Volume 24, Number 2 September 2015

ISMS BULLETIN

Official Publication of the

Indian Society for Medical Statistics (ISMS)

Editorial Office
Department of Medical Biometrics and Informatics (Biostatistics)
Jawaharlal Institute of Postgraduate Medical Education and Research,
Puducherry – 600 006.

Phone: + 91- 4132296185, 6186 **Mail:** ismsbulletin2013@gmail.com

ISMS Bulletin[$] Editors

(1986 to 2012)

T. Krishnan, ISI, Calcutta (1986-93)[#]

Arvind Pandey, IIPS, Mumbai (1994-98)

P. Venkatesan, TRC, Chennai (1999-2003)

N. K. Tyagi, MGMC, Wardha (2004-05)

R. M. Pandey, AIIMS, New Delhi (2006-07)

N. S. K. Nair, Manipal University, (2008-12)

[$] Till 1995, it was "**ISMS Newsletter**" and **1996** onwards got upgraded as "**ISMS Bulletin**".
Since the year **2006,** only electronic-version instead of print-version is officially being circulated.

[#] Initial three issues (1986-87) were coordinated by **B. N. Mukherjee**, ISI Calcutta.

Note: No Publications were made during the years 2002 to 2005 & 2007, 2009 and 2012.

(*Chief*) **Editor** (2013–2017)

Ajit Sahai

Professor of Biometrics & Head,

Department of Medical Biometrics and Informatics,

JIPMER, Puducherry–600 006.

Website Address: http://www.isms-ind.org/ *** E-Mail:** ismsbulletin2013@gmail.com

Contents

Message from the Editor

Ajit Sahai
E-mail ID: *ajit.sahai@gmail.com*

It is an exciting event in favour of the Society that the first ever professional editorial has been written by his Excellency, Professor C. R. Rao and his team of internationally renowned Statisticians.

Thrilled with his blessings, we decided to dedicate this issue of the Bulletin to his global recognition and the untiring leading role continued to be played by him even at the age of ninety five (95) years. In case the next ISMS conference is finalized to be held at ISI Kolkata, we must ensure to celebrate his presence back at his own institution.

I express my gratitude to the contributors of original papers and articles to this issue of the Bulletin. Even at the cost of repetitions, I personally invite senior members of the Society for submission of technical notes, review articles as well original papers and other useful information.

Though the Bulletin is gradually getting set to improve, there is an urgent need to form a core-group to restart 'book reviews' and 'review of journal articles' as well to write professional editorials. We expect from our earlier Presidents, Editors, General-Secretaries, ISMS Fellows & other awardees to come forward to write at least their brief overview of decades of professional experiences gained by them.

I may like to request further to our esteemed members to share on regular basis the feedback information on the seminars and workshops organized by them during the block of each six months' period preceding to the publication of the bulletin and also the details of the activities planned to be organized in future six months' period.

We would continue to publish the series on distinguished 'ISMS-Fellows' of international imminence as well complete the other series on 'Biostatistics Departments in Medical Institutions in India' before starting similar series on health institutions in India. This issue of the bulletin has included JIPMER, Pondicherry, despite the fact that the three premier medical institutions are still left due to inordinate delays in reporting. These institutions are AIIMS, New Delhi, SGPGIMS, Lucknow and KLE University, Belgaum (*Belagavi*).

However, it is very much desirable to initiate another series on Biostatistics Departments in International Medical Schools. The consolidated information must prove to be quite useful to young researches. But the question is who is going to volunteer help to the Editor?

Also, we must think very seriously to have 'chief or managing', 'associate or joint' and 'regional or assistant editors' apart from continuing to construct international advisory board and the editorial board requesting elite group of dedicated biostatisticians.

I am very concerned with the fact that now the time has come for me to request more and more participation of young minds to take over from me the responsibility of sharing editing management at the earliest as I have already completed three years.

We shall discuss the modalities of bringing out further improvements and enhancements in the professional standards of the Bulletin during forthcoming annual conference of the Society at KLE University, Belgaum *(Belagavi)*, Karnataka (13^{th} to 16^{th} October, 2015 – conference website http://www.isms-kle-india.org).

In accordance with the advice of some of the editorial board members we have applied for getting the recognition by International Standard Serial Number (ISSN) – India office for Online ISMS Bulletin (Registration No - 3993).

Message from the President

'On Road-to-Health'

Abhaya Indrayan
E-mail ID: *a.indrayan@gmail.com*

About a year ago when I took over as President, my message was 'fasten your seat belt'. Have we made any progress since then on this promise?

First we are grateful to our Editor, Dr Ajit Sahai, for taking so much pains and raise the level of our Bulletin to this height. Sometimes I feel this is becoming more than we can digest. This issue has an Editorial with legendry Prof CR Rao as one of the authors. The Bulletin has been carrying articles of substantive nature for the past couple of years, and is being regularly published.

Now about recent achievements, the Society formed three committees last year. These have done considerable work as per the mandate with spectacular results.

The first is the Constitution and Awards Review Committee with Dr BL Verma as Chair and Dr A Indrayan, Dr Ajit Sahai and Dr Anil C Mathew as members. Dr GD Shukla was coopted as member. This committee has worked immensely hard and has submitted its report on revision of the Constitution. As of date, this revision is under review by our Executive Council, and will be presented to the General Body at ISMS-group. Considerable work has been done for review of ISMS awards also, and we hope this report will also be ready soon.

The second committee under the Chair of yours truly is on Biostatistics Education in Medical and Allied Colleges. The other members of this committee are Dr BL Verma, Dr Ajit Sahai with Dr CM Pandey and Mr Manoj Diwakar as coopted members. The Committee literally burnt midnight oil to review the present structure of courses prescribed by various Councils, sieve the biostatistics and research methodology components as well as teaching and other infrastructure facility currently prescribed, and in framing the new ones that can meet the aspirations of the members of our Society. This committee has finalized its recommendations for bachelor's, master's and other courses in medical colleges, pharmacy colleges, homeopathic colleges and ayurvedic colleges. These recommendations have been approved by the Executive Council (EC) as well, and placed on our website for any comments by the members. We hope to communicate these to the respective Councils soon.

The third committee is for pursuing the registration of the Society. This is driven by Dr AK Bansal. It was a herculean task to get hold of all the documents for the past three years that the rules require, and to get signature of all the members of the EC by sending the same paper by speed post to all by turn. Some EC members returned the signed document quickly and safely packed in a hard folder but some took long time and multifolded the paper. Anyway, that is well that ends well, and Dr Bansal has submitted all the required papers to the Registrar.

You will note that all these issues were pending for decades, and seem to be reaching to the logical end now with the cooperation of all of you.

Our General Secretary, Dr Anil Mathew informs that he has now prepared a hand-written register of members. I took the task of onto myself to complete the electronic register and this was completed sometime ago, and is being regularly updated. Dr Antonisamy is looking after the website very well and we now seem to have a stable website.

Among other recent achievements are getting PAN, forming ISMS-group on Goggle, and endowment wise income and expenditure statement. We are now in frequently consulting the EC through email and seeking inputs from the General Body. Hopefully we are marching on the road-to-health.

Website: http://indrayan.weebly.com

Editorial

What is Medical Statistics and how is it useful?

Ranajit Chakraborty[1], Dabeeru C. Rao[2], and C. Radhakrishna Rao[3]

[1]Center for Computational Genomics, Institute of Applied Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Room - CBH-249, Fort Worth, Texas 76107, USA

[2]Division of Biostatistics, Washington University School of Medicine in St. Louis, Box 8067, 660 S. Euclid Ave., St. Louis, MO 63110, USA

[3]CRRAO AIMSCS, Hyderabad Univ.Campus, Prof. C.R.Rao Road, Hyderabad, 500046

What is Medical Statistics and how is it useful?

1. What is Medical Statistics? – A brief introduction

'Medical statistics' has a dual meaning. First, it is a branch of statistical science for analysis and interpretation of biomedical data. Second, it also refers to compilations of medical data; collected, compiled, and released by Governmental and/or Health agencies for various uses, including public health.

This editorial briefly reviews the current state of knowledge and use of medical statistics as a scientific discipline. In Section 2, we outline how medical statistics as a discipline relates to and differs from other disciplines of biomedical sciences, such as epidemiology, public health, medical forensics, bioinformatics, computational biology, and clinical research. Starting with a brief overview (in Section 3) of the Indian Society of Medical Statistics' role in developing and promoting this discipline, the subsequent 5 sections (Sections 4 through 8) discuss current issues in medical statistics that are emerging due to advances in collection, storage, and interpretation of such data. Next with a short history of medical statistics as a data compilation subject, we illustrate how governmental and health agencies use medical statistics data for uses in public health and other health service areas (Section 9). In Section 10, we discuss the role of medical statistics as a discipline as well as a data resource in the context of dealing with some societal problems (such as in DNA forensics and missing person identification in mass disaster cases) and in Section 11, we end with the potential utility of medical statistics in 'evidence-based medicine' or 'precision medicine' which is being projected as the current paradigm of prevention, control, and treatment of medical problems.

2. Medical Statistics as a Discipline – How it started, its identity/distinction with Biostatistics, Epidemiology, Bioinformatics, Computational Biology, etc.

