SSuN Best Practice Note

Strategy B: Random Sampling

PS19-1907 SSuN Strategy B

(B) Protocol-based enhanced case-based STD surveillance:

1. ***Systematic collection of enhanced data elements for reported cases of STD.***
2. ***Ascertainment of patient demographics, behavioral risk, STD treatment, clinical characteristics and HIV/STD preventive services for a probability sample of selected cases reported in the jurisdiction.***
3. ***Provision of technical assistance to state and local STD surveillance and program staff in monitoring and improving the quality of STD surveillance data.***

Why SSuN includes this strategy

Many cases routinely reported by providers and laboratories are missing information such as patient race, Hispanic ethnicity, sexual orientation, gender identity and other important clinical and behavioral characteristics.

These data are critical to understanding patterns in case incidence, differences in the burden of disease and to help direct prevention, control and health equity efforts. Conducting enhanced investigations on a ***random sample*** (also known as a probability sample) of cases allows us to develop valid estimates the distribution of cases by these important characteristics.

Key definitions: What do we mean by a 'random sample'?

A random or probability sample is defined as a smaller number of cases picked at random from the universe of all cases reported. What makes these cases useful for analysis is that each reported case has the same 'probability' of getting picked for the sample; as long as this is true, the distribution of characteristics found from investigating just this sample of cases would closely mirror what we would find if we could get the same information from all reported case.

The ratio of sampled to all reported cases is known as the **sample fraction**. In SSuN, the proportion of cases successfully investigated *among the sampled cases* is known as the **effective sample**. It is critically important that there are no **intentional or inadvertent** biases in the sample, which means that there should be no restrictions on which reported cases get sampled. The only exception to this rule is that the sample size may be allowed to differ by geographic unit by design.

All cases, regardless of any characteristic such as patient gender, race, ethnicity, age, source of report or method of data entry (i.e., electronic versus hand entered) must have the same probability of being sampled. It is acceptable to establish a different sample size by pre-defined geographic criteria (such as patient's county of residence) because this can be adjusted for in analysis by creating design weights, but within the geographic unit, no other criteria should be used to adjust the sample size.

Considerations for implementation

How do I pick a 'random sample' of my cases?

- The best practice is to 'roll the dice' as close as possible to the initial report and to capture the result permanently in the electronic case record:

- Use random number functions, built in to your surveillance data management system. All data management software (e.g., Oracle, SQL, etc.) have mathematical functions built in that can be accessed through stored procedures or through custom programming.
- This function must randomly generate a number between 0 and 1.0 (or between 0 and 100, depending on the specific function used); this range should never be altered or changed once sampling has begun.
- Create a permanent container (i.e., variable or separate column in data tables) to hold the result of a random number generator.
- Assure that this random number is only run and captured ONCE for any given record.

❖ Selecting a 'sample' is a second step that compares the randomly generated number to a fixed range; records where the randomly generated number falls into that fixed range (the sampling fraction) are considered 'in the sample'. This second step should also be done once, and a binary indicator or flag set to indicate that the case is 'in the sample'. Having this second step adds flexibility to allow for different sample fractions for different diseases or for different counties, regions or areas.

❖ Build a flexible way to use the random number generated for each record to sample cases differently depending on your local project design. For example, sample size can:

- vary by area or county
- vary by disease
- change over time to increase or decrease the sample size as conditions warrant.

Evaulation Activities Using best practices…

❖ To prevent any accidental biases to affect your sample, a best practice is to regularly (monthly or quarterly) compare the distribution of all cases to the distribution of cases in your sample by sex, age, geographic area, etc. to monitor representativeness of the sample. Small differences (<2%), especially when the number of cases is fairly small, are to be expected and may just be random variations. However, over time, as cases accumulate, these minor differences should become smaller not larger.

❖ If there are significant differences between the sampled cases and all cases, re-evaluate your process to assure that all steps are working correctly, double check your sample fraction and assure that the random number generator is functioning.

Fully automating the generation of a random number, and creation of a flexible method to change and modify the size of your sample of cases based on this random number can be a powerful tool for evaluating and enhancing surveillance activities across a range of diseases and data quality issues!

Other resources

❖ SSuN Cycle 4 Protocols: https://www.cdc.gov/std/funding/ssun/default.htm

CDC's Division of STD Prevention, Surveillance and Special Studies Team created this series of documents for CDC staff working on PS19-1907 STD Surveillance Network (SSuN) and for applicants and recipients of that NOFO, to help clarify strategies outlined and to support project implementation.