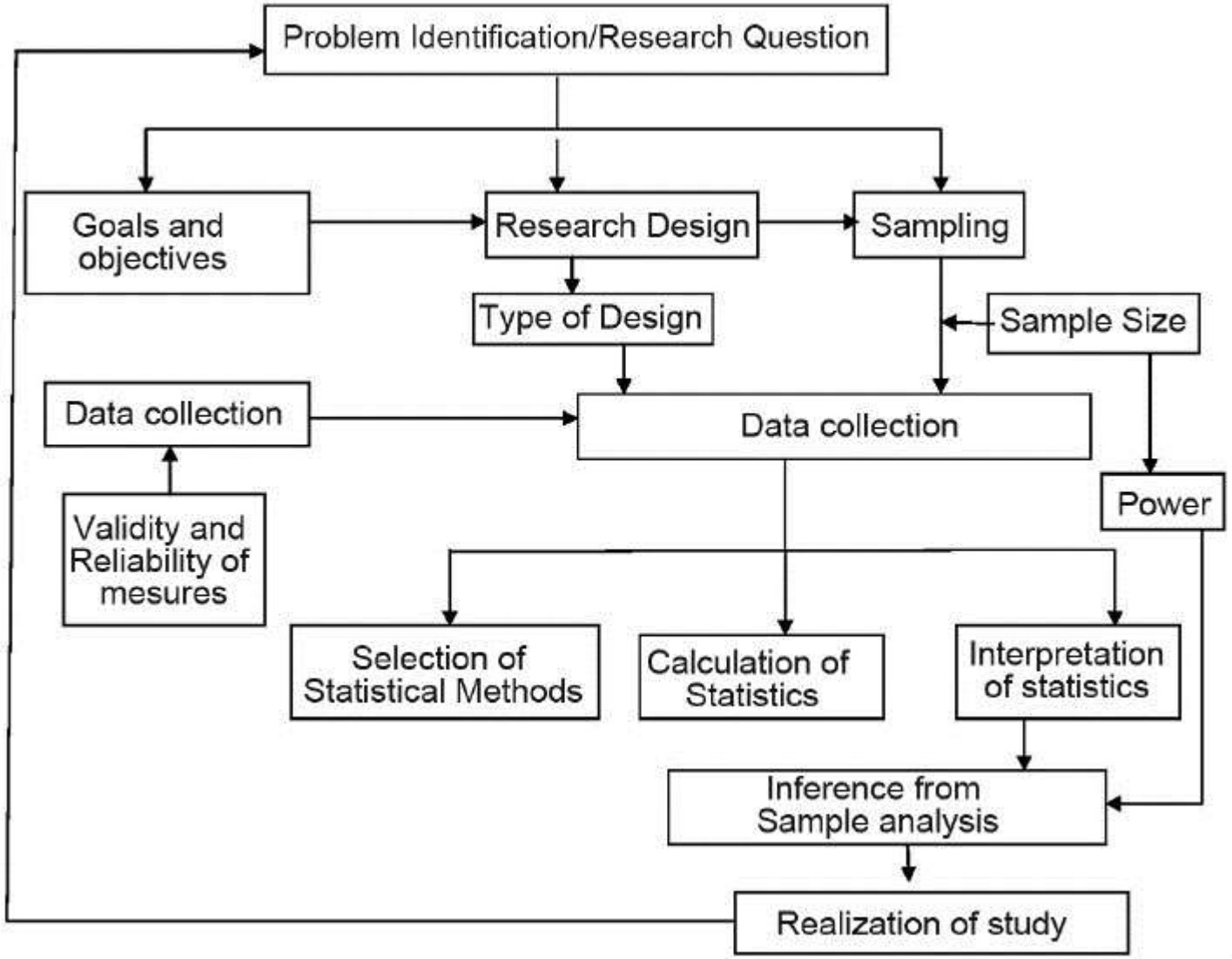
Sampling in Statistics

Dr Soumya Swaroop Sahoo
Associate Professor
Dept of Community and Family Medicine

Contents

Overall framework of Research Design



Why Sampling?

A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)

Why sample?

- Resources (time, money) and workload
- Gives results with known accuracy that can be calculated mathematically
- Can be externally validated to likewise representative population

Sampling Process

The sampling process comprises several stages:

- Defining the **population of concern**
- Specifying a **sampling frame,** a set of items or events possible to measure
- Specifying a **sampling method** for selecting items or events from the frame
- Determining the **sample size**
- Implementing the sampling plan
- Sampling and data collection
- Reviewing the sampling process

Essentials of Sampling

Representativeness: ensure by random selection

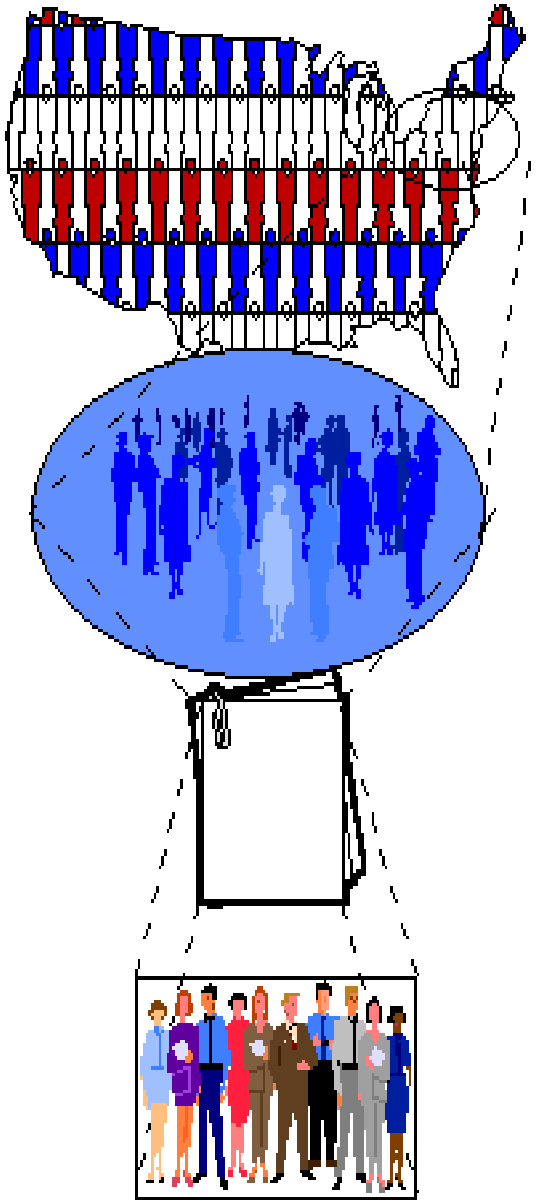
Adequacy: sample size

Independence: same chance of selection

Homogeneity: no basic difference in nature of units

Who do you want to generalize to?