The Oxford dictionary of statistical terms (Dodge, 2003) describes the discipline of Medical Statistics as a subject that deals with applications of statistics to medicine and health sciences, including epidemiology, public health, forensic medicine, and clinical research. Medical statistics has been a recognized branch of statistics in the United Kingdom for more than 40 years but the term has not come into general use in North America, where the broader term 'biostatistics' is more commonly used. However, "biostatistics" relates to all applications of statistics to biology. In olden times, it encompassed the design of biological experiments, especially in medicine, pharmacy, agriculture and fishery. Biostatistical investigations, thus, dealt with collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from the results. A major branch of this is medical biostatistics (Indrayan 2012), which is exclusively concerned with medicine and health. Kirkwood (2003) describes Medical Statistics as "the science of summarizing, collecting, presenting and interpreting data in medical practice, and using them to estimate the magnitude of associations and test hypotheses" is almost indistinguishable from the above connotation of Medical Biostatistics. It has a central role in medical investigations. It not only provides a way of organizing information on a wider and more

formal basis than relying on the exchange of anecdotes and personal experience, but also takes into account the intrinsic variation inherent in most biological processes." (Kirkwood, 2003). This definition obviously makes the discipline of Medical Statistics encompassing the goals and scope of epidemiology, public health, forensic medicine, and clinical research, these later subjects developed as a branch of medical statistics to serve the purpose of exploring principles and methodologies more focused in nature.

For example, epidemiology is generally described as the science that studies patterns, causes, and effects of health and disease conditions in defined populations. In that sense, epidemiology is a critical element of public health, with its applications guiding to build strategies of policy decisions and evidence-based practice of medicine by identifying risk factors for disease and targets for preventive healthcare. Epidemiologists focus on study design, collection, and statistical analysis of data, and interpretation and dissemination of results (including critical review of data from individual studies as well as meta-analyses and synthetic reviews of several studies on problems with common objectives). Epidemiology has helped develop methodology used in clinical research, public health studies, and, to a lesser extent, basic research in the biological sciences (Porta 2014). In addition, epidemiological studies being interdisciplinary to a large extent often rely on unique statistical principles and methods for understanding the disease processes by distinguishing cause-effect relationships from associations (e.g., correlation versus structural equation modeling).

The father of Western Medicine, the Greek physician Hippocrates is often credited as the originator of the concept of epidemiology (Grammaticos and Diamantis, 2008), since he used several concepts of modern epidemiology in his systematic organization of the field of clinical medicine. However, John Snow, an English physician of early nineteenth century (1813-1858) is considered as one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854. His findings also paved the way of using epidemiological findings for developing public health policies for disease prevention, since his work culminated in fundamental changes in the water and waste systems of London, which led to similar changes in other cities, and a significant improvement in general public health around the world.

Although epidemiological studies are generally bread and butter of public health departments and health institutions of hygiene and disease, in most universities Biostatistics and Epidemiology, together, form the core of training programs of studying medical statistics in public health. In this context, another recent development of formally subdividing the medical statistics field further is worth noting. As elaborated in the preface and 1st chapter of the Handbook of Statistics, Volume 28: Bioinformatics in Human Health and Heredity (Chakraborty, Rao and Sen , 2012), modern discipline of Medical Statistics is also now encompassed in two related but subtly different subjects, called Bioinformatics and Computational Biology. The biological Information Science and Technology Initiative Consortium (BISTIC) of the US National Institutes of Health provided definitions of these two subjects as follows. Bioinformatics covers "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data", while Computational Biology focuses on the "development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological , behavioral, and social systems" (http://www.bisti,nih.gov/). Chakraborty, Rao, and Sen (2012) wrote that "imbedded in these definitions, the distinctions are obvious; namely, studies under bioinformatics would emphasize applications of principles of information sciences and related technologies that would increase the power of understandability and utility of diverse and complex forms of life science data, while mathematical and computational tools are intended to be foci of attention in computational biology to address analytical and experimental issues of biological research in computational biology." Nonetheless, there is clear affinity between these two modern fields and Medical Statistics.

3. Brief History, Role, and Mission of the Indian Society of Medical Statistics

The Indian Society of Medical Statistics (ISMS) was established in 1983 as a registered society under Societies Registration Act XXI of 1860. It started and still continues as a scientific body consisting of biostatisticians, medical teachers, researchers, demographers and scientists from other related disciplines. For scientists working in medical field it is the only Society of its kind in the South East Asia region. Broad goals of ISMS are to provide a common platform to medical researchers, statisticians and computer specialists working in medical institutions to interact and benefit mutually towards development and application of biostatistical techniques.

The website of ISMS (www.isms-ind.org) lists the society's objectives as:

- To facilitate communication and build links among scientists engaged in medical research and teaching of Biostatistics
- To enhance the development and use of statistics in medicine and public health.
- To hold scientific meetings in the form of conference once a year for diffusion of modern techniques of biostatistics.
- To conduct scientific conferences, seminars, training workshops on health statistics in relation to specific national problems.
- To publish its own half-yearly Bulletin viz.; "ISMS BULLETIN", Journal, Proceedings and Abstracts of research papers for regular communications with its members
- The Society has more than five sponsored award schemes for the encouragement of its members towards research publications and one Fellowship award to bestow honor on its distinguished scientists who have made splendid contribution for the promotion of research in medical statistics.

Consequently, ISMS has been serving the field of medical statistics by fostering collaboration among scientists whose research and training goals match with the mission of this society. Through these activities ISMS is continuing to promote the discipline of Medical Statistics further by expanding its applications and by enhancing its translational value for solving problems of societal relevance in India and other South Asian countries.

4. A Brief Review of Medical Statistics

As prefaced by Rao et al. (2008) in "Epidemiology and Medical Statistics" (Handbook of Statistics Volume 27), data generation and statistical methodology research are joined at the hip:

"The history of statistics suggests two important lessons. First and foremost, methodological research flourished in the hands of those who were not only highly focused on problems arising from real data, but who were themselves immersed in generating highly valuable data, like the late Sir Ronald A. Fisher. Second, although theoretical statistics research can be nurtured in isolation, only an applied orientation has made it possible for such efforts to reach new heights. Throughout the history of statistics, the most innovative and path breaking methodological advances have come out when brilliant statisticians have confronted the practical challenges arising from real data. The computational revolution has certainly made many of the most difficult computational algorithms readily available for common use, which in turn made the data-centric approach to methodological innovation sustainable. That the data-centric approach enriched statistics itself is amply demonstrated by the varied applications of statistical methods in epidemiology and medical studies." As documented in "Epidemiology and Medical Statistics" (Rao et al. 2008), application of statistical methods in epidemiology and medical statistics is broad and varied. The edited volume included 27 chapters, each focusing on a major area, covering statistical challenges in biomedical research, issues relating to causal effects, statistical methods for biomarkers, survival analysis, cluster analysis, factor analysis, structural equation models, longitudinal data analysis, meta-analysis, and multiple comparisons. As discussed in some length, biomarker research is very important for exposure assessment, possible early disease diagnosis (before the onset of disease), as well as for the evaluation of possible therapies (e.g., Prentice, 2008). Rubin (2008) discussed methods for addressing one of the central issues in epidemiology and medical statistics, namely, how to infer causation especially in drug treatments, for example, whether the new drug really reduces heart disease.

As discussed by Looney and Hagan (2008),

analysis of biomarker data often requires statistical methods that are typically not covered in introductory statistics textbooks. Some biomarker studies need non-standard analysis methods especially because of challenges arising on account of non-normality of the biomarker data, heterogeneity, and censoring among other things. Also, typically, biomarker studies engage small sample sizes which do not lend themselves to statistical methods based on large sample approximations. For these reasons, Looney and Hagan (2008) emphasize using exact methods when available; they also recommend distribution-free and robust methods for many situations.

Likewise, topics of relevance to medical statistics have been covered in at least two other volumes of Handbook of Statistics. For example, in volume 8: Statistical Methods in Biological and Medical Sciences (Rao and Chakraborty, 1991). This volume encompasses discussions ranging from current trends of design and analyses of epidemiological studies, including sample size determination in clinical research (Breslow, 1991; Bock and Toutenburg, 1991), methodologies for detecting familial aggregation of biological traits (Rao, 1991) and genetic linkage for complex biomedical traits (Lathrop and Lalouel, 1991), estimation of biological relatedness from genetic data (Thompson, 1991) to evolutionary studies (Saitou 1991, Lessard, 1991, Chakraborty and Rao 1991). More recently, Volume 28 of the Handbook of Statistics (Rao, Chakraborty and Sen 2012) covered analytical bioinformatics methods (Pierre-Louis et al. 2012, Pinheiro et al. 2012, Ogasawara 2012, Bannerman-Thompson et al. 2012), Genetics and DNA Forensics (Chakraborty and Ghosh 2012, Gu 2012, Gupta and Ray 2012, Zhang et al. 2012, Bertoni et al. 2012, Guha et al. 2012) and current issues in epidemiology, including quality of life (Mesbah 2012, Tunes-da-Silva et al. 2012), epistemology and management of obesity (Chakraborty and Chakraborty 2012) that include both classical statistical as well as Bayesian inferential methods (Rodin et al. 2012).

5. *The General Quality of Statistics Applied in Medical Research*

There is substantive evidence that statistical methods are used widely and extensively in medical research. From time to time, it has been observed that the statistical standards are generally not consistently good. As Strasak et al. (2007) point out, a growing body of literature exhibit occurrences of statistical errors in many medical journals. However, there was no comprehensive study contrasting the top medical journals of basic and clinical science for recent practice in their use of statistics. Strasak et al. (2007) undertook such a study and their findings are reviewed here to highlight the general state of statistical quality in medical research.

They screened all original research articles in Volume 10, Numbers 1–6 of *Nature Medicine* (*Nat Med*) and Volume 350, Numbers 1–26 of *The New England Journal of Medicine* (*NEJM*) for their statistical content. Types, frequencies, and complexity of various applied statistical methods were systematically recorded. A 46-item checklist was used by the authors to evaluate the statistical quality for a subgroup of papers.

