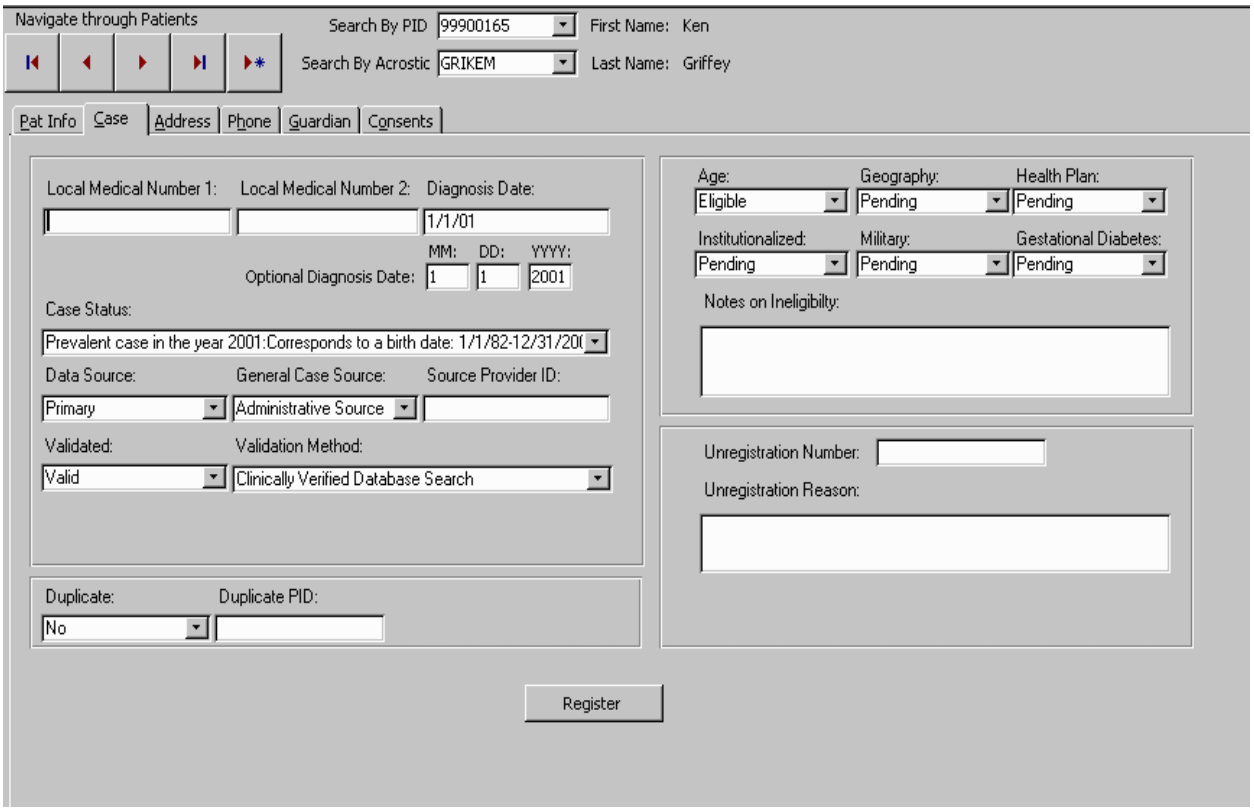
When a unique (unduplicated) case is validated, eligible, and NOT ineligible based on current information, the case should be registered. When the local center receives and enters adequate information in the TDBS, that center will register the case by clicking the Register button on the Case information screen on the TDBS. This screen is opened from the Patient menu and the Case Tab.

All data required for registration is entered and stored in the TDBS. The process and source of information used to enter data on potential participants is center specific. These data may be gathered locally from secondary data sources (e.g., database searches, clinician referrals, chart reviews, death certificate reviews) and primary data sources (e.g., participant self-referral, participant and/or parent/guardian survey, and/or the Initial

Participant Survey [IPS]). Primary and secondary source definitions may differ by center. Some centers may choose to administer the IPS prior to registration as part of their case ascertainment efforts.

Information regarding the TDBS is described in Section 4. Figure 2 illustrates the CASE TAB of the TDBS.

Figure 2 - Case Tab of the TDBS



When the Register Button is 'clicked,' the TDBS will evaluate if the information entered meets the eligibility criteria and provide a message indicating 1) registration is completed or 2) provide information regarding information that makes the participant not eligible to be registered. If the case is registered, the 'Register' button will be replaced by an un-register button.

On a regular basis, each center will upload new registration and un-registration data to the Coordinating Center using the TOOLS menu and clicking on the export routine. The new data is then made available to the Coordinating Center via the data upload utilities on the SEARCH website. These utilities are provided based on security rights assigned.