The Theoretical Population



What population can you get access to?

The Study Population

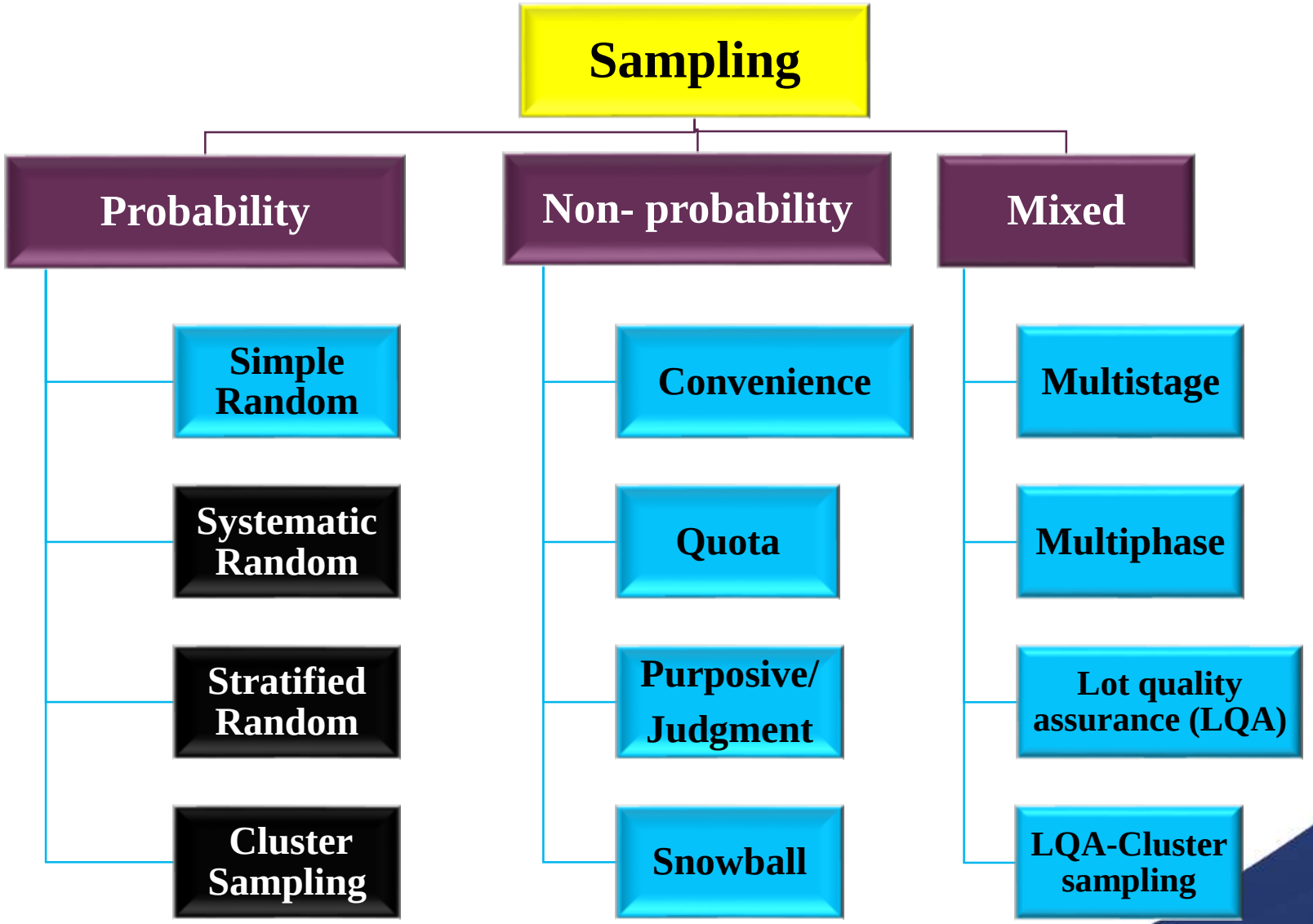
How can you get access to them?