It was noted that 94.5 percent (95% CI 87.6–98.2) of *NEJM* articles and 82.4 percent (95% CI 65.5–93.2) of *Nat Med* articles contained inferential statistics. *NEJM* papers were significantly more likely to use advanced statistical methods ($p < 0.0001$). Statistical errors were identified in a considerable proportion of articles, although not always serious in nature. Documentation of applied statistical methods was generally poor and insufficient, particularly in *Nat Med*.

As compared to the 1983 review (Emerson and Colditz, 1983), a large increase in usage and complexity of statistical methods was observed for *NEJM* papers but not necessarily for *Nat Med* papers, as the results of the study indicate that basic science sticks with basic analysis. Since statistical errors seem to continue in the medical literature, only serious attention to statistical methodology can raise standards. We believe that every medical research journal should periodically (e.g., every 10 years) undertake review of the statistical quality of papers published in their journal by 3rd party statisticians and publish the results in their journals. Such practices will likely lead to ever improving quality.

6. *Quality of Medical Data versus Choice of Statistical Methods of Analysis*

It happens all too often where statistical analysts of

data, any data, argue whether the statistical methods used are the best possible, with optimum statistical power, and robust inference. What many overlook sometimes is whether the data at hand is of optimal quality and whether anything could be done to improve the quality of the data at hand. In typical clinical and epidemiological studies, it is possible to anticipate and do something about reducing potential quality issues [*quality assurance* (QA)]. For example, when measuring blood pressure (BP) in medical research studies, standard guidelines recommend that the subject be seated in a reclined chair for 5 minutes before taking BP readings. This QA measure controls variability due to measurement (error) by standardizing the measurement protocol. In the same vein, it is further recommended to take multiple BP readings (e.g., three), not just one. Taking the average of the 3 readings reduces its variance, thereby enhancing the quality of the data analyzed.

Despite good QA measures in place, errors do creep into data collected. For example, despite the QA measures discussed above, those recording the data sometimes make mistakes (such as adding an extra digit or dropping a digit, leading to bad outliers). Thus, it is also necessary to apply *quality control* (QC) procedures to data collected. Common procedures include filtering out extreme outliers (e.g., those outside 4 or 5 standard deviations from the average), data audits whereby copies of any data forms are obtained from the study center and double checked against the data set at hand (to identify any data entry errors), verifying that one's systolic BP is greater than his/her diastolic BP, etc.

As researchers with extensive QA and QC experience with medical data may argue, compromised data quality may hinder the quality of medical research more than how sophisticated the statistical methods used are. While the use of appropriate and valid statistical methods is essential, data quality is paramount to the integrity of biomedical research.

7. Statistical Issues in Genome-Wide Studies of Diseases

Extraordinary advances in genomics and computational capacity have revolutionized the approach to the study of human disease, thereby promoting biomedical research and the development of novel statistical methods at an unprecedented scale. In recent years, the advent of Genome-Wide Association Studies (GWAS) using millions of single nucleotide polymorphisms (SNPs) has achieved rapid progress in identifying hundreds of genetic variants (SNPs) associated with many common complex human diseases and disease-related traits including blood pressure and hypertension, lipids, and diabetes among others. Most of the identified genetic variants have small effect sizes and explain only small proportions of the trait variance. Recent emphasis on more sophisticated approaches involving gene-environment interactions is beginning to look more promising for identifying larger sets of genetic variants with larger expected effects. Interaction analyses are generally based on the joint model: $E[Y] = \alpha + \beta_g G + \beta_e E + \beta_{ge} GE$, where β_g and β_e, respectively, are the genetic (SNP) and environmental main effects and β_{ge} is their multiplicative interaction effect. If GWAS approaches lead to major discoveries in terms of identifying specific genes and their effects, it can help promote precision medicine.

GWAS studies have identified and resolved several statistical issues (e.g., Zhang and Liu, 2011; Yi, 2010; Kraft et al., 2007; Voorman et al., 2011; Thomas, 2010; Manning et al., 2011). Some of the statistical solutions are appropriate for the needs while others can benefit from additional improvements. These include: (a) How to optimally correct for multiple testing involving millions of genetic markers in the entire genome, not all of which are independent?; (b) What are the most powerful ways to test for interactions?; (c) What are the optimal statistical designs for synthesizing information from large numbers of disparate studies by striking a desirable balance between type I and type II errors? A few examples of real studies are briefly outlined below.

A Large-Scale Investigation of Gene-Age Interactions in Blood Pressure (Simino et al., 2014): Although age-dependent effects on blood pressure (BP) have been reported, Simino et al. (#614) were the first to systematically investigate these in a large-scale GWAS. They leveraged the infrastructure of three well-established consortia which are active in the genetic studies of BP (namely, CHARGE, GBPgen, and ICBP) to conduct a genome-wide search of genetic variants with

age-dependent effects on BP traits. In a two-stage design using 99,241 individuals of European ancestry, in their investigation we identified 20 genome-wide significant loci by using joint tests of the SNP main effect and SNP-age interaction. Several SNPs exhibited large age interactions, with opposite directions of effect in the young versus the old. The changes in the genetic effects over time were small but not negligible. A secondary analysis revealed 22 additional loci with evidence of age-specific effects (e.g., only in 20 to 29-year-olds). Age-dependent effects identified through novel statistical approaches can provide insight into the biology and temporal regulation underlying BP associations.

Electronic Medical Records and Genomics (eMERGE) Network Identifies a Strong Protective Genetic Variant for LDL Cholesterol in African Americans (Rasmussen-Torvik et al., 2012): The eMERGE Network, a consortium of five US institutions (Northwestern University, Marshfield Clinic, Mayo Clinic, Vanderbilt University, and the Group Health Cooperative, University of Washington, and Fred Hutchinson Cancer Research Partnership), has pursued high density GWAS of several disease traits compiled from electronic medical records (EMR). eMERGE investigators have worked collaboratively to develop algorithms for extracting phenotypes from EMRs for use in genetic research. Each site in the consortium conducted a GWAS on a specific phenotype derived from their EMR records. Members of the consortium then combined all genotyped samples across the network into one merged dataset and conducted further GWAS analyses by working across the Network to extract and harmonize phenotypes from all five sites' databases. Rasmussen-Torvik et al. (2012) presented analyses using a subset of African American individuals in the eMERGE consortium. One SNP in a highly studied lipid gene (rs7412 in APOE) was found to be strongly associated with LDL-C in alternative datasets based on medication information. A decrease of 20.0 mg/dL per minor allele was observed in one data set (unusually large effect). Use of median LDL-C extracted from EMR after exclusions for medications and comorbidities appears to have increased the percentage of trait variance explained by genetic variation which illustrates the benefits of very large EMR data sets despite the statistical issues involved in harmonizing the data.

8. Opportunities and Challenges with Big Data in Medical Statistics

Technological advances in many domains including the ease of generating data, the highly affordable storage and transfer media, and the massively parallel computing, have all made it possible and even led to the exponential growth in the volume of data generated; from kilobytes (KB; 10^3 bytes) to megabites (MB; 10^6) to gigabytes (GB; 10^9) to terabytes (TB; 10^{12}) to petabites (PB; 10^{15}) to exabytes (EB; 10^{18}) to zettabytes (ZB; 10^{21}) to yottabites (YB; 10^{24}) (Wikipedia). The EMC International Data Corporation (EMC/IDC) Digital Universe Studies indicates that in 2013, there was close to 4 ZB of information stored with that amount doubling every 2 years (Gantz and Reinsel, 2011), perhaps already 8 ZB by 2015. For a sense of the magnitude, a ZB is 10^{21} bytes; that could equate to stored data amounts of around several hundred CD-ROMS for each and every person on earth – man, woman and child (Frické, 2015). Perhaps 'big data' is already in need of a 'bigger' name?

Is more data always better? Not necessarily so especially if the quantity of data is inversely proportional to the quality of data. The sources, sizes, and quality of big data vary. Heuristically, one may consider four categories. Data collected through direct measurement such as laboratory measurements, genotyping, sequencing, and imaging data may be regarded as of the highest possible quality. Data collected through other traditional means such as surveys, questionnaires, and measurements in the clinic may be of second tier in data quality. Data extracted from semi-controlled environments such as the electronic medical records (EMR) may be regarded as of the third tier whereas all other direct capture data sets (such as telephone conversation records) may be regarded as the lowest possible quality. Many of these big data, including large scale genotype and sequence data, imaging data, and EMR data, are of particular relevance to medical statistics. They involve some outstanding statistical issues such as optimal methods of analyzing these voluminous data sets. In particular, the EMR data, which are subject to multiple sources of error/noise, may

benefit from a different modeling framework as compared to that for observational data.

As noted by Frické (2015), some of the big data sources may yield larger sample sizes at less cost and make it possible to do more extensive testing of theories. On the negative side, data-driven science may encourage passive data collection, as opposed to collecting data from experimental means, and consequently perhaps less sound statistical maneuvering of the data.