7.3. CASE UN-REGISTRATION

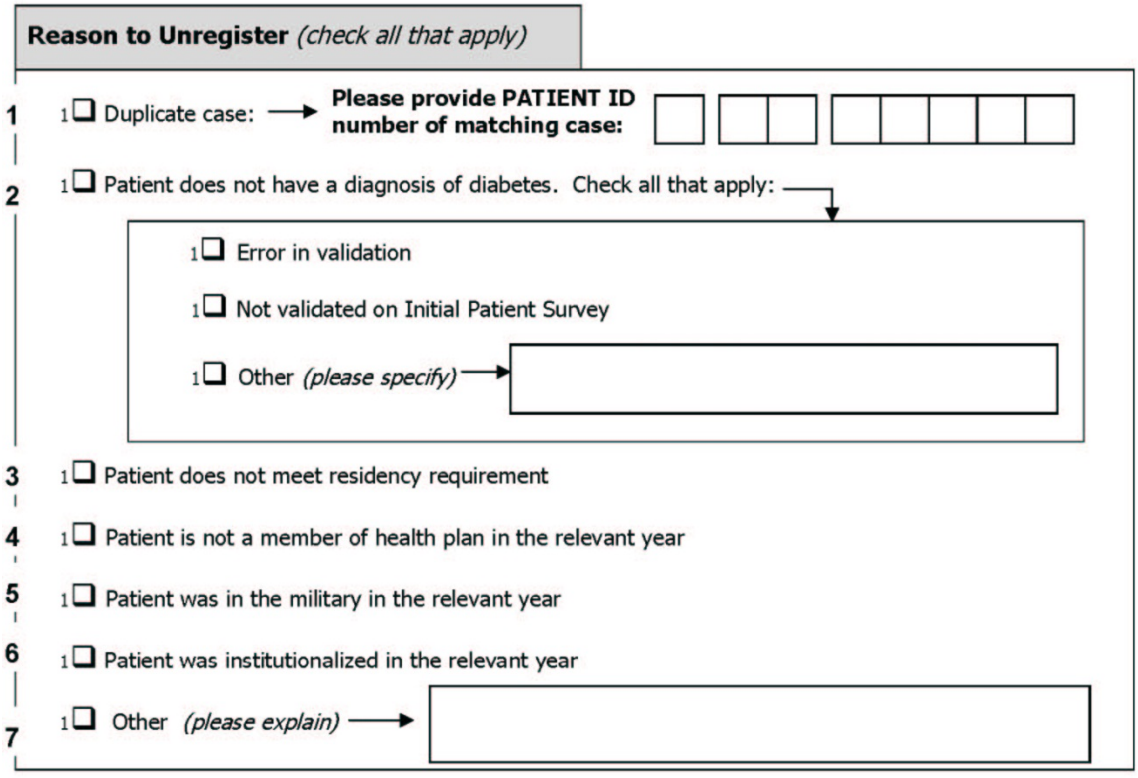
During the process of data collection and participant contacts, a case may be discovered to be ineligible or duplicated. When an already registered case is later found not to meet one or more of the eligibility criteria, or is shown to be a duplicate case, the case will be un-registered, will not be counted for incidence estimates and will not be invited to further participate in data collection. Un-registration information will be collected using the Un-registration Form.

NOTE: If a participant or their parent states (on the IPS or some other communication method) that he or she does not have DM but the provider report or medical record review indicates that the participant does have DM, then the case should NOT be Unregistered.

7.3.1. *The Un-registration Form*

The un-registration process is a multi-step task. This task includes the completion of the Un-registration Form (shown in Figure 3), data entry into the web-based data management system and ending with data entry into the tracking database. The intent of this form is to document the reason(s) the case is not eligible and request it to be un-registered.

Figure 3 - Un-registration Form



Item 1: Refers to the case being a **duplicate.**

- Check this box if this Participant is determined to be a duplicate case and place the matching Participant's assigned identification number in the boxes provided. The Participant identification affixed to the top of the Un-registration Form is the identification number of the Participant to be unregistered. The matching identification number is that of the Participant to be retained in the SEARCH study.

Item 2: Refers to the Participant not having a **diagnosis of diabetes**.

- Check this box if you determine the Participant has not been diagnosed with diabetes, and
    - Check the corresponding box(es) determining where the error in diagnosis occurred.

      Note: If parent/participant selects **"No"** for diabetes diagnosis on the IPS, but there is reliable information (provider diagnosis/medical record notation) indicating the Participant actually has diabetes, the Participant SHOULD NOT be un-registered. Note within the tracking database, this Participant should not be contacted further. The reason should be documented in the tracking database as the Participant believes they do not have diabetes ("denies diabetes").

Item 3: Refers to the Participant not meeting the **residency** requirement.

- Check this box if the Participant did not meet the residency requirement - specifically, if they were not a resident of the county and state.

Item 4: Refers to the Participant's non-membership of a **health plan member**.

- Check this box if the Participant is not a member of a health plan in the relevant year.

Item 5: Refers to the Participant being in the military.

- Check this box if the Participant was a member of the military at diagnosis for incident cases.

Item 6: Refers to the Participant being institutionalized.

- Check this box if the Participant was institutionalized at diagnosis for incident cases.

Note: There are no SEARCH questions or uploaded fields that refer to information relating to Participant institutionalization. If this information becomes known from available records or anecdotal information, the Participant

should be unregistered and the reason for un-registration should be noted in the TDBS.

Item 7: Refers to **Other** reasons the Participant may not be eligible.

- If there are other reasons the Participant is not eligible for the SEARCH study, check this box and specify the reason.

The Un-registration ID Number box is provided to record the un-registration number provided by the web-based data management system as explained below.

7.3.2. *Unregistering the case*

Once a participant is determined to be ineligible and the Case Un-registration Form is completed, the subject will be unregistered from the web-based data management system **and** the tracking database.

- The information on the Case Un-registration Form will be entered on the appropriate screen of the web-based data management system.
- When finished, click on the button at the bottom of the screen labeled **Unregistered**.
- The computer will respond with a confirmation message and **un-registration number**. This number must be written on the bottom of the Un-Registration Form.
- After the Participant has been unregistered on the web-based data management system, the Participant should be unregistered in a center's local TDBS by flagging the Participant's record as un-registered.

To un-register the participant using the ACCESS TDBS:

- Access the Patient Information screen and click the Case tab.
- Enter the **Un-registration** number. This will confirm the participant has already been unregistered from the central database.
- Click on the button on the bottom of the screen labeled unregister.
- The program will ask for confirmation that you wish to unregister the participant. Click on the appropriate response.
- Enter the reason the participant was unregistered in the box provided.

7.4. UN-REGISTERING PARTICIPANTS IDENTIFIED AS NOT HAVING DIABETES

In some situations, cases originally registered in SEARCH may be determined later on to actually not have diabetes. This may occur as a result of a short-term elevation in blood glucose resulting in an inappropriate diagnosis of diabetes by a health care provider. If after the clinic site has reviewed the available data on this participant, including a confirmatory

assessment by a pediatric endocrinologist, and determined that it is not a true case of diabetes, the study site should go through the un-registration procedures for this participant in the SEARCH database.

7.5. LATE REGISTERED CASES

Systematic re-ascertainment will NOT continue in SEARCH 4.