The Sampling Frame

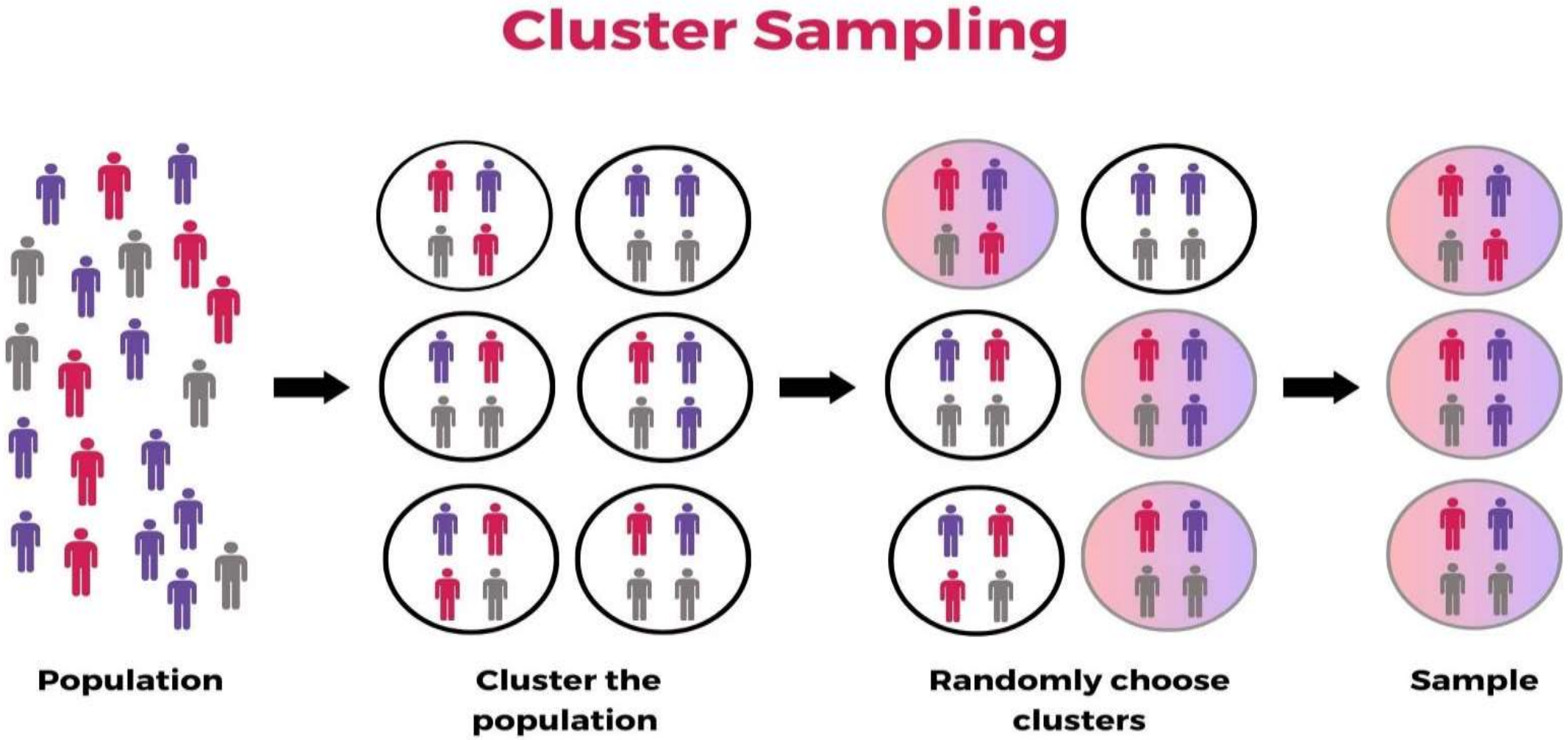
Who is in your study?

The Sample

SAMPLING BREAKDOWN


Sampling
Probability
Non- probability
Mixed
Simple Random
Systematic Random
Stratified Random
Cluster Sampling
Convenience
Quota
Purposive/ Judgment
Snowball
Multistage
Multiphase
Lot quality assurance (LQA)
LQA-Cluster sampling

Cluster Sampling



Non Probability Sampling

Advantages

- Pilot studies
- Better suited for exploratory research
- Study unknown traits
- Urgent public policy and decisions
- Reach out to inaccessible populations

Disadvantages

- Questions on representativeness of the population
- Biased selection
- Personal prejudice
- Limited generalizability
- Potential of overlooking subgroups

Non Probability Sampling

Brazilian Journal of Otorhinolaryngology 2022;88(S4):S163–S169

Brazilian Journal of
OTORHINOLARYNGOLOGY
www.bjorl.org

ORIGINAL ARTICLE

Feasibility of a snowball sampling survey to study active surveillance for thyroid microcarcinoma treatment among endocrinologists and surgeons of Brazil

I.M. Silva, T.Q. Nogueira, D.N. Couto et al.

- Snowball sampling strategy is feasible and able in reaching hard-to-reach groups, such as doctors of different specialties, in different work environments or in large countries.
- In Brazil, total thyroidectomy seems to be the most indicated treatment for thyroid papillary microcarcinoma.

KEYWORDS
Microcarcinoma;
Slow-risk thyroid cancer;
Active surveillance;
Survey

Abstract
Objectives: This study aims to investigate if a sampling method using virtual networks is feasible to survey AS adoption among this ''hard-to-reach'' population of Brazilian doctors.
Methods: An online piloted 11-point structured survey questionnaire (designed using Googleforms®) probed the actual treatment patterns for adult patients with PTMCs, including treatment decision-making nonoperative options, was undertaken between 10 November and 30 November 2020. Participants were reached by the mobile phone Application (APP) and a snowball sampling strategy was used to recruit a total of 4783 members (maximum number of potential reach), which is the total of doctors of the all 21 social media WhatsApp® groups.
Results: From a total of 4783 members (maximum number of potential reach), there were 657 (13.7%) doctors (actual reach) who clicked the web link of the questionnaire, out of whom 512 (10.7%) fully completed the online survey. Among the survey respondents, 361 were endocrinologists (70.5%) and 151 were surgeons (29.5%). Overall, for low-risk PTMCs in an elderly patient, 118 responders (23%) recommend AS, while 390 (76%) recommend immediate surgery as the management, including lobectomy (18.5%) and Total Thyroidectomy (58.2%). The present responders tended to recommend surgery for PTMCs that were located adjacent to the dorsal surface of the thyroid, were multiple, or raised the size during the follow-up.
Conclusion: Using snowball sampling strategy as an innovative route to conduct surveys was feasible and applicable but the rate of response was still very low. Our data also suggests the need to investigate if AS is embraced by Brazilian doctors.

Sampling in Retrospective studies

Open access Original research

BMJ Open

Retrospective study of cancer patients' predictive factors of care in a large, Hungarian tertiary care centre

Márton Koch,[1] Éva Szabó,[2] Csaba Varga,[1,3] Viktor Soós,[1] Lilla Prenek,[1] Lili Porcsa,[1] Szabolcs Bellyei,[4] Kyra Girán,[5] János Girán,[6] István Kiss,[6] Éva Pozsgai [6,7]

To cite: Koch M, Szabó É, Varga C, *et al*. Retrospective study of cancer patients' predictive factors of care in a large, Hungarian tertiary care centre. *BMJ Open* 2023;**13**:e070320. doi:10.1136/bmjopen-2022-070320

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2022-070320).

MK and ÉS contributed equally.

Received 18 November 2022
Accepted 27 March 2023

ABSTRACT

Objectives To identify predictive factors of multiple emergency department (ED) visits, hospitalisation and potentially preventable ED visits made by patients with cancer in a Hungarian tertiary care centre.