Big Data in the Context of Medical Statistics: Although the *Big Data Research and Development Initiative* of the US Government, launched in 2012, (www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf) has a broad aim of enhancing the speed of discovery in science and engineering, one of the objectives of the original announcement was to lower costs and improve health outcomes in the country. As a consequence, in the context of medical statistics, availability of large-scale genomic sequencing platforms, medical imaging devices of ultra-fine resolution, electronic medical records, and mobile health monitoring technologies prompted infusion of additional funds by the US National Institutes of Health in their initiative named as *Big Data to Knowledge* (MD2K) program. This has already resulted in opening promising avenues for understanding the onset and progression of diseases, and in exploring new therapeutic modalities and their speedy applications for improving heath and health care. Some examples of these Big Data generating activities in medical research in the USA are:

- *Human Connectome Project* (http://www.neuroscienceblueprint.nih.gov/connectome/) , which is now mapping neural pathways of human brain function with the objective of detecting abnormal brain circuits that would ultimately help in diagnosis and treatment of neurologic and psychiatric disorders,
- *Cancer Genome Atlas* (TCGA, http://cancergenome.nih.gov/), which is rapidly assembling comprehensive genomic deep sequencing data (both at DNA and RNA level) for various site-specific cancers, identifying differences between normal and malignant tissues of the same patients. TCGA data are already being routinely used in developing individual-specific treatment modality in many Cancer Centers of the USA.
- *Physionet* (http://www.physionet.org/), which is a research resource for characterizing complex physiological signals, which houses its bank of Big Data, called *Physiobank*. Together with its accompanied software tool, called *PhysioToolkit*, this publicly available resource (obtained through licensing) has already enhanced our understanding of cardiac rhythms and gait dynamics that would be helpful in understanding mechanisms of aging diseases influencing geriatric health. Data from the Physionet project are already being used for developing new biomarkers for Alzheimer's disease (AD) and monitoring progression of AD by MR imaging of amyloid perturbation in patients.

These examples of Big Data resources (or initiatives) in the context of medical statistics show that in spite of challenges of assembling and use of Big Data, in the context of medicine, the Big Data initiatives are already proving their utility of translating medical knowledge from laboratory to bed-side applications.

9. Medical Statistics as a Data Compilation

The data compilation component of Medical Statistics has gone through a major evolution due to wide-spread use of the internet and public access of databases involving medical data. The reference librarians at the US National Library of Medicine® (NLM®) made an important contribution towards this by developing the concept and lists of specific Medical Subject Headings (MeSH) terms. Using these MeSH terms as key words, search engines such as PubMed can now be readily used to find links to national as well as international data resources on specific diseases, health insurance, hospitals, medical procedures, and pharmaceuticals.

The web-site of NLM®, http://www.nlm.nih.gov/services/statistics.html, is an important source of finding the major databases on health and medical statistics and their frequently used purposes. The ones of special relevance to US listed in this web-site are:

- Agency for Health Research and Quality (AHRQ) for Health Care Utilization (HCUP)

and Medical Expenditures (MEPS)
- CDC National Center for Health Statistics (NCHS) for US health statistics, including community health status indicators
- Centers for Medicare & Medicaid Services (CMS) for spending, enrollment and claims data
- FedStats links to statistical information from over 100 federal agencies and programs, including the US Census bureau for population-related statistics
- NIH Institutes and Offices for disease specific statistical information starting points
- NLM MedlinePlus® Health Statistics topic page for links to selected health statistics sources; MedlinePlus Health topic pages for links to disease-specific statistics sources
- NLM National Information Center on Health Services Research and Health Care Technology (NICHSR) for finding and using Health Statistics for a self-study course on the subject and for links to Data Tools and Statistics, Health Services and Sciences Research Resources (HSRR), and more.
- Partners in Information Access for the Public Health Workforce for links to individual state statistics, county and local health data.
- US Census Bureau for vital health and other statistics.
- Injuries, Illnesses, and Fatalities (IIF) for annual reports on workplace injuries, illnesses, and fatal injuries.

Some of the International Health Statistics Sources, cited in the same web-site, are:

- World Health Organization (WHO) Global Health Observatory (GHO) for data, tools, analysis and reports by country or health topic.
- United Nations UNdata and Statistical Databases for links to Millennium Development Goals (MDG) Indicators and other international sources of data.
- Pan American Health Organization for Regional Core Data and other international statistical information.
- EU Health Statistical Reports for reports from Eurostat, the statistical office of the European Union.
- UK National Statistics for data from United Kingdom by topic or department for statistical release announcements.

These Medical Statistics data resources are not always simply to assist agencies or policy makers to document public health burdens of medical problems. For example, the US National Health and Nutrition Examination Survey (NHANES) database, maintained and curetted by the Centers of Disease Control and Prevention (CDC) is a program of studies designated to assess the health and nutritional status of children and adults in the US is an unique and valuable resource for epidemiological studies, since it combined data from interviews and physical examination, providing numerous opportunities to investigate the role of life style and other environmental factors on phenotypes of biomedical interest. The National coverage of subjects enrolled in NHANES database provide scope for establishing baseline estimates of prevalence of diseases or conditions of public health impact (e.g., cholesterol and lipid abnormalities, high blood pressure, and infection of Hepatitis C, HIV, and other sexually transmitted diseases), which are used for studying etiology of these conditions as well as for assessment of health disparity within US.

Likewise, the cancer mortality database, created and maintained by the Cancer surveillance program at the International Agency for Research on Cancer (IARC) is extracted from the World Health Organization (WHO) cancer database, and used for various purposes other than simple cancer surveillance. For example, this database has been used to explain the age pattern of cancer mortality in general as well as that of familial polyposis coli by multistage risk models of carcinogenesis in several studies (Chakraborty and Weiss 1988, Weiss and Chakraborty 1984).

Several medical statistics data resources have also been used to model effects of concerns about access to individualized public databases consisting of medical information. For example, protecting numerical data for data perturbation and imputation methods, which are recoverable for research purposes, but help in preserving data privacy, has been discussed by Sarathy and Muralidhar (2012) and Reiter (2012).

10. *Use of Medical Statistics in Integrating Topics of Societal Relevance with Basis Science: DNA Forensics, Genetic Counseling and Genetic Etiology of Diseases*

Analytical as well as laboratory methods of Medical Statistics have helped in integration of several seemingly unrelated subjects of research interest. For example, since late thirties of the last century genetic markers have been used for determining paternity of disputed children (Essen-Möller 1938). Since then, statistical logic behind paternity testing using genetic markers has been extended to determine biological relatedness (Thompson 1991) and for solving other DNA forensic cases, including missing person identification, and interpretation of DNA mixtures (Chakraborty and Deka 2006). These methodologies have been extended to include data on multiple persons typed from families (Ge et al. 2010). Interestingly, the same problem is relevant in the context of genetic counseling, when stated relationships between pedigree members are to be assessed from genetic data (Elston and Stuart 1971) or for studying co-segregation of genetic traits for linkage studies (Lathrop and Lalouel 1991).

Likewise, methods for detection and interpretation of DNA mixtures from crime scene samples (Budowle et al. 2009) have also statistical principles that mimic problems of separating malignant versus normal cells in diagnosis of cancer. Without discussions of such research under the unified forum of Medical Statistics the similarities of these problems would have remained un-noticed. Medical Statistics training programs in many schools have made it possible to show these types of similarities of seemingly unrelated problems increasing the translational value of medical statistics research.

11. *Medical Statistics in the era of Precision Medicine – present/future restructuring of medical statistics data for evidence-based medicine*

Finally, the recently proposed paradigm shift in initiatives called precision medicine (Collins and Varmus 2015) has also implications for furthering the use and training of Medical Statistics. Often also called evidence-based medicine, or individualized medicine, this paradigm calls for synthesis of patient data encompassing large-scale genomic knowledge (including data on genomic sequence, proteomics, metabolomics, and diverse cellular assays) with present and past data on health condition and life style features of patients to developing appropriate individual-specific prevention and treatment strategies for any medical condition. Obviously, this involves challenging tasks of assimilating medical statistics data of very large dimension (referred to earlier as big data) as well as novel technologies, termed as mobile health technologies, such as electronic health records.

While precision medicine is some distance away from becoming a norm of all medical diagnosis or treatment, considerable progress has been made in some contexts. For example, therapeutic modalities of cancer treatment is now very frequently designed considering patient-specific data utilizing the Cancer Genome Atlas (TCGA) database assembled by the US National Institutes of Health (https://tcga-data.nci.nih.gov/tcga/). Likewise, individual-specific radiation exposure, repeatedly faced by children due to common frequently used computed tomography (CT)-based screening devices, has come into heated debate in recent years (Brenner and Hall 2007; Boice 2015), whose resolution appears to involve the concept of 'reverse causation' derived from studies done in Medical Statistics (Chakraborty 2001).

These examples indicate that the discipline of Medical Statistics as well as Medical Statistics as a resource of databases will have to evolve with changes in technology in the field of Biomedical Sciences and Computational tools to enhance its translational value not only in medicine but also in other applications of societal value.

12. *Conclusions*

The discussions above show that the scope and mission of Medical Statistics has evolved over the years. Quantitative assessment of biomedical data has become refined not only because of advances and ease of collection of precise data, systematic archival, storage, retrieval and analyses of such data has progressed to an extent that was inconceivable even a decade back. Big data is no longer a problem either for collection, or for analyses and visualization or dissemination of the results. Bioinformatic tools for data protection and limiting access to authorized personnel increased the

establishment of large medical data resources and their use in collaborative platforms not only for researchers, but also for agencies and authorities who can use such data for policy making to address medical and public health problems. Societal applications of medical statistics have also helped in encompassing related disciplines to this field to make today's training, research, and applications of medical statistics truly multidisciplinary.