7.6. CASE ASCERTAINMENT COMPLETENESS

The completeness of ascertainment for each site will be estimated by dividing the number of identified cases by the estimated total number obtained from the capture-recapture analysis. The capture-recapture corrected estimate will be computed by dividing the observed incidence rate by the estimated capture-recapture rate. This corrected estimate can be seen as a ratio of 2 random variables. Pooled estimates that borrow information across site, sex and age groups will be used to guarantee that the capture-recapture rate and its associated standard error can be computed for all combinations of the variables considered in the analysis. Stratification by site, diabetes type, race/ethnicity, sex and age group can sometimes lead to small cell count causing convergence failures in the maximum likelihood estimation routines. Pooled estimation performed assuming a log-linear model 84 makes it possible to obtain the maximum likelihood estimates in these cases and simplifies the derivation of the standard error associated with the estimated percentage completeness. We will consider models that include site, diabetes type, race/ethnicity, sex and age group and all relevant interaction effects between them as covariates. The standard error associated with the incidence rates can also be derived in a similar fashion within the log-linear model framework, such that the delta method can be used to derive the standard error associated with capture-recapture corrected incidence rate. We will consider both first and second order Taylor expansions, and compare the accuracy of each set of estimates. This approach is similar to that as applied to derive capture-recapture corrected incidence estimates. More information is provided in the technical report which is attached in Appendix A.

Appendix A: Technical Report

SEARCH TECHNICAL REPORT

Estimation of Completeness of Case Ascertainment Using Capture-Recapture

Richard F. Hamman, MD DrPH[1]
Dana Dabelea, MD, PhD[1]
Angela D. Liese, PhD[2]
Lenna L. Liu, MD, MPH[3]
Jennifer W. Talton, MS[4]
Andrea Anderson, MS[4]
Mark A. Espeland, PhD[4]
Kristi Reynolds, PhD[5]
Scott Isom, MS[4]
Jasmin Divers, PHD[4]

July, 2015
based on prior reports from 2005, 2007 and 2013

REVISED January 2016
***As of January 2016, the 2 modes of ascertainment in SEARCH will be defined as “inpatient or emergency” and “outpatient”.

[1] University of Colorado Denver, Colorado School of Public Health, Department of Epidemiology
[2] Arnold School of Public Health, University of South Carolina, Department of Epidemiology and Biostatistics
[3] University of Washington School of Medicine, Department of Pediatrics
[4] SEARCH Coordinating Center, Wake Forest University School of Medicine, Department of Biostatistical Sciences
[5] Kaiser Permanente Southern California

Background

The purpose of this technical report is to summarize SEARCH activities to estimate the completeness of case ascertainment using the capture-recapture (C-R) method. The goal is to estimate the total size of the population of youth with diabetes aged 0-19 in a population, when the size of that population is not known. Case ascertainment through multiple sources provides a count of the number of cases found, but the number not identified remains unknown and must be estimated. The C-R method[1] was developed from animal biology to estimate the size of rodent populations, but it has been applied extensively to human disease situations [2-7]. The unknown total population size is estimated based on the number of cases found in more than one source (e.g., duplicate records from multiple hospitals, health care offices, and other sources). The approach is shown in Figure 1. Since two or more sources are required for C-R, it was not possible to use it in the SEARCH sites primarily utilizing one data source. These include the Kaiser Permanente Southern California site and participating Native American Tribes. The Kaiser Permanente site uses information from multiple health databases (laboratory, pharmacy, inpatient and outpatient encounters) and direct case reports from pediatric endocrinologists but these sources are not independent. Native American tribes used a single source, the Indian Health Service RPMS record system. This report further explores the use of two mode or multiple-mode sources in a systematic way for all four geographic sites.

Figure 1. Estimation of the total (unknown) population size.

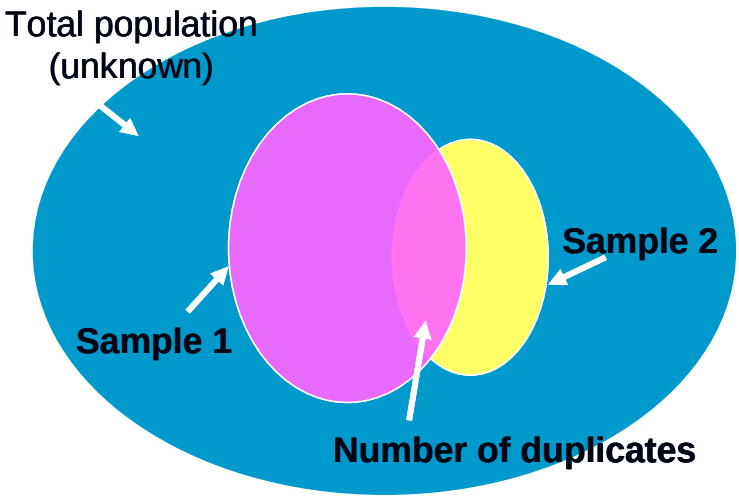


Table 1 summarizes the data display for a simple two source situation, where N can be estimated algebraically as shown in the formula if it is assumed that the two sources identify cases independently (this assumption cannot be verified without additional data).

Table 1: Summary of capture-recapture calculations

Source 1	Source 2		
	Yes	No	Total
Yes	A	B	A+B
No	C	X=?	?
Total	A+C	?	N=?

$$\widehat{N} = \frac{(A+B+1)(A+C+1)}{A+1}$$

The addition of 1 to each cell counts prevents the estimated totals from taking nonsensical values like 0 and infinity.

Assumptions

Traditional C-R methods, such as we have adopted, make the following assumptions[1]: Cases are:

- From the same space and time. This means that geographic and temporal residence is the same for all members of the population and can be determined similarly in all cases.

- Identical with respect to how likely they are to be identified. This assumption means that every case has the same probability of being identified by a given source, i.e., that some cases are not inherently easier or less difficult to identify than others. This is rarely met in health care studies.

- Independently ascertained by separate modes. The assumption of independence of sources is rarely met in disease ascertainment but can be dealt with using log-linear models with interaction terms to estimate and model the source dependence when more than two modes of ascertainment are involved. When only two ascertainment modes are available, the assumption of independence cannot be assessed.

- Matched between modes of ascertainment. The assumption of equal matching between modes of ascertainment assumes that sufficient data are available on personal identifiers from each source to be 'certain' that cases identified in multiple sources are, or are not, the same person. This may vary across sites and within sites across sources, depending on the amount of personal information provided by a source.