Design Observational, retrospective study.

Setting A large, public tertiary hospital, in Somogy County, Hungary, with a level 3 emergency and trauma centre and a dedicated cancer centre.

Participants Patients above 18 years with a cancer diagnosis (International Classification of Diseases, 10th Revision codes of C0000–C9670) who visited the ED in 2018, who had received their diagnosis of cancer within 5 years of their first ED visit in 2018 or received their diagnosis of cancer latest within the study year. Cases diagnosed with cancer at the ED (new cancer diagnosis-related ED visits) were also included, constituting 7.9% of visits.

Primary outcome measures Demographic and clinical characteristics were collected and the predictors of multiple (≥2) ED visits within the study year, admission to inpatient care following the ED visit (hospitalisation),

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This is a comparatively large, comprehensive study on patients with cancer visiting the emergency department (ED), where data regarding a wide range of parameters were collected.

⇒ The analysis of multiple aspects of ED visits made by patients with cancer is unique.

⇒ This is a retrospective study from a single centre; therefore, further studies are needed to confirm our results.

huge strain on the healthcare system. A large percentage of patients with cancer present to emergency departments (ED) due to a variety of medical conditions ranging from life-threatening conditions such as sepsis and unspecific symptoms such as pain or nausea.[2]

Since patients with cancer have been shown to use the ED more frequently than patients

Mixing probability and non probability sampling? Issues

Cancer Epidemiol Biomarkers Prev. Author manuscript; available in PMC: 2019 Sep 1.

Published in final edited form as: Cancer Epidemiol Biomarkers Prev. 2019 Jan 14;28(3):471–477. doi: 10.1158/1055-9965.EPI-18-0797

Weighting Non-probability and Probability Sample Surveys in Describing Cancer Catchment Areas

Ronaldo Iachan [1], Lewis Berman [2], Tonja M Kyle [3], Kelly J Martin [4], Yangyang Deng [5], Davia N Moyse [6], Deirdre Middleton [7], Audie A Atienza [8]

Author information · Article notes · Copyright and License information

PMCID: PMC6401258 NIHMSID: NIHMS1517558 PMID: 30642842

Abstract

Background: The Population Health Assessment initiative by National Cancer Institute (NCI) sought to enhance cancer centers' capacity to acquire, aggregate and integrate data from multiple sources, as well as to plan, coordinate, and enhance catchment area analysis activities.

Methods: Key objectives of this initiative are pooling data and comparing local data with national data. A novel aspect of analyzing data from this initiative is the methodology used to weight datasets from sites that collected both probability and non-probability samples. This article describes the methods developed to weight data which cancer centers collected with combinations of probability and non-probability sampling designs.

Results: We compare alternative weighting methods in particular for the hybrid probability and non-probability sampling designs employed by different cancer centers. We also include comparisons of local center data with national survey data from large probability samples.

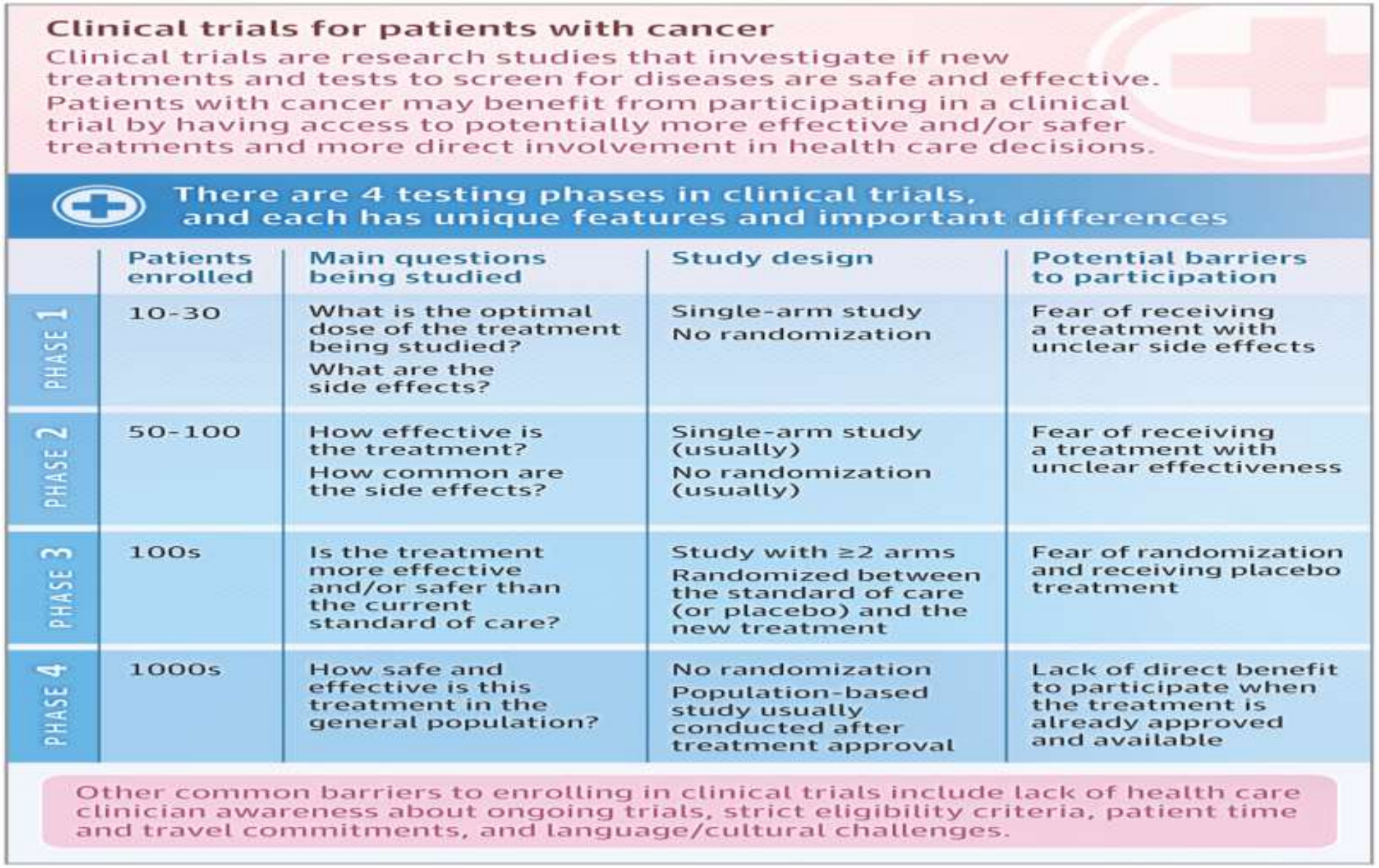
Corresponding Author: Tonja M. Kyle, ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850, 301-572-0820, 301-572-0986 (fax), Tonja.Kyle@icf.com.