The Indian Society of Medical Statistics (ISMS), with its 32 years of existence, has also broadened its mission and activities in conjunction of such ongoing changes in Medical Statistics. With increased numbers of training centers of Medical Statistics in the country, and with increased international collaborations through conferences and workshops, ISMS should continue such activities to ensure that researchers, students, and policy makers can utilize the newly evolving tools of conducting medical statistics studies and be active participants in international projects involving large scale data analyses and applications. The society should also take active part in modernizing the contents of medical statistics training programs in India as well as in neighboring countries to enhance the translational value of research and applications of medical statistics to solve public health as well as societal problems that are relevant in South and South-East Asia.

References:

Bannerman-Thomson, H., Rao, M. B., and Chakraborty, R. (2012) Multiple Testing of Hypothesis in Biomedical Research. In*: Handbook of Statistics, Vol. 28 – Bioinformatics in Human Health and Heredity* (Rao, C.R., Chakraborty, R., and Sen, P. K. eds.), Chapter 8, pp.201-238, Elsevier, Amsterdam.

Bertoni, B., Velazquez, T., Sans, M., and Chakraborty, R. (2012) A Molecular Information Method to Estimate Population Admixture. In*: Handbook of Statistics, Vol. 28 – Bioinformatics in Human Health and Heredity* (Rao, C.R., Chakraborty, R., and Sen, P. K. eds.), Chapter 13, pp. 339-353, Elsevier, Amsterdam.

Boise JD, Jr. Radiation epidemiology and recent paediatric computed tomography studies. Ann ICRP 2015 June; 44(1 Suppl), 236-248.

Bock J and Toutenburg H. Sample size determination in clinical research. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* (pp. 515-538), Elsevier B.V., the Netherlands.

Brenner DJ and Hall EJ. Computed tomography – an increasing source of radiation. New Eng Jour Med, 2007, 357; 2277-2284.

Breslow N. Statistical design and analysis of epidemiologic studies: Some directions of current research. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* (pp. 125-144), Elsevier B.V., the Netherlands.

Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: Defining the relevant features for guidelines for tha assessment of mixed DNA profiles in forensic casework. Jour Forens Sci, 2009; 54, 810-821.

Chakraborty R. A rooster crow does not cause the sun to rise: Book review of: Causality: Models, Reasoning and Inference (J. Pearl). *Hum Biol.* 2001,**73**:621-624.

Chakraborty BM and Chakraborty R. Bioinformatics of Obesity. (Rao CR, Chakraborty R and Sen PK, Eds) *Bioinformatics in Human Health and Heredity (Handbook of Statistics*,2012. *Vol. 28*, Chapter 17, pp. 433-477, Elsevier, Amsterdam.

Chakraborty R and Deka R. DNA forensics: A population genetic and biological anthropological perspective. *Physical (Biological) Anthropology* (Rudan P, Ed.), *in Encyclopedia of Life Support Systems (EOLSS)*, 2006. Develeoped under the auspices of the UNESCO, Eolss Publishers, Oxford, UK (http://www.eolss.net).

Chakraborty R and Rao CR (1991) Measurement of genetic variation for evolutionary studies. In: *Handbook of Statistics, Volume 8: Statistical Methods for Biological and Medical Sciences*. (C.R. Rao and R. Chakraborty, eds.). New York: North-Holland, pp. 271-316.

Chakraborty S and Ghosh M. Applications of Bayesian neural networks in prostate cancer studies. Rao CR, Chakraborty R and Sen PK

(Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)*

Chakraborty R and Weiss KM (1988) Age-specific risks for cancer as determined by multi-stage models of carcinogenesis. *Statistics in Medicine.* (T. Krishnan, ed.). Bombay, India: Himalaya Publishing House, pp. 64-91.

Collins FS and Varmus H. A new initiative on precision medicine. New Eng Jour Med 2015, 372; 793-795.

Dodge Y (2003) *The Oxford Dictionary of Statistical Terms*, OUP. ISBN 0-19-850994-4

Elston RC and Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered 1971, 523-542.

Emerson JD and Colditz GA. Use of Statistical Analysis in the *NewEngland Journal of Medicine. New England Journal of Medicine. 1983;* 309, 709–713.

Essen-Möller E. Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis; theortische Grundlagenn. Mitt anthrop Ges 1938; 68, 9-53.

Frické M. Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology*. 2015; 66(4), 651–661.

Ge J, Budowle B and Chakraborty R. DNA identification by pedigree likelihood ratio with population substructure and mutations. *Investigative Genetics* 2010 Oct 4;1 (1):8, http://www.investigativegenetics.com/content/1/1/8 [PMID: 21092343].

Gu X. Statistical methods for detecting functional divergence of gene families. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 263-272), Elsevier B.V., the Netherlands.

Gupta M and Ray S. Sequence pattern discovery with applications to understanding gene regulation and vaccine design. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 273-307), Elsevier B.V., the Netherlands.

Gantz J and Reinsel D. (2011). Extracting value from chaos. Retrieved from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Grammaticos PC, Diamantis A (2008). Useful known and unknown views of the father of modern medicine, Hippocrates and his teacher Democritus. *Hell J Nucl Med* **11** (1): 2–4.PMID 18392218.

Guha S, Ge J and Chakraborty R. (2012) Effects of Inclusion of Relatives in DNA Databases: Empirical Observations from 13K SNPS in Hap-Map Population Data. In*: Handbook of Statistics, Vol. 28 – Bioinformatics in Human Health and Heredity* (Rao, C.R., Chakraborty, and Sen, P. K. eds.), Chapter 14, pp. 355-366, Elsevier, Amsterdam.

Indrayan A (2012). *Medical Biostatistics*. CRC Press. ISBN 978-1-4398-8414-0.

Kirkwood, Betty R. (2003). *Essential medical statistics*. Blackwell Science, Inc., 350 Main Street, Malden, Massachusetts 02148–5020, USA: Blackwell. ISBN 978-0-86542-871-3.

Kraft P, Yen YC, Stram DO, Morrison J, and Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Human Heredity*. 2007; 63: 111-9.

Lathrop GM and Lalouel JM. Statistical methods of linkage analysis. Rao CR, Chakraborty R (Eds.) , *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* 1991(pp 81-123). Elsevier B.V., the Netherlands.

Lessard S. Stastistical models for sex-ratio evolution. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* 1991 (pp. 347-372), Elsevier B.V., the Netherlands.

Looney SW and Hagan JL. (2008). Statistical Methods for Assessing Biomarkers and Analyzing Biomarker Data. Rao CR, Miller JP, Rao DC (Eds.), *Epidemiology and Medical Statistics (Handbook of Statistics Volume 27)* (pp. 109 – 147). Elsevier B.V., the Netherlands.

Manning AK, LaValley M, Liu CT, Rice K, An P, Liu Y, Miljkovic I, Rasmussen-Torvik L, Harris TB, Province MA, Borecki IB, Florez JC, Meigs JB, Cupples LA, and Dupuis J. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. *Genetic Epidemiology*. 2011; 35(1): 11-8. doi: 10.1002/gepi.20546.

Mesbah M. Measurement and analysis of quality of

life in epidemiology. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 369-400), Elsevier B.V., the Netherlands.

Miquel Porta (2014). *A Dictionary of Epidemiology* (6th ed.). New York: Oxford University Press. ISBN 978-0-19-997673-7.

Ogasawara H. Asymptotic expansions of the distributions of the least square estimators in factor analysis and structural equation modeling. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 163-200), Elsevier B.V., the Netherlands.

Pierre-Louis BJ, Suchindran CM, Chen P-L, Cole SR and Morrison CS. A statistical appraisal of biomarker selection methods: Applicable to HIV/AIDS research. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 99-128), Elsevier B.V., the Netherlands.

Pinheiro A, Pinheiro HP and Sen PK. The use of Hamming distance in bioinformatics. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* (pp. 129-162), Elsevier B.V., the Netherlands.

Prentice RL (2008). Statistical Methods and Challenges in Epidemiology and Biomedical Research. Rao CR, Miller JP, Rao DC (Eds.), *Epidemiology and Medical Statistics (Handbook of Statistics Volume 27)* (pp. 1 – 27). Elsevier B.V., the Netherlands.

Rasmussen-Torvik LJ, Pacheco JA, Wilke RA, Thompson WK, Ritchie MD, Kho AN, Muthalaqu A, Hayes MG, Armstrong LL, Scheftner DA, Wilkins JT, Zuvich RL, Crosslin D, Roden DM, Denny JC Jarvik GP, Carlson CS, Kullo IJ, Bielinski SJ, McCarty CA, Li R, Manolio TA, Crawford DC, Chisholm RL. High Density GWAS for LDL Cholesterol in African Americans Using Electronic Medical Records Reveals a Strong Protective Variant in APOE. *Clinical and Translational Science*. 2012; 5(5):394-399.

Rao CR and Chakraborty R (1991). *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)*. Elsevier B.V., the Netherlands.

Rao CR, Chakraborty R and Sen PK (2012) *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)*. Elsevier B.V., the Netherlands.

Rao CR, Miller JP, and Rao DC. (2008). *Epidemiology and Medical Statistics (Handbook of Statistics Volume 27)*. Elsevier B.V., the Netherlands.

Rao DC. Statistical considerations in applications of path analysis in genetic epidemiology. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* (pp. 63-80), Elsevier B.V., the Netherlands.

Rubin DB (2008). Statistical Inference for Causal Effects, With Emphasis on Applications in Epidemiology and Medical Statistics. Roa CR, Miller JP, Rao DC (Eds.), *Epidemiology and Medical Statistics (Handbook of Statistics Volume 27)* (pp. 28-63). Elsevier B.V., the Netherlands.