- Cases have been validated. This assumes that cases truly have diabetes and that this can be determined in each source.

- Cases are from a closed population. This assumption means that cases in the total population are not moving in or out of the population during the time interval.

Methods

In each of the geographic sites (Colorado, Ohio, South Carolina, and Washington) cases were identified from multiple sources (CO: 13+; Ohio: 20; SC: 41; WA: 26). A "source" was defined as any location where cases were reported. Sources were then aggregated. First, all individual small practices were usually grouped into 'practices', but these were initially maintained distinctly from larger pediatric endocrine practices, HMOs providing larger numbers of cases, etc. Individual hospitals were also maintained separately. Matching across sources was done on a regular basis as cases were reported to identify potential duplicate records. Initial computerized listings were generated, sorted and compared using available personal health identifiers (PHI), and then manual matching was completed. The amount of PHI available to conduct the matching across sites differed by site, with some sites unable to identify names at the first receipt of data. Once matching was accomplished across sources, the sources were further grouped into 'modes' of ascertainment. After exploratory analyses, all provider sources were aggregated, as were all hospital system records and modes were defined for all sites as 'provider' and 'hospital'. Several sources were large health care systems that included both ambulatory and inpatient facilities (e.g., Children's Hospital, Seattle). In these cases, manual review of records categorized youth by whether they had been cared for in either one or both portions of the system to allow better classification of the mode of ascertainment. As of January 2016, the 2 modes of ascertainment in SEARCH will be defined as "inpatient or emergency" and "outpatient". A SEARCH study participant is considered to be "inpatient or emergency" if the participant was ascertained during a hospital visit that included at least one overnight stay or during treatment by an emergency department or transport service. Cases ascertained during an outpatient visit that did not involve a hospital stay and occurred in a setting suitable for regular follow-up care will be classified as outpatient. For example, a participant identified during an emergency room visit will be classified as inpatient or emergency. A participant ascertained through a private doctor's office visit will be classified as outpatient. The inpatient or emergency mode will include most of the sources that were initially classified as 'hospital' including hospital admissions, surgical admissions, ICU/PICU admissions, in-hospital observation less than 24-hours, emergency room visits, and air/water/ground ambulance transfers. Similarly, the outpatient mode will consist of the majority of sources previously included in the 'provider/other' source including primary/specialty care offices, community outreach clinics, diabetes treatment centers, and also hospital based research or outpatient clinics, and other visit settings not associated with inpatient or emergency care.

The list of sources and corresponding mode of ascertainment is provided in the appendix.

Once two modes were identified and their duplicates noted, log linear models[8,24,25] were fitted to the data to estimate the total (unknown) population. These estimates were computed separately for prevalent 2001, prevalent 2009 and incident 2002 to 2009 youth. The models were fitted using all the data that was available in each subset adjusting for relevant covariates including and site. Multiple mode interaction models were evaluated systematically for each of the four geographic sites. For models with more than two modes, an estimate of the 'best' model was based on identifying the minimum value (best fit) of an information criterion statistic[16] defined as:

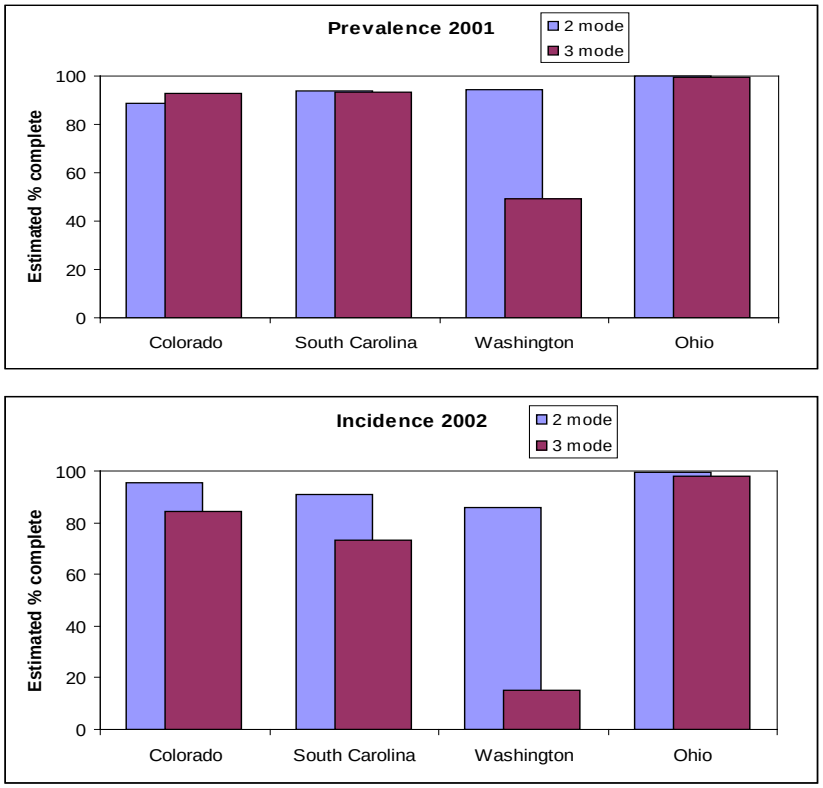
$$IC = G^2 - c \times df$$

Where:
G^2: Likelihood ratio statistic (-2 logarithm of the ratio of the likelihood of the fitted model to the likelihood of the saturated model);
df: Number of degrees of freedom for the comparison of any fitted model with the saturated model;
c: A constant that varies with the method use to estimate the information criterion. For the minimum Akaike information criterion (AIC), c = 2.
The percent completeness of ascertainment for any group was estimated as the number of observed cases divided by the total number estimated from C-R. Estimates of the ascertainment rates pooled across clinical sites were produced from a global log-linear model [8], which allowed for separate intra-site performance. The rates were estimated using maximum likelihood and the standard errors were estimated using the delta method [9].