Conflict of Interest: The authors declare no potential conflicts of interest.

Conclusions: This hybrid approach to calculating statistical weights can be implemented both within cancer centers that collect both probability and non-probability samples with common measures. Aggregation can also apply to cancer centers that share common data elements, and target similar populations, but differ in survey sampling designs.

Impact: Researchers interested in local versus national comparisons for cancer surveillance and control outcomes should consider various weighting approaches, including hybrid approaches, when analyzing their data.

Clinical Trials in Cancer Research



Clinical trials for patients with cancer

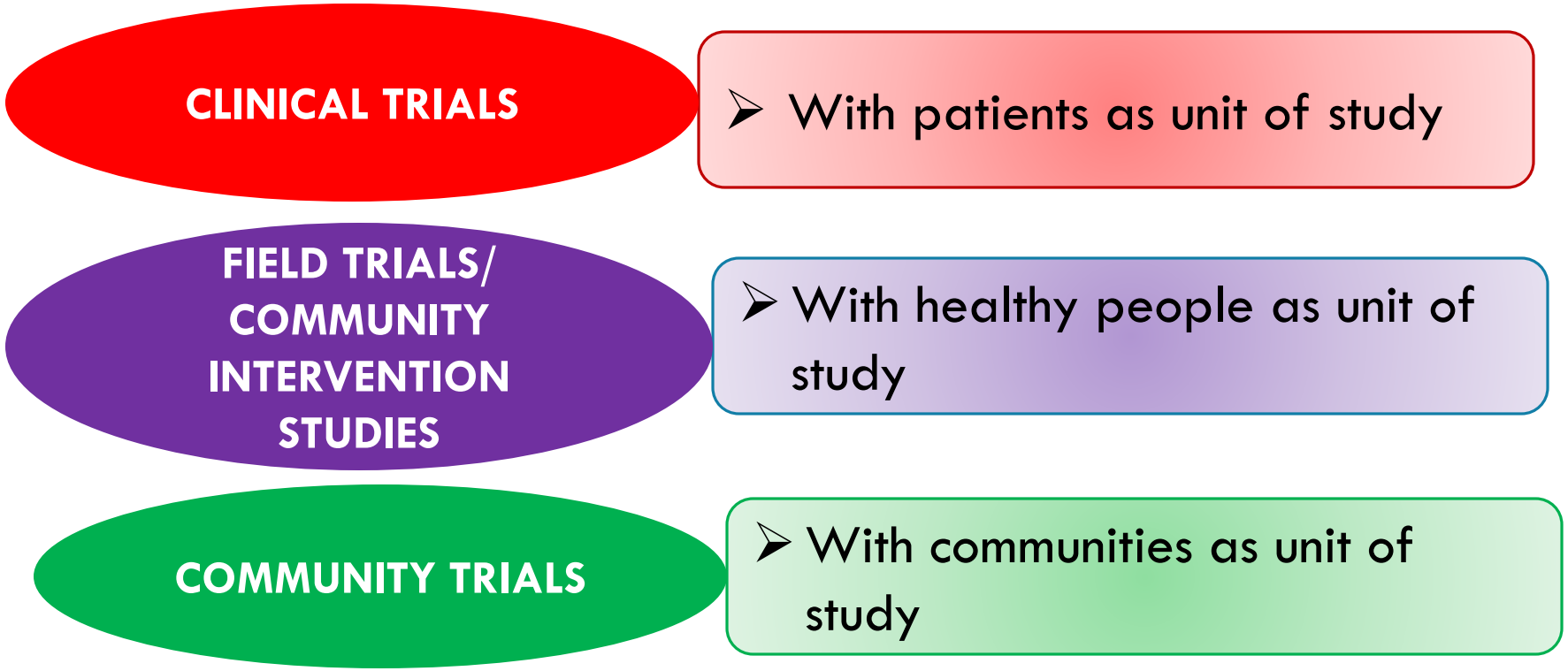
Clinical trials are research studies that investigate if new treatments and tests to screen for diseases are safe and effective. Patients with cancer may benefit from participating in a clinical trial by having access to potentially more effective and/or safer treatments and more direct involvement in health care decisions.

There are 4 testing phases in clinical trials, and each has unique features and important differences

	Patients enrolled	Main questions being studied	Study design	Potential barriers to participation
PHASE 1	10-30	What is the optimal dose of the treatment being studied? What are the side effects?	Single-arm study No randomization	Fear of receiving a treatment with unclear side effects
PHASE 2	50-100	How effective is the treatment? How common are the side effects?	Single-arm study (usually) No randomization (usually)	Fear of receiving a treatment with unclear effectiveness
PHASE 3	100s	Is the treatment more effective and/or safer than the current standard of care?	Study with ≥2 arms Randomized between the standard of care (or placebo) and the new treatment	Fear of randomization and receiving placebo treatment
PHASE 4	1000s	How safe and effective is this treatment in the general population?	No randomization Population-based study usually conducted after treatment approval	Lack of direct benefit to participate when the treatment is already approved and available

Other common barriers to enrolling in clinical trials include lack of health care clinician awareness about ongoing trials, strict eligibility criteria, patient time and travel commitments, and language/cultural challenges.

Experimental/Interventional studies



RCTs

- **Superiority trials**: new treatment intervention (drug, technique) is superior to (better than) the control condition
- **Non Inferiority trials:** non-inferiority trial is to show that treatment A is **not worse** than the treatment B
- **Equivalence trials**: the researcher aims to show that an intervention is **not too different from the comparator** (neither better nor worse by more than a predefined margin). It is defined “as a difference in performance of two interventions for which the patient will not detect any change in effect when replacing one drug by the other.”