Simino J, Shi G, Bis J, Chasman DI, Ehret GB, Gu X, Guo X, Hwang SJ, Sijbrands E, Smith AV, Verwoert GC, Bragg-Gresham JL, Cadby G, Chen P, Cheng CY, Corre T, de Boer RA, Goel A, Johnson T, Khor CC; LifeLines Cohort Study, Lluís-Ganella C, Luan J, Lyytikäinen LP, Nolte IM, Sim X, Sõber S, van der Most PJ, Verweij N, Zhao JH, Amin N, Boerwinkle E, Bouchard C, Dehghan A, Eiriksdottir G, Elosua R, Franco OH, Gieger C, Harris TB, Hercberg S, Hofman A, James AL, Johnson AD, Kähönen M, Khaw KT, Kutalik Z, Larson MG, Launer LJ, Li G, Liu J, Liu K, Morrison AC, Navis G, Ong RT, Papanicolau GJ, Penninx BW, Psaty BM, Raffel LJ, Raitakari OT, Rice K, Rivadeneira F, Rose LM, Sanna S, Scott RA, Siscovick DS, Stolk RP, Uitterlinden AG, Vaidya D, van der Klauw MM, Vasan RS, Vithana EN, Völker U, Völzke H, Watkins H, Young TL, Aung T, Bochud M, Farrall M, Hartman CA, Laan M, Lakatta EG, Lehtimäki T, Loos RJ, Lucas G, Meneton P, Palmer LJ, Rettig R, Snieder H, Tai ES, Teo YY, van der Harst P, Wareham NJ, Wijmenga C, Wong TY, Fornage M, Gudnason V, Levy D, Palmas W, Ridker PM, Rotter JI, van Duijn CM, Witteman JC,

Chakravarti A, and Rao DC. Gene-Age Interactions in Blood Pressure Regulation: A Large-Scale Investigation with the CHARGE, Global BPgen, and ICBP Consortia. *American Journal of Human Genetics. 2014;* 95(1), 24–38.

Saitou N. Statistical methods for phylogenetic tree reconstruction. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* 1991 (pp. 317-346), Elsevier B.V., the Netherlands.

Strasak AM, Zaman Q, Marinell G, Pfeiffer KP and Ulmer H. The Use of Statistics in Medical Research. *The American Statistician.* 2007; 61:1, 47-55.

Thomas D. Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics.* 2010; 11(4): 259-272.

Thompson EA. Estimation of relationships from genetic data. (Rao CR and Chakraborty R, Eds) *Statistical Methods in Biological and Medical Sciences (Handbook of Statistics Volume 8)* (pp. 255-269), Elsevier B.V., the Netherlands.

Tunes-da-Silva G, Pedroso-de-Lima AC and Sen PK. Quality of life perspectives in chronic disease and disorder studies. Rao CR, Chakraborty R and Sen PK (Eds). *Bioinformatics in Human Health and Heredity (Handbook of Statistics Volume 28)* 2012 (pp. 401-431), Elsevier B.V., the Netherlands.

Voorman A, Lumley T, McKnight B, and Rice K. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One.* 2011; 6(5): e19416.

Weiss, K.M., and Chakraborty, R. (1984) Multistage risk models and the age pattern in familial polyposis coli. *Cancer Investigation* **2**:443-448.

Yi N. Statistical analysis of genetic interactions. *Genetics Research (Camb).* 2010; 92: 443-459.

Zhang G, Jin L and Chakraborty, R (2012) Single-Locus Association Analysis with Ordinal Tests. In*: Handbook of Statistics, Vol. 28 – Bioinformatics in Human Health and Heredity* (Rao, C.R., Chakraborty, R., and Sen, P. K. eds.), Chapter 12, pp. 309-338, Elsevier, Amsterdam.

Zhang Y and Liu JS. Fast and Accurate Approximation to Significance Tests in Genome-Wide Association Studies. *Journal of the American Statistical Association*. 2011; 106:(495): 846-857.

Prof. C. R. Rao,

29 Old Orchard Street

Williamsville, NY 14221

(crr1@psu.edu)

International Advisory Board – (*Continued…*)

C. Radhakrishna Rao

Sc.D., FRS
Padma Vibhushan Awardee
Professor Emeritus and Hon. Advisor to CRRAO AIMSCS
India Science Medal Awardee
National Medal of Science Laureate, USA
E-mail ID: *crr1@psu.edu*

"One could say that if Europe is the mother of differential calculus based on deterministic analysis, India could well be called the mother of statistics. When I think of modern statistics, Dr. C.R. Rao features on the top of the list. He once said that "statistics is the technology of finding the invisible and measuring the immeasurable."

Abdul Kalam, former President of India,
Bharata Ratna, in a speech delivered at
University of Hyderabad

"C.R. Rao is a great name from the golden age of statistics. His work was done in India, and his intellect shaped statistics worldwide."

Julian Champkin, editor, *Significance*, a
Journal of Royal Statistical Society, UK and
American Statistical Association, USA, Vol.8,
No. 4, Nov 2011, 175-178.

EARLY LIFE AND ACADEMIC QUALIFICATIONS

Calyampudi Radhakrishna Rao (C. R. Rao) was born on 10 September 1920 in Huvanna Hadagali, Karnataka State, India. He is the eighth child in a family of six brothers and four sisters who were raised by his parents, C.D. Naidu and Lakshmikantamma, and was named Radhakrishna following the tradition of naming the eighth child in a family after God Krishna, who is the eighth child.

He studied in schools at Gudur, Nuzvid, Nandigama, and Visakhapatnam, in Andhra Pradesh and completed his M.A. degree in mathematics in 1940 at the Andhra University in Waltair, Andhra Pradesh, with a first class and first rank. He has the unique distinction of having the first rank in the final examination in all classes from primary to college. He joined the Indian Statistical Institute (ISI) in January 19 41 as a student of one year training course in statistics. On the advice of Professor Mahalanobis, the director of ISI, he took admission in 1941 for MA course in statistics of Calcutta University, and obtained MA degree in statistics with a first class and first rank with record marks which remains unbeaten for the last 80 years. He was among the first few people in the world to hold a Master's degree in Statistics at that time. In 1943 he was offered a research scholarship at ISI. Soon after he joined the ISI at the age of 20, he started doing research in statistics by himself with no guide and publishing papers. This was the beginning of his career in statistics. He took compulsory retirement from ISI in 1980 at the age of 60 after 40 years of active service.

EARLY YEARS IN RESEARCH: BREAKTHROUGH PAPERS

Two of Rao's papers, one published in 1945 at the age of 25 (Bulletin of Calcutta Math. Soc., Vol. 37) and another published in 1947 at the age of 27 (Proc. Cambridge Phil. Soc., Vol.44) have been included in the book on *Breakthrough Papers in Statistics: 1889-1990*.

The paper published in 1945 attracted worldwide attention and generated technical terms bearing Rao's name **- Cramer-Rao inequality, Rao Blackwell theorem, Rao metric and Rao distance**. These technical terms appear in text books on statistics, econometrics and engineering. The following quote refers to the breakthrough paper by Rao published in 1945 at the age of 25.This paper generated a large number of papers by others pursuing some ideas given in the paper.

"The article focuses on an important piece of work of the world renowned Indian statistician, Calyampudi Radhakrishna Rao. In 1945, C.R. Rao

(25 years old then) published a path breaking paper, which had a profound impact on subsequent statistical research. It opens up a novel paradigm by introducing differential geometric modeling ideas to the field of statistics. In recent years, this contribution has led to the birth of a flourishing field of Information Geometry"

-- Frank Nielsen, Professor in Computer Science, Ecole Polytechnique, France, in his paper on Cramer_Rao Bound and Information geometry, *Connected at infinity*, 2, 18-37, 2013.

The paper published in 1947 generated the technical term **Score Statistic** which appears in text books on econometrics and has been the subject of discussions at various conferences.

INVITATION FROM CAMBRIDGE UNIVERSITY, UK

On a request from the Anthropology Department of Cambridge University, UK, to Professor Mahalanobis to send someone to analyze, by using the methods of multivariate analysis, measurements made on ancient human skeletons brought from *Jebel Moya* in North Africa by the University Museum of Archeology and Anthropology, to trace the origin of the people who lived there. Rao was sent to Cambridge in 1946 as he had the required expertise. Rao had by 1946 at the age of 26 about 30 research publications. The analysis of multiple measurements was not well developed at that time. Rao worked in Cambridge at the University museum of Archeology and Ethnology for two years (1946-1948) as a visiting scholar, developed some new methods of analysis of multiple measurements and used them to analyze the data. The results of the work done at the museum were published in the book *Ancient Inhabitants of Jebel Moya*, Cambridge University Press, 1955, under the joint authorship of Rao and two anthropologists.

Rao received Ph.D. degree of Cambridge University at the end of his 2-year assignment in 1948, based on new multivariate methodology (MANOVA) he developed while analyzing skeletal data, and Sc.D. in 1974, based on peer review of his publications. In 1974, he was made an Honorary Life Fellow of Kings College, Cambridge University, a rare honor.

During the two year period (1946-1948) of stay in Cambridge, in addition to working on the Jebelmoya project, Rao published a number of research papers in Biometrica and J. Royal Statistical Society introducing new methods of statistical analysis.

Rao had about 50 published papers during the forties of the last century on various aspects of statistics on statistical methodology.

The 1940s were ungrudgingly C.R. Rao's. His 1945 paper, which contains the Cramer-Rao Inequality, Rao-Blackwell Theorem, and the beginning of differential geometry of parameter space will guarantee that, even had he done nothing else-but there was much".