Results

Four different ascertainment modes were initially defined for the analysis. In the "provider" mode, practices were split into "endocrine" and "other", and "hospitals" were divided into "hospitals only" and "integrated practices". There were too few cases in some locations in each of the four modes to successfully use this approach, so "hospitals" was used as a single mode, and there were two practice modes (endocrine and other) to allow a three mode model. Three mode models were explored allowing all possible 3-way interactions, and the 'best' model was chosen with the lowest IC value. The overall estimate of completeness of ascertainment was then compared to the 2 mode model (from which no IC value can be calculated). As shown in Figure 2 below, there was wide variability in the 3 mode estimates across centers, ranging (for prevalence) from 44 to 99% complete. This 3 mode estimate can be compared to the range of the 2 mode estimates from 89 to ~100% across sites. For incidence estimates, the 3 mode models ranged from 15 to 98%, whereas the 2 mode model was much more consistent – from 86% to ~ 100%. Results of the different models within site also showed substantial heterogeneity. Across sites, there were several different patterns of interactions between sites – that is, there were not consistent types of modes that interacted across centers. In Ohio, the 2 and 3- mode approaches gave almost identical results, since all estimates were > 97%. The widest changes within a single site occurred in Washington, where for prevalence, the 2 mode estimate was 94% and the 3 mode was 49%; for incidence it was 86% for 2 mode and 15% for the "best" 3 mode model.

Figure 2. Estimates of case ascertainment completeness for prevalence (2001) and incidence (2002) using a 2 mode (blue bar) and 3 mode (red bar) capture-recapture model, by geographic site.



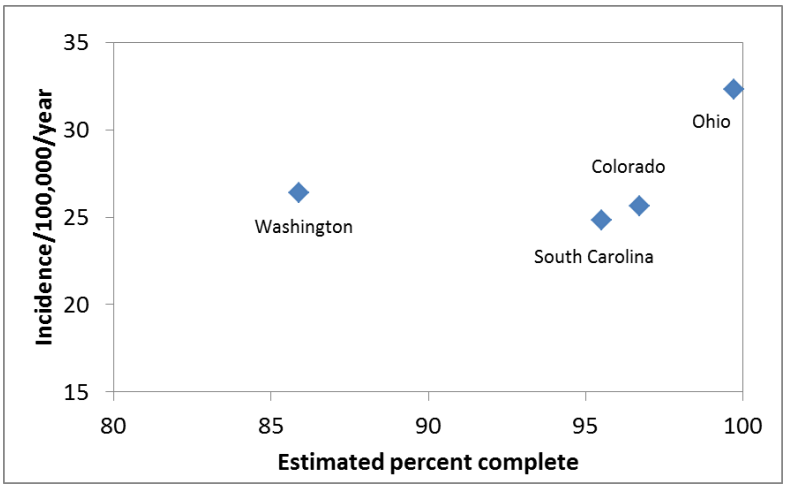
Based on the heterogeneity of results from the 3 mode (interaction) models, which appeared to be due largely to site specific differences in patterns of care, reporting, location of duplicate cases, and statistical variability, it was decided that the 2 mode model provided more consistent results with better face validity. For example, the 3 mode model suggested that Washington missed over 1500 cases in 2002, more than 3 times the number actually identified. Another rationale for choosing the 2 mode estimate comes from the consistency of the incidence rates by site. This consistency is shown for total incidence (all types) in 2002 in Table 2. If the 3 mode model were correct, it would suggest that rates in Washington and South Carolina would be substantially lower than actually observed.

Table 2. Total incidence rates of diabetes (all types) by geographic sites, 2002.

Center	Youth with DM	Population Denominator (Person-years)	Incidence Rates (per 100,000/year)	95% CI (per 100,000/year)
Ohio	355	1,097,960	32.3	29.1-35.9
South Carolina	539	2,170,362	24.8	22.9-27.0
Washington	509	1,927,958	26.4	24.2-28.8
Colorado	655	2,553,884	25.7	23.8-27.8

The incidence rate in Ohio was ~ 25% higher than rates in the other three geographic sites and Ohio also had the highest estimated completeness. However, rates for South Carolina, Washington and Colorado were quite similar (24.8-26.4) while estimates of completeness ranged from 86% (Washington) to 97% (Colorado) (Figure 3). It cannot be ruled out that higher rates in Ohio were due in part to slightly higher estimated completeness, however, over this narrow range, it did not appear to influence rates in the other three sites.

Figure 3. Incidence rates (per 100,000/yr.) by estimated completeness from capture-recapture (2 mode) by site, SEARCH 2002 incidence



For these capture-recapture analyses, we therefore chose a 2 mode model to estimate completeness by site and overall. Table 3 shows the results using C-R for the four geographic sites using a two mode ascertainment model (‘hospital’ vs. ‘provider sources’ combined) for prevalent and incident cases, by site and age group.

Table 3. Summary of percent completeness of ascertainment by capture-recapture analysis by year, site and age-group.
SEARCH 2001-2002

		Age group at diagnosis				Total	95% CI	
Year	Site	0 - 4	5 - 9	10 – 14	15 – 19		LL	UL
2001	Colorado [1]	95.3	91.6	92.0	83.9	88.8		
Prevalent	Washington	95.9	92.5	89.8	87.3	89.3		
	Ohio	100.0	100.0	99.9	99.4	99.8		
	South Carolina [1]	94.4	96.6	99.6	94.9	97.0		
	All sites [2]	96.3	94.1	93.3	90.9	92.2	91.1	93.3
2002	Colorado [3]	99.0	98.8	93.7	91.9	96.9		
Incident	Washington	94.4	87.3	83.7	79.6	85.9		
	Ohio	--[4]	99.9	100.0	97.9	99.7		
	South Carolina [3]	97.2	98.3	96.2	86.0	95.5		
	All sites [2]	97.5	95.1	92.8	87.9	93.8	91.9	95.6

(1) Prevalence sub-areas of state (2) Weighted average using observed cases at each site as weight (3) Entire state (4) Too few cases to estimate

The C-R analyses suggest that over all four sites, both prevalent and incident cases are at least 91% ascertained. Ascertainment appeared somewhat lower in the older than younger age groups, reflecting clinic experience at the difficulty of identifying and recruiting older youth. Analyses will be updated periodically and reported in relevant manuscripts.