RANDOMISATION:
STEPS IN A TYPICAL RANDOMISATION PROCESS

1. Sequence generation

- Generate the random allocation sequence by ranom procedures
- Participants should be assigned to comparison groups in the trial on the basis of a chance (random) process characterised by unpredictability

2. Allocation concealment

- Develop allocation concealment mechanism (such as numbered, identical bottles or sequentially numbered, sealed, opaque envelopes)

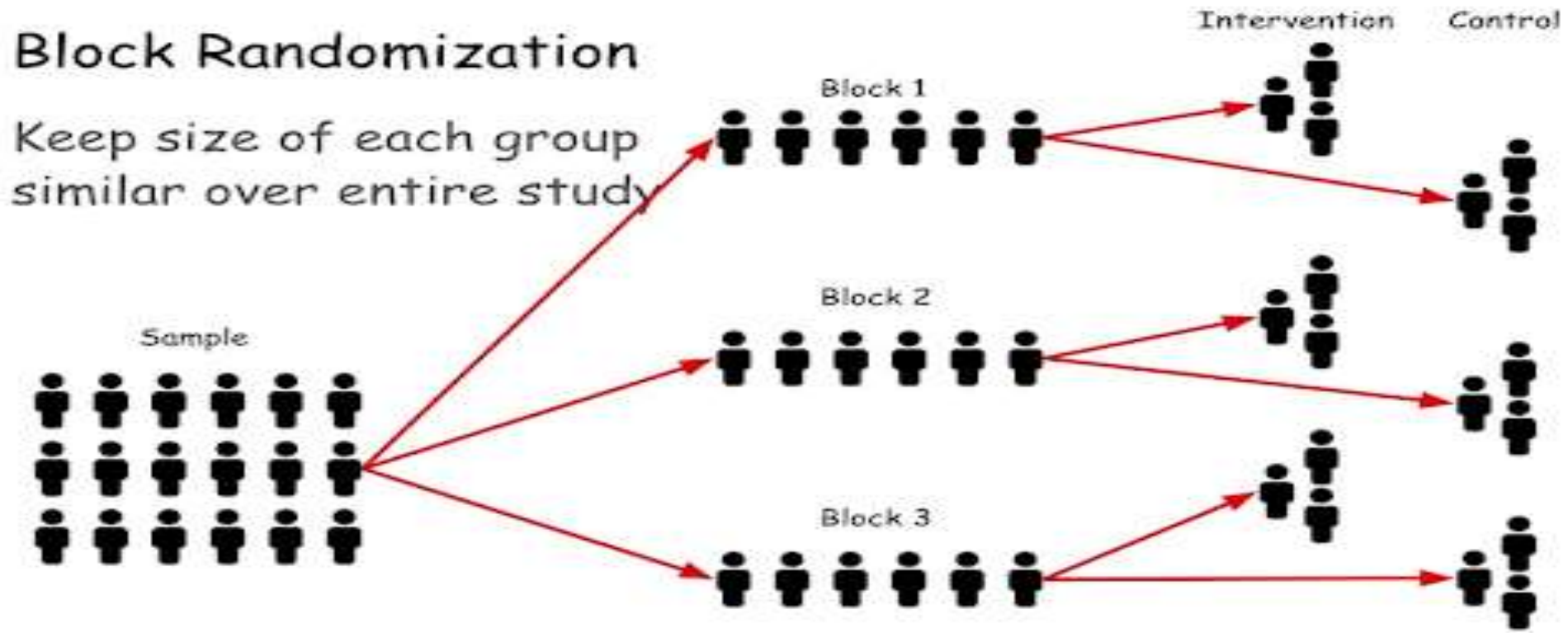
3. Implementation

- Enrol participants
- Assess eligibility
- Discuss the trial
- Obtain informed consent
- Enrol participant in trial
- Ascertain intervention assignment
- Administer intervention