- Terry Speed, FRS, Professor, University of Berkeley, USA, in Institute Mathematical Statistics Bulletin, USA, Jan-Feb, 2010 issue

Rao is considered as one of those who developed statistics as an independent discipline of great practical importance.

"The first half of the 20th century was the golden age of statistical theory, during which our field grew from ad hoc origins, similar to the current state of computer science to a firmly grounded mathematical science. Men of the intellectual caliber of Fisher, Neyman, Pearson, Hotelling, Wald, Cramer, and Rao were needed to bring statistical theory to maturity."

Brad Efron, National Medal of Science Laureate, USA, Professor of Statistics and Health Research Policy, Stanford University, USA, in Royal Statistical Society (UK) News, 22, 1995

INDIAN SCHOOL OF STATISTICS: DEVELOPMENT OF EDUCATIONAL AND RESEARCH ACTIVITIES AT THE INDIAN STATISTICAL INSTUTUTE

Rao returned to ISI in 1948 after 2 years stay in Cambridge and worked in various capacities over a period of 40 years (1940-1980), (as Full Professor at the age of 28-29, Professor and Head of the Division of Research and Training ,Director, Jawaharlal Nehru Professor, and National Professor).

"Government has appointed C.R. Rao, an eminent statistician, as National Professor. Professor Rao is an outstanding and creative thinker in the field. He was appointed by Professor Mahalanobis as full-fledged professor of the Indian Statistical Institute at the early age of 29 in recognition of his

creativity". – Secretary, Government of India

When Rao joined the ISI in early forties, statistics was not considered as an independent discipline and no university offered courses at the master's level. Rao developed numerous courses in statistics, started bachelor's and master's degree programs at the ISI. Rao also initiated for the first time in India, a Ph.D. program in theoretical statistics and probability. D. Basu, the first Ph.D. he guided, was the first Ph.D. in theoretical statistics in India. Up to date he has guided the research work of 50 students for Ph.D., who in turn have produced about 500 Ph.D.'s *(website: The Mathematics Genealogy Project C R Rao).* With encouragement from Professor Mahalanobis, he laid the foundation for all research and educational activities at ISI and created what was called as the **Indian School of Statistics.** Rao is considered as one of those who developed statistics as an independent discipline of great practical importance.

SERVICE TO GOVERNMENT

NATIONAL STATISTICAL: SYSTEM: Under the direction of Professor Mahalanobis he helped in establishing the State Statistical Bureaus. Indian National Statistical System is considered to be one of the best in the world.

MEMBERSHIP OF GOVERNMENT COMMITTEES: He was a member of several government committees during the period when Smt. Indira Gandhi was prime-minister. Some of the important committees he served on are: Chairman of the National Committee on Statistics (1962-69), Chairman of the demographic and communication action research program committee (for population control) (1968-69), Chairman of the committee on Mathematics, Atomic Energy Commission (AEC) (1969-78), Member of COST (Committee on Science and Technology (1969-71), Member of Justice Sarkar Committee to enquire into overall functioning of CSIR.

LATER YEARS IN RESEARCH AND EDUCATION – USA (University of Pittsburg, Pennsylvania State University, University at Buffalo)

After taking compulsory retirement in India at the age of 60 with 40 years of service at ISI, Rao tried to get a suitable job in India to continue his research activities without administrative responsibilities. As opportunities to work in India were denied he accepted offers of distinguished professorships in USA with very limited teaching responsibilities. He worked for another 25 years as University Professor at the University of Pittsburgh and Eberly (Chair) Professor of Statistics at The Pennsylvania State University continuing his research in diverse areas of Statistics. He retired from active service at the age of 80 but continued his research activities as Director of the Center for Multivariate Analysis at The Pennsylvania State University.

He has 475 research publications (201 working at ISI for 40 years and 274 working in USA for 25 years). He has one published paper in 2014 in the Proc. Nat. Acad. Sc., USA, at the age of 95, and some are under publication.

CR RAO ADVANCED INSTITUTE OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE (CRRAO AIMSCS)

Rao was very interested in encouraging research in India which has a poor rank even below that of some developing countries in terms of research publications. Rao thought of establishing a research institute in Hyderabad and developing programs of research in Mathematics, Statistics and Computer Science. This was in 1990 when Chandrababu Naidu was the Chief Minister of Andhra Pradesh. He immediately responded to Rao's request by sanctioning Rs. 50 lakhs. With an additional donation of Rs.25 Lakhs from Mrs. C.R. Rao and a gift of 5 acres of land by Pro.f Hasnain, Vice Chancellor of the University of Hyderabad, in the University campus, and a grant of Rs.15 crores from Dr. Manmohan Singh, Rao established the institute named after him by the governing council of the institute as CRRAO ADVANCED INSTITUTE OF MATHEMATIC, STATISTICS AND COMPUTER SCIENCE (CRRAO AIMSCS). It took 10 years of planning and execution of dedicated work to complete the project.

The C.R. Rao Gallery highlighting Rao's academic career in the institute was designed by Dr. Tejaswini Rao to encourage students to study statistics. The Gallery was inaugurated by the Nobel Laureate and Knighted Dr. Venkataraman Ramakrishnan, of the Molecular Laboratory, University of Cambridge, UK in December 2013.

STATISTICS OLYMPIAD: Rao started the Statistical Olympiad program organized by AIMSCS, the first in the world, to encourage the study of statistics. He plans to establish a museum of statistics to spread statistical literacy in India.

PUBLICATIONS AND TECHNICAL TERMS BEARING RAO'S NAME

Rao is the author of 14 books and 475 research papers most of which are published in high impact journals. One of his books**,** *Linear Statistical Inference* is translated into several European languages, Japanese, and Chinese. Another book *Statistics and Truth* is translated into German, Spanish, Japanese, Taiwan Chinese, Mainland Chinese, Korean, and Turkish languages.

Some of his publications motivated further research by others. Some of the technical terms arising from his research which appear in most of the text books on statistics, econometrics, and engineering are *Cramer Rao Inequality, Rao Blackwellization, Rao's Score Test, Orthogonal Arrays described by Forbes Magazine as a new "new manthra" for American Industries, Rao's G-inverse, MANOVA (Multivariate Analysis of Variance), and Rao's U test for adequacy of a subset of Measurements.*

Other technical terms which appear in specialized papers are *Rao distance, Fisher Rao metric, Rao's Quadratic entropy, ANODIV (analysis of Diversity), Rao's paradox in multivariate analysis, Rao's paradox in sample surveys, Khatri Rao product and Lau, Rao, Linnik, and Shanbagh Theorems, and Burbea Rao Divergence measures.*

Other technical terms bearing Rao's name which are based on papers by others are *Steriological Rao Blackwell Theorem, Quantum Cramer Rao Bound, Fisher Rao Theorem, and Rao Measure.*

Rao edits a series of Handbooks on various topics in statistics. To date he has edited 33 Handbooks.

38 HONORARY DOCTORATE DEGREES

Rao was awarded 38 Honorary Doctorate degrees from universities in 19 countries, spanning 6 continents: Europe (10 countries, *11 degrees*): Germany, Russia, Switzerland, Poland, Serbia, Spain, Finland, Portugal, Greece, and Cyprus; North America (2 countries, 8 degrees): USA 6 and Canada 2; South America (2 countries, *2 degrees*): Brazil and Peru; Asia (3 countries, *15 degrees*): India, Sri Lanka, and Philippines; Australia (*1 degree*) and Africa (*1 degree*).

PROMOTION OF RESEARCH IN ECONOMETRICS IN INDIA

Rao founded The Indian Econometric Society (TIES) in 1971 and conducted yearly seminars on Data Base of Indian Economy. He established Mahalanobis National and International Medals to be awarded to econometricians at the annual conference of TIES.

AWARDS OF MEDALS AND OTHER HONORS

Highest Awards of Medals in India, USA, and UK

- **Samuel Wilks Medal of American Statistical Association 1989, the highest award given to a statistician in USA,** "for major contributions to the theory of multivariate statistics and applications of that theory to problems of biometry; for worldwide activities as advisor to national and international organizations; for long time conscientious as a teacher, editor, author, and founder of academic institutions; and for the great influence he has had on the application of statistical thinking in different scientific disciplines, embodying over a career of more than 40 years the spirit and ideals of Samuel S. Wilks."
- **National Medal of Science 2003, the highest award given to a scientist in USA:** Awarded by the president of USA with the citation "*as a prophet of new age for his contributions to the foundations of statistical theory and multivariate statistical methodology and their applications, enriching the physical, biological, mathematical, economic, and engineering sciences".*
- **India Science Award 2009, the highest recognition given to a scientist in India**, *"for major contribution(s) of a path-breaking nature.*" The award, given by the Prime Minister of India, carries a gold medal and cash prize of Rs. 25 lakhs.
- **Guy Medal in Gold of the Royal Statistical Society 2011, the highest award given to a statistician in UK**: This award is given once in

3 years to "*those who are judged to have merited a significant mark of distinction by reason of their innovative contribution to theory or application of statistics."* This is the first time the medal was given to an Asian and second time given to a Non-British citizen during the last 118 years since the inception of the medal.

- **Jerzy Splawa-Neyman medal** from the Polish Statistical Association "in recognition of his outstanding contributions to the theory, applications and teaching of statistics. "The medal was established in 2012 to celebrate the 100th anniversary of the Polish Statistical Association. "'We are sending our best words of appreciation and admiration of your extraordinary scientific work and your invaluable commitment to the statistical education all over the world," – President of the Association.