Limitations

These analyses have a number of limitations, and while the C-R method is often touted as the best way to estimate completeness of ascertainment [2,10], several authors have identified significant problems with the method [1,11-23]. In the context of the current US healthcare system and HIPAA regulations, several of the limitations of the method were encountered. These include: a) possible incomplete matching across sources due to restrictions on access to names for matching in some states (thus violating the assumption that cases can be matched in all sources); b) uncertainty about the residence location of some cases (thus violating the assumption that cases were from the study area); and c) design of the ascertainment system for efficiency (thus avoiding sources of likely duplicate cases). Each of these problems is known to inflate the estimated number of total cases in the C-R analysis, leading to an underestimate of the percent completeness. In addition, given the multiple sources of information used to identify cases, it was possible to arbitrarily combine these sources into two modes in many alternate ways instead of the one chosen: 'hospital' vs. 'provider'. If this was done on the identical dataset, it was possible to drive the estimates of completeness from 72.7% to 86.5% completeness (in South Carolina as an example). An example of another problem came from Colorado. Preliminary analyses conducted in December 2003 suggested that prevalent cases were 87% complete, and that there were approximately 1246 estimated cases if all cases had been identified. By December of 2004, Colorado had identified a total of 1366 prevalent cases;

however, the C-R estimate dropped to 81.5% complete. Addition of duplicates changed this to 88.8% complete. We, therefore, believe that the C-R estimates shown in Table 2 are a 'lower bound' on the completeness of ascertainment in these four sites. While some redesign of the case ascertainment system might provide better estimates of completeness, inherent limitations of access to records in all sources with incomplete personal identifiers make the use of C-R in the US difficult. Nonetheless, given the large geographic areas covered, and the multiple providers and hospitals contacted and used during case ascertainment, SEARCH achieved at least 90% ascertainment of prevalent and incident cases across all four geographic sites.

Systematic evaluation of models allowing interaction terms between 3 ascertainment modes did not improve estimates of ascertainment completeness, and were inconsistent with the observed rate consistency. Given the limitations noted above, the 2-mode model will continue to be used.

In January of 2013, capture-recapture estimates were updated to reflect on-going case ascertainment and inclusion of later registered cases. In order to estimate completeness by race/ethnicity and type of diabetes, the approach taken for fitting the log-linear model was revised. The current approach relies on adjusted models instead of the stratified models that were used previously. As the number of stratification variables increased, the cell counts observed in some cases were too small, which prevented the maximum likelihood estimation routines from converging. The adjusted models do not suffer from this limitation since they use all the available data[24-25].

Conclusions and use of results in SEARCH

Capture-recapture methods in the four geographic sites resulted in an overall estimate of completeness of at least 90% for both prevalence and incidence. No estimates are possible in the California and Native American sites. Given the closed nature of these data systems and the comparable methods used to identify cases in these health systems, it seems likely (though untested) that ascertainment rates were at least as good, if not better, than in the geographic sites. It is likely, given the limitations of the use of C-R methods as implemented in SEARCH, the estimates of completeness of ascertainment are a lower bound on the actual completeness.

References

1. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev.* 1995;17:243-264.
2. LaPorte RE, McCarty DJ, Tull ES. Counting birds, bees, and NCDs. *Lancet.* 1992;339:494-495.
3. Cormack RM, Chang YF, Smith GS. Estimating deaths from industrial injury by capture-recapture: a cautionary tale. *Int J Epidemiol.* 2000;29:1053-1059.
4. Corrao G, Bagnardi V, Vittadini G, Favilli S. Capture-recapture methods to size alcohol related problems in a population. *J Epidemiol Community Health.* 2000;54:603-610.
5. Ballivet S, Salmi LR, Dubourdieu D. Capture-recapture method to determine the best design of a surveillance system. Application to a thyroid cancer registry. *Eur J Epidemiol.* 2000;16:147-153.

6. Tilling K, Sterne JA, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol.* 2001;30:1351-1359.
7. Verstraeten T, Baughman AL, Cadwell B, Zanardi L, Haber P, Chen RT. Enhancing vaccine safety surveillance: a capture-recapture analysis of intussusception after rotavirus vaccination. *Am J Epidemiol.* 2001;154:1006-1012.
8. Bishop YMM, Fienberg SE, Holland PW. Chapter 6. *Discrete multivariate analysis.* Cambridge. MA: MIT Press; 1975.
9. Espeland MA. A general class of models for discrete multivariate data. *Communications in Statistics: Simulation and Computation.* 1986;15:405-424.
10. LaPorte RE, McCarty D, Bruno G, Tajima N, Baba S. Counting diabetes in the next millennium. Application of capture-recapture technology. *Diab Care.* 1993;16:528-534.
11. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol.* 1992;135:1060-1067.
12. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol.* 1996;25:474-478.
13. Nanan DJ, White F. Capture-recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Dis Can.* 1997;18:144-148.
14. Domingo-Salvany A, Hartnoll RL, Maguire A et al. Analytical considerations in the use of capture-recapture to estimate prevalence: case studies of the estimation of opiate use in the metropolitan area of Barcelona, Spain. *Am J Epidemiol.* 1998;148:732-740.
15. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol.* 1999;52:917-926.
16. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol.* 2000;152:771-779.
17. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? *Int J Epidemiol.* 2000;29:536-541.
18. Hook EB, Regal RR. On the need for a 16th and 17th recommendations for capture-recapture analysis. *J Clin Epidemiol.* 2000;53:1275-1277.
19. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med.* 2001;20:3123-3157.
20. Tilling K. Capture-recapture methods--useful or misleading? *Int J Epidemiol.* 2001;30:12-14.
21. Laska EM, Meisner M, Wanderling J, Siegel C. Estimating population size and duplication rates when records cannot be linked. *Stat Med.* 2003;22:3403-3417.
22. Carle, F., Gesuita, R., and Gregorio, F. Effect of overlapping among sources on the validity of capture-recapture estimates of prevalence of diabetes. 39th Annual Meeting EDEG . 2004.
23. Verlato G, Muggeo M. Capture-recapture method in the epidemiology of type 2 diabetes: a contribution from the Verona Diabetes Study. *Diab Care.* 2000;23:759-764.
24. Chao A, Tsay PK, Lin SH et al.: The applications of capture-recapture models to epidemiological data. Statist Med 2001; 20:3123-3157.