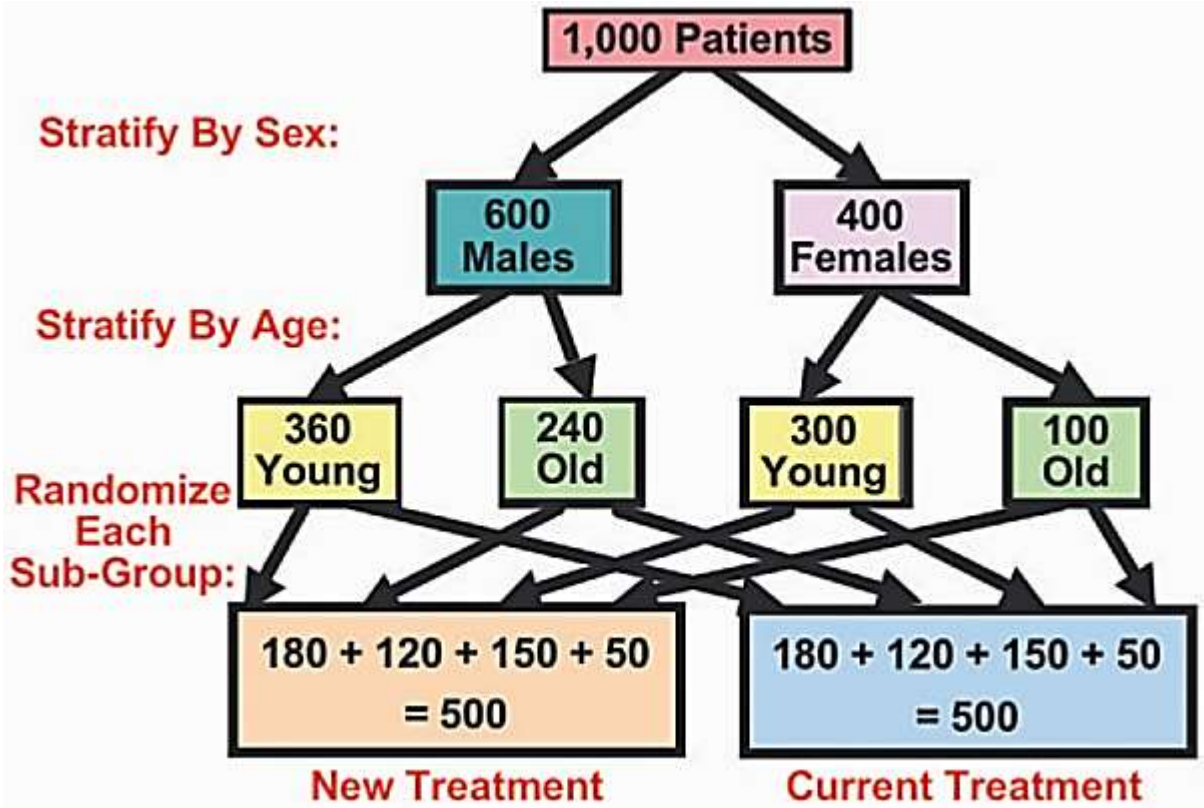
RANDOMISATION: SEQUENCE GENERATION

Type of randomisations		Examples
Simple	Randomization with no constraints to generate an allocation sequence.	*"We generated the two comparison groups using simple randomization, with an equal allocation ratio, by referring to a table of random numbers".*
Restricted	Generate a sequence to ensure particular allocation ratios to the intervention groups	
Blocked	Blocking ensures that the numbers of participants to be assigned to each of the comparison groups will be balanced within blocks of, for example, 5 in one group and 5 in the other for every 10 consecutively entered participants.	*"We used blocked randomization to form the allocation list for the two comparison groups. We used a computer random number generator to select random permuted blocks with a block size of eight and an equal allocation ratio".*
Stratified	Stratified randomisation is achieved by performing a separate randomisation procedure within each of two or more strata of participants (e.g., categories of age or baseline disease severity)	
Minimisation	Minimisation assures similar distribution of selected participant factors between study groups. It incorporates both the general concepts of stratification and restricted randomization	

Block Randomization



Stratified Randomization



RANDOMISATION:
Allocation concealment

Technique of ensuring that implementation of the random allocation sequence occurs without knowledge of which patient will receive which treatment, as knowledge of the next assignment could influence whether a patient is included or excluded based on perceived prognosis.

Ways to ensure concealment:

1. **Central randomization:** In this technique the individual recruiting the patient contacts a central methods center by phone or secure computer after the patient is enrolled.
2. **Sequentially numbered, opaque, sealed envelopes (SNOSE):** The envelopes receive numbers in advance, and are opened sequentially, only after the participant's name is written on the appropriate envelope.

Cluster RCT

original reports abstract

Cluster Randomized Trial to Facilitate Breast Cancer Early Diagnosis in a Rural District of Rwanda

Lydia E. Pace, MD[1,2]; Jean Marie Vianney Dusengimana, MPH[3]; Lawrence N. Shulman, MD[4]; Lauren E. Schleimer, MD[2]; Cyprien Shyirambere, MD[3]; Christian Rusangwa, MD[3]; Gaspard Muvugabigwi, BS[5]; Paul H. Park, MD[1,2]; ChuanChin Huang, ScD[1,2]; Jean Bosco Bigirimana, MPH[3]; Vedaste Hategekimana[6]; Vestine Rugema[5]; Aline Umwizerwa[5]; Nancy L. Keating, MD[1,2]; and Tharcisse Mpunga, MD[5]

PURPOSE Feasible and effective strategies are needed to facilitate earlier diagnosis of breast cancer in low-income countries. The goal of this study was to examine the impact of health worker breast health training on health care utilization, patient diagnoses, and cancer stage in a rural Rwandan district.

METHODS We conducted a cluster randomized trial of a training intervention at 12 of the 19 health centers (HCs) in Burera District, Rwanda, in 2 phases. We evaluated the trainings' impact on the volume of patient visits for breast concerns using difference-in-difference models. We used generalized estimating equations to evaluate incidence of HC and hospital visits for breast concerns, biopsies, benign breast diagnoses, breast cancer, and early-stage disease in catchment areas served by intervention versus control HCs.