25. Cormack RM. Loglinear models for capture-recapture. Biometrics 1989; 45:395–413.

Supplementary Table 1: Sources included in each mode of ascertainment in South Carolina

South Carolina	Old classification	New classification
Anderson Area Medical Center	Hospital	Inpatient/ER
Anmed Child Health Center	Hospital	Inpatient/ER (same facility as Anderson Area Medical Center)
Beaufort Hospital	Hospital	Inpatient/ER (don't obtain cases from here any longer)
Carolinas Hospital - Florence	Hospital	Inpatient/ER (don't obtain cases from here any longer)
Greenville Health System	Hospital	Inpatient/ER and Outpatient
Greenville Memorial Hospital	Hospital	Inpatient/ER (falls under Greenville Health System)
Lexington Medical Center	Hospital	Inpatient/ER (don't obtain cases here any longer)
Mauldin Medical Center	Hospital	Not applicable (don't obtain cases here any longer)
McLeod Hospital	Hospital	Inpatient/ER
McLeod Regional Medical Center	Hospital	Inpatient/ER (same as McLeod Hospital)
Orangeburg Hospital	Hospital	Inpatient/ER
Palmetto Bapstist Medical Center Easley	Hospital	Inpatient/ER (same as Palmetto Health Easley)
Palmetto Baptist Medical Center Columbia	Hospital	Inpatient/ER (same as Palmetto Health Baptist)
Palmetto Health Alliance/RMH	Hospital	Inpatient/ER (same as Palmetto Health Richland)
Palmetto Health Baptist	Hospital	Inpatient/ER
Palmetto Health Easley	Hospital	Inpatient/ER
Palmetto Health Richland	Hospital	Inpatient/ER
Roper St Francis Hospital	Hospital	Inpatient/ER (don't obtain cases here any longer)
Spartanburg Regional Healthcare System	Hospital	Inpatient/ER
Spartanburg Regional Medical Center	Hospital	Inpatient/ER (same as Spartanburg Regional Healthcare System)
The Regional Medical Center of Orangeburg and Calhoun Counties	Hospital	Inpatient/ER (same as Orangeburg Hospital)
(new) Palmetto Health Parkridge	Hospital	Inpatient/ER
Amrhein	Other	Outpatient (Same as GHS Pediatric Endocrinology)
Broome	Other	Outpatient (no longer active)
Carolina Diabetes and Kidney Center	Other	Outpatient (no longer active)
Coulter	Other	Outpatient (no longer active)
GHS Pediatric Endocrinology	Other	Outpatient
Heinze	Other	Outpatient (no longer active)
Hoffman	Other	Outpatient (no longer active)
Jackson	Other	Outpatient (same as USC Pediatric Endocrinology)
Jocelyn Myers	Other	Outpatient (no longer active)
Laurel Endocrine-Brennan	Other	Outpatient (no longer obtain cases from there)
McLeod Pediatric Subspecialists	Other	Outpatient
McLeod/Woodberry	Other	Outpatient (same as McLeod Pediatric Subspecialists)

Mendes	Other	Outpatient (formerly with McLeod Pediatric Subspecialists)
MUSC	Other	Inpatient/ER and Outpatient
Parker	Other	Outpatient
Raine	Other	Outpatient (no longer obtain cases from there)
Schwartz	Other	Outpatient (formerly with USC Pediatric Endocrinology)
USC Pediatric Endocrinology	Other	Outpatient or Inpatient (if seen during rounds)
Willi	Other	Outpatient (formerly with MUSC)
Benedict College	Other	Outpatient (no longer obtain cases from there)
Black River Community Health Care	Other	Outpatient (no longer obtain cases from there)
Brooks Health Center	Other	Outpatient (no longer obtain cases from there)
C.S.R.A. Renal Services	Other	Outpatient (no longer obtain cases from there)
Care-South Carolina	Other	Outpatient
Carolina Health Greenwood	Other	Outpatient (no longer obtain cases from there)
Carolina Peds	Other	Outpatient (no longer obtain cases from there)
Catawba Longhouse	Other	Outpatient (no longer obtain cases from there)
Children and Family HealthCare Center (USC College of Nursing)	Other	Outpatient (no longer obtain cases from there)
CSRA Renal Services	Other	Outpatient (no longer obtain cases from there)
Debbie Yoman	Other	Outpatient (no longer obtain cases from there)
Diabetes Education Center in Lancaster	Other	Outpatient (no longer obtain cases from there)
Doctors Care (statewide)	Other	Outpatient (no longer obtain cases from there)
Eau Claire	Other	Outpatient
Eau Claire Cooperative Health Center	Other	Outpatient
Eau Claire Cooperative Health Centers	Other	Outpatient (same as Eau Claire and Eau Claire Cooperative Health Center)
Family Health Care Center--Orangeburg	Other	Outpatient (same as Family Health Centers, Inc)
Family Health Centers, Inc.	Other	Outpatient
Family Practice Center-Palmetto Health	Other	Outpatient
Franklin Coulter	Other	Outpatient (no longer active)
Grand Strand Ped	Other	Outpatient (no longer obtain cases from there)
Grand Strand Pediatrics	Other	Outpatient (no longer obtain cases from there)
Lexington Pediatrics	Other	Outpatient (no longer obtain cases from there)
Longcreek Family Practice	Other	Outpatient (no longer obtain cases from there)
Orangeburg Hospital Diabetes Educator	Other	Outpatient (no longer obtain cases from there)
Orangeburg Hospital-Diabetes Educator	Other	Outpatient (no longer obtain cases from there)
Pediatric Associates, P.A.	Other	Outpatient (no longer obtain cases from there)
Pediatric Associates, PA	Other	Outpatient (no longer obtain cases from there)
Richland Community Health Care Association	Other	(no longer active)
SandHills Pediatrics-Wessinger	Other	Outpatient (no longer obtain cases from there)
Sea Island Pediatrics P.A.	Other	Outpatient (no longer obtain cases from there)
Self Report	Other	
The Pediatric Clinic	Other	Outpatient (no longer obtain cases from there)
Undefined	Other	
USC Central Billing	Other	Outpatient (no longer obtain cases from there)

USC Department of Family & Preventive Medicine	Other	Outpatient (no longer obtain cases from there)
USC OB/GYN clinic (1801 Sunset)	Other	Outpatient (no longer obtain cases from there)
USC-Central Billing-Dr. Bryant	Other	Outpatient (no longer obtain cases from there)
Yoman	Other	Outpatient (no longer obtain cases from here)

Supplementary Table 2: Sources included in each mode of ascertainment in Ohio

Ohio	Old classification	New classification
FHH	Hospital	Inpatient or emergency
StLuke	Hospital	Inpatient or emergency
UniversityHosp	Hospital	Inpatient or emergency
Christ	Hospital	Inpatient or emergency
Mercy	Hospital	Inpatient or emergency
MRH (name has changed to Atruim)	Hospital	Inpatient or emergency
Jewish	Hospital	Inpatient or emergency
CCHMC	Hospital	Inpatient or emergency
StElizabeth	Hospital	Inpatient or emergency
GoodSam/ Bethesda	Hospital	Inpatient or emergency
McCullough	Hospital	Inpatient or emergency
FHH	Hospital	Outpatient
StLuke	Hospital	Outpatient
UniversityHosp	Hospital	Outpatient
Christ	Hospital	Outpatient
Mercy	Hospital	Outpatient
MRH (name has changed to Atruim)	Hospital	Outpatient
Jewish	Hospital	Outpatient
CCHMC	Hospital	Outpatient
StElizabeth	Hospital	Outpatient
GoodSam/ Bethesda	Hospital	Outpatient
McCullough	Hospital	Outpatient
EndoAdult	Other	Outpatient
EndoPeds	Other	Outpatient
PrimaryMDs	Other	Outpatient
CDEs	Other	Outpatient
Universities	Other	Outpatient
Other	Other	Outpatient
CintiHealthDept	Other	Outpatient
Anthem	Other	Outpatient
Aetna	Other	Outpatient

KYMedicaid	Other	Outpatient
BCMH	Other	Outpatient
CareSource	Other	Outpatient
Anthem	Other	Inpatient or emergency
Aetna	Other	Inpatient or emergency
KYMedicaid	Other	Inpatient or emergency
BCMH	Other	Inpatient or emergency
CareSource	Other	Inpatient or emergency

Supplementary Table 3: Sources included in each mode of ascertainment in Colorado

Colorado	Old classification	New classification
St Mary's in Grand Junction	Hospital	Inpatient or emergency
Exempla Hospitals	Hospital	Inpatient or emergency
The Children's Hospital/The Children's Hospital Colorado	Hospital	Inpatient or emergency
Centura Hospitals	Hospital	Inpatient or emergency
Boulder Community Hospital	Hospital	Inpatient or emergency
Pueblo, CO Hospitals/Metro Community Hospital	Hospital	Inpatient or emergency
Denver Health inpatient or emergency		
Barbara Davis Center	Other	Outpatient
Pediatric Endocrine Associates	Other	Outpatient
San Luis Valley/Valley Wide Health System	Other	Outpatient
Western Ped. in Grand Junction	Other	Outpatient
Salud Family Health Centers	Other	Outpatient
Denver Health outpatient	Other	Outpatient
Kaiser Permanente	Other	Outpatient
Providers/San Luis Valley Case Reports	Other	

Supplementary Table 4: Sources included in each mode of ascertainment in Washington

Washington	Old classification	New classification
Boldt	Hospital	
CHRMC	Hospital	
CHRMC inpatient	Hospital	
Harborview Medical Center	Hospital	
Madigan Medical Center	Hospital	
Mary Bridge	Hospital	
Mary Bridge inpatient	Hospital	
Providence St. Pete's	Hospital	
Seattle Children's inpatient	Hospital	
Swedish Medical Center	Hospital	
UW Medical Center	Hospital	
Valley Medical Center	Hospital	
Virginia Mason	Hospital	
	Hospital	
	Hospital	
	Hospital	
	Hospital	
	Hospital	
CHCKC	Other	
CHRMC Endo Clinic	Other	
CHRMC outpatient	Other	
Diabetes Care Center	Other	
Dr McGowen	Other	
Green	Other	
Joslin	Other	
Mauseth	Other	
MB outpatient	Other	
Minor & James clinic	Other	
N Sea Pub Health	Other	
Neighborcare	Other	
Ped Asso Olympia	Other	
PSNHC	Other	
SeaMar	Other	
Seattle Children's outpatient	Other	
Summit View Clinic	Other	
Swedish Joslin	Other	
UW Physicians Network	Other	
ADA	Other	
Camp Leo	Other	

GHC	Other	
Newspaper Advertisements	Other	
Other	Other	
SKWIDDS	Other	
	Other	
	Other	
	Other	
	Other	

Appendix B: ICD-9 and ICD-10 Codes for Case Ascertainment

	ICD-9 Codes						ICD-10 Codes				
	WA	**OH**	**CA**	**CO**	**SC**						
T1, T2, MODY	250.xx	250.xx	250.xx	250.xx	250.xx		E10.xx	E11.xx	E13.xx		
Persistent neonatal diabetes	775.1	775.1	775.1	775.1	775.1		P70.2				
Secondary diabetes	249.xx	249.xx		249.xx			E08.xx	E09.xx	E13.xx		
diabetic cataract			366.4		366.41		E08.xx	E09.xx	E10.xx	E11.xx	E13.xx
diabetic retinopathy			362.0x				E08.xx	E09.xx	E10.xx	E11.xx	E13.xx
polyneuropathy in diabetes			357.2				E08.xx	E09.xx	E10.xx	E11.xx	E13.xx
pre-existing diabetes in pregnancy			648.0x				O24.1	O24.0	O24.3	O24.8	

New ICD-10 codes to be used by all sites:	**E08.xx**	**E09.xx**	**E10.xx**	**E11.xx**	**E13.xx**	**P70.2**	**O24.0**	**O24.1**	**O24.3**	**O24.8**

10/19/2016