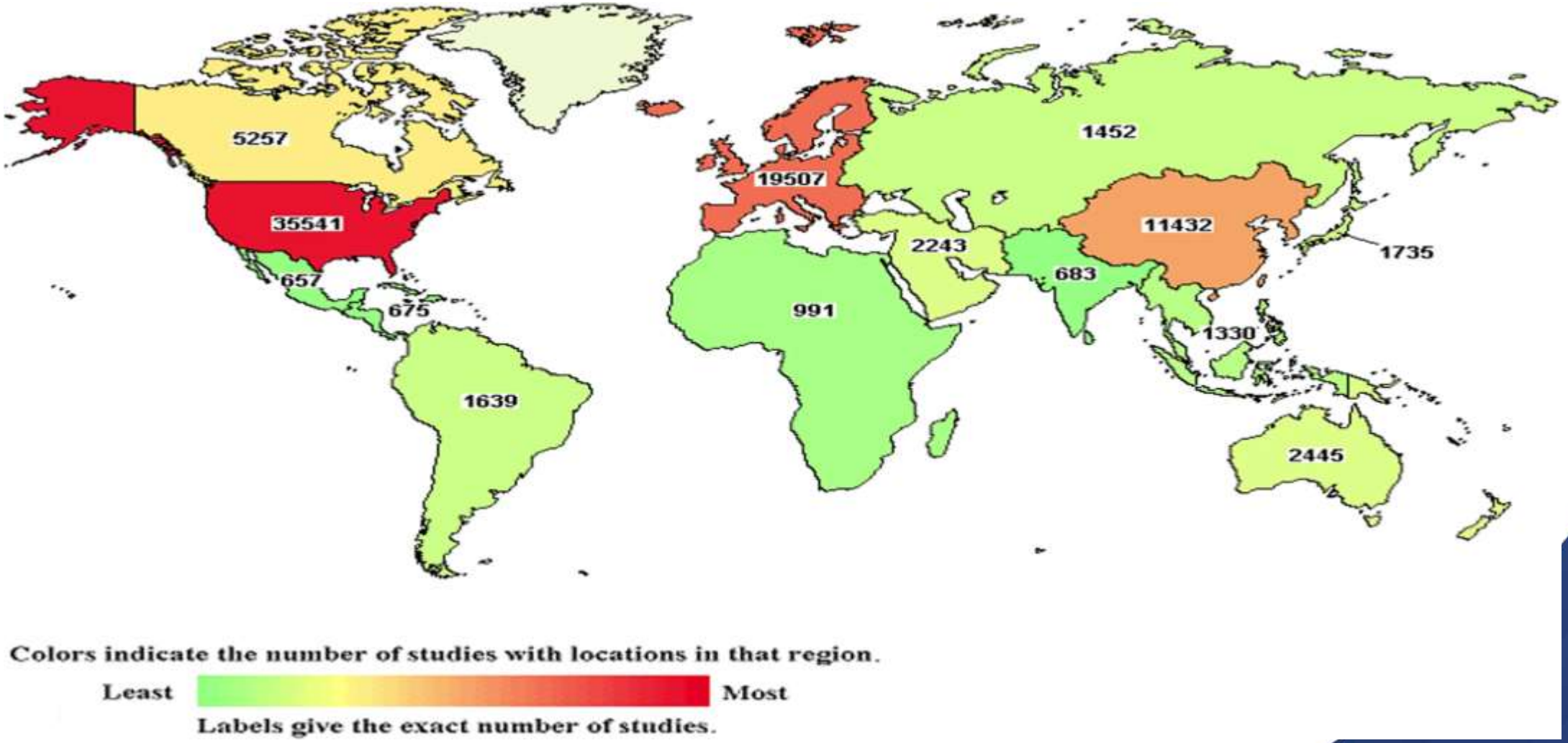
RESULTS From April 2015 to April 2017, 1,484 patients visited intervention HCs, and 308 visited control HCs for breast concerns. The intervention led to an increase of 4.7 visits/month for phase 1 HCs ($P = .001$) and 7.9 visits/month for phase 2 HCs ($P = .007$) compared with control HCs. The population served by intervention HCs had more hospital visits (115.1 *v* 20.5/100,000 person-years, $P < .001$) and biopsies (36.6 *v* 8.9/100,000 person-years, $P < .001$) and higher breast cancer incidence (6.9 *v* 3.3/100,000 person-years; $P = .28$). The incidence of early-stage breast cancer was 3.3 per 100,000 in intervention areas and 0.7 per 100,000 in control areas ($P = .048$).

CONCLUSION In this cluster randomized trial in rural Rwanda, the training of health workers and establishment of regular breast clinics were associated with increased numbers of patients who presented with breast concerns at health facilities, more breast biopsies, and a higher incidence of benign breast diagnoses and early-stage breast cancers.

Advantages:

- Preferred when the target of the intervention is a collective or system rather than a particular person, such as a patient
- when there is a significant potential for contamination in the study
- many interventions are naturally applied at a group level

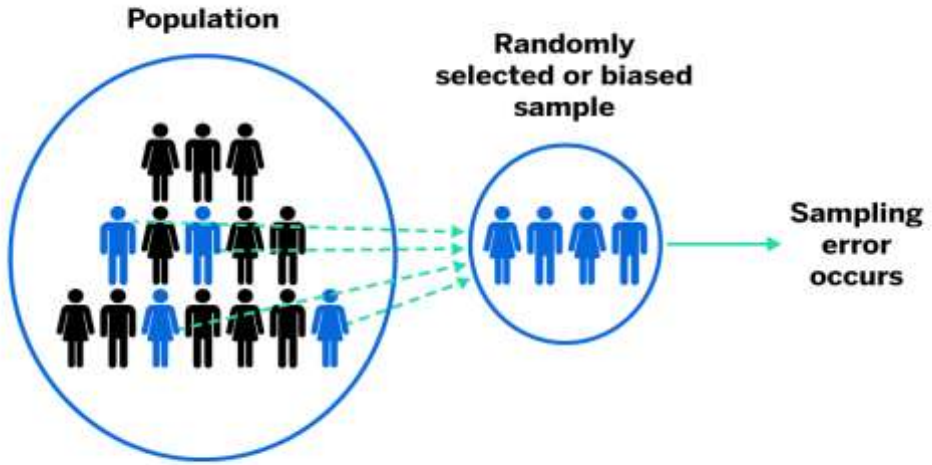
Landscape of Oncology clinical trials



(Source: U.S. National Library of Medicine. https://clinicaltrials.gov/ct2/results/map?cond=Oncology&map=

Sampling Errors

- Occurs due to non representativeness of the sample selected for observation.
- It reflects the difference between a result derived from a sample study and the “true value” obtained if the entire target population was studied.



Sampling errors- Types

Biased : When the selection of sample is based on the personal prejudice or bias of the investigator then the results are prone to bias errors.

Unbiased: Unbiased Errors arise due to a chance, i.e. the investigator has not intentionally tampered with the sample

Reasons:

- Faulty selection of sampling method.
- Faulty demarcation of sampling units.
- Variability of the population which has different characteristic.
- Substituting one sample for other sample due to difficulties in collecting the sample.

Steps for minimizing sampling errors

- Increasing sample size
- Dividing the population into groups.
- Random selection, results in the elimination of bias.
- Performing an external record check.

References

- Essential Medical Statistics Betty Kirkwood and Sterne
- Statistical Methods by SP Gupta
- Basic & Clinical Biostatistics by Dawson and Beth
- Wells JC, Sharma S, Del Paggio JC, et al. An Analysis of Contemporary Oncology Randomized Clinical Trials From Low/Middle-Income vs High-Income Countries. *JAMA Oncol*. 2021;7(3):379–385. doi:10.1001/jamaoncol.2020.7478
- Abu Rous F, Desai A, Singhi EK. Clinical Trials for Patients With Cancer. *JAMA Oncol*. 2024;10(3):416. doi:10.1001/jamaoncol.2023.5778
- Gao P, Chen J, Hong Z,. Landscape of cancer clinical trials in India–a comprehensive analysis of the Clinical Trial Registry-India. The Lancet Regional Health-Southeast Asia. 2024 May 1;24.

Thank You