Other Medals

- **Megnadh Saha Medal 1967 and Srinivasa Ramanujan Medal 2003** of the Indian National Science Academy *"for outstanding contributions to science*".
- **Jagdishchandra Bose Gold Medal 1979,** of Bose Institute
- **Mahalanobis Birth Centenary Gold Medal 1996** awarded by the Indian Science Congress Association

Other Awards

- **Bhatnagar Award 1963**, of the Indian Council of Scientific and Industrial Research for contributions to science, from Prime Minister Nehru.
- **International Mahalanobis Prize 2003,** "*for lifetime achievement in statistics and the promotion of best statistical practice* "awarded by the International Statistical Institute.
- "**Prof. C.R. Rao Road"**, The road from IIIT to Alind Factory, Lingampally covering University of Hyderabad is named as "Prof. C.R. Rao Road, in recognition to the services rendered by the legendary figure of Indian Science, Prof. C.R. Rao, Padma Vibhushan, world's renowned Statistician".
- The Department of Statistics and Programme Implementation (DOS&PI) instituted in 1999 "a **National Award in honor of C.R. Rao**, the renowned statistician of the country". The award given once in two years carries an amount of one lakh of rupees and is reserved for young statisticians for the work done during the preceding three years in any field of statistics.(GOVT. Notification)

Professional Awards

Rao received recognition from all statistics societies for his pioneering contributions to statistical theory and applications. He was elected to Royal Society (FRS), UK, the National Academy of Sciences, USA, American Academy of Arts and Science, Indian National Science Academy, Indian Academy of Sciences, National Academy of Sciences, India, Lithuanian Academy of Sciences, and Third World Academy of Science. He was made an Honorary Member of the European Academy of Sciences, the International Statistical Institute, International Biometric Society, Royal Statistical Society (UK), Finnish Statistical Society, Portuguese Statistical Society, Institute of Combinatorics and Applications, American Statistical Association, and World Innovation Foundation, Life fellow of Kings College, Cambridge.

He has been the president of the International Statistical Institute, the first one from Asia, Institute of Mathematical Statistics, USA, the first one from outside USA and the International Biometric Society, the first one from Asia.

A PLACE IN THE HISTORY OF STATISTICS

"*Rao's early contributions to statistical theory earned for him a place in the history of statistics".* (Website: Founders of statistics, www.math.unm.edu/~ppathak/Founders/)

Rao's name appears in all websites on "History of Statistics" as one of those who developed statistics as a distinct discipline.

"C.R. Rao is among the worldwide leaders in statistical science over the last five decades. His research, scholarship, and professional service have had a profound influence in the theory and applications of statistics and are incorporated into standard references for statistical study and practice."

S. Karlin, Member of National Academy of Science and National Medal Science Laureate, USA

Ranajit Chakraborty

PhD, FACE (USA), FIASc (India), FCNAS (Chile)

E-mail ID: *Ranajit.Chakraborty@unthsc.edu*

Currently Dr. Ranajit Chakraborty is the Director of the Center for Computational Genomics (CCG) at the Institute of Applied Genetics, and a Professor of Molecular and Medical Genetics at the University of North Texas Health Science Center at Fort Worth, Texas. He received his PhD degree in Biostatistics and Population Genetics in 1971 from the Indian Statistical Institute at Kolkata, India, under the guidance of Professor C. R. Rao, FRS. Since 1971 he served as faculty member at numerous universities in India and USA and had been a visiting professor in India, Germany, Chile, Japan, and Sweden. His national and international awards include: Fellow of American College of Epidemiology (FACE) in 1980; US Federal Bureau of Investigation Award for "Research in DNA Forensics during the Decade of DNA 1989-1998"; Foreign Member of the Chilean National Academy of Sciences (FCNAS), Santiago, Chile (in 2004); Honorary Member of the Mediterranean Academy of Forensic Sciences, Italy (in 2004); and Honorary Fellow of the Indian Academy of Sciences (FIASc in 2006). With over 570 research articles and 8 edited books, he contributed in the areas of molecular population genetics, complex disease genomics, strategies of gene mapping, statistics of parentage and relationship testing, radiation risk estimation, and DNA and Microbial forensics. He also co-authored 8 National and International Committee/ Governmental Reports. His discoveries lead to 6 US patents. He served as associate editor for over 23 international journals. He has graduated 58 PhD and MS students and trained 37 post-doctoral fellows. He also served in numerous national and international committees, including being on the US DNA Advisory Board (1996-2000), New York State DNA Subcommittee (1996-2011), International Commission of Radiological Protection (2004-current), grant reviewer and study section member of US National Institutes of Health (NIH), National Science Foundation (NSF), and National Institute of Justice (NIJ) continually since 1976. He helped in designing the Combined DNA Index System (CODIS) database of US Federal Bureau of Investigation and still serves the CODIS project as its Subject Matter Expert (SME). Since 1991 he testified in many courts nationally as well as internationally presenting valid DNA evidence as well as challenging them when they are inappropriately generated.

His current Research Interest are:

Statistical and Population Genetics - Complex Disease Genomics – Gene-Environment Interaction – Admixture Mapping - Molecular Epidemiology – DNA and Microbial Forensics – Missing Person Identification by DNA testing - Radiation Risk Estimation – Precision Medicine - Health Disparities – Human Genome Diversity

Dabeeru C. Rao

E-mail ID: *rao@wubios.wustl.edu*

https://biostatistics.wustl.edu/facultyandstaff/Pages/DCRao.aspx

Dr. Dabeeru C. Rao, PhD, earned bachelor's and master's degrees in statistics and a doctorate in statistical genetics from the Indian Statistical Institute in Kolkata, India in 1971, under the guidance of Professor C. R. Rao, F.R.S. After a one year post-doctoral fellowship at the University of Sheffield, England, he spent the next 8 years carrying out his career defining research at the University of Hawaii, Honolulu. He then joined Washington University in St Louis in 1980 where he is Professor and Director of the Division of Biostatistics at Washington University School of Medicine in St. Louis. He is also a professor of biostatistics in genetics and psychiatry, and mathematics.

Rao is the principal investigator of several National Institutes of Health (NIH) grants that support his research and has authored over 600 scientific articles. In addition to active research, Rao is also dedicated to training and mentoring activities. He developed and directs two master's degree programs in biostatistics (MSIBS) and genetic epidemiology (GEMS). He also directs a post-doctoral training program (T32) in cardiovascular genetic epidemiology, and a Summer Institute program (PRIDE) to enhance diversity in the biomedical workforce, both supported by NIH grants.

In his 40+ years as a researcher, Rao has made substantive contributions to the study of human disease, most notably hypertension and related metabolic diseases. Early on, he has played a watershed role in the development of the field of genetic epidemiology, including the development of statistical methods for evaluating the roles of nature and nurture in human disease. He is a member and Past-President of the International Genetic Epidemiology Society (IGES, 1996) and was the founding Editor-in-Chief of the society's journal, *Genetic Epidemiology* (1984-91). Rao received an "IGES Leadership Award" from the IGES in 1997, and a "Champion of Public Health" award from the School of Public Health and Tropical Medicine, Tulane University, New Orleans, in 2005

Rao is best recognized for a lifetime of distinguished contributions to advancing and promoting statistics in human genetics and for contributions to training biostatisticians and statistical geneticists. These contributions led to his election as a Fellow of the American Association for the Advancement of Science (AAAS) in 2012 and as a Fellow of the American Statistical Association (ASA) in 2013.

L. Jeyaseelan

E-mail ID: *ljey@hotmail.com*

Dr. L. Jeyaseelan is Professor of Biostatistics at the Dept. of Biostatistics, Christian Medical College, Vellore, India - 632002. He was the former chair of the department (on rotation). He is International Clinical Epidemiology (INCLEN) Fellow. He has established Biostatistics Resource and Training Centre (BRTC) in 1998 that trains medical faculty in Bio statistical and epidemiological methods. This centre has trained over 3000 medical faculty and PG students through workshops/short courses on various topics since 2001. He established Clinical Data Management Centre (CDMC) in 2007, which undertakes data management for vaccine trails and large scale clinical trials. This centre has been equipped with powerful softwares such as Oracle clinical, Promasys and a team of Programmers and Biostatisticians. A WHO sponsored study on Measles vaccination got published in New England Journal of Medicine and another study on Vitamin D for TB in Lancet in 2015. These studies data management, analyses and writings were done at this centre. He developed sample size calculation software "nMaster", which is marketed now in India. He is currently developing a software for Diagnostic Test data analyses named "Dx Test". Dx Test software Version 1.0 has been released to selected researchers. He got involved as PI, CO-PI on national level studies such as Prevalence of STD and HIV in India, Study on Abuse in Family Environment, Antibiotics Prescription Pattern study, and National Estimation of Blood requirement in India etc. He serves as a member of Data Safety Monitoring Board (DSMB) in DBT India and WHO. He has about 125 papers published in national and international journals. He is specialized in Clinical Trials methods, Bayesian Methods, Bootstrap, Categorical and Survival data analyses. There are 4 students currently doing PhD under his guidance, two are waiting for Viva Voce and a student was awarded PhD. His interest has been in teaching and doing methodological research.

S. B. Sarmukaddam

E-mail ID: *sanjeev.sarmukaddam@gmail.com*

Dr. S. B. Sarmukaddam is M.Sc. in Biometry from Department of Statistics, University of Pune; D.B.S. (P.G. Diploma in Biostatistics) from ICMR's Institute for Research in Medical Statistics, New Delhi; M.P.S. (Master of Population Studies) from International Institute for Population Sciences, Bombay; and Ph.D. in Biostatistics (topic: Some Inferential Problems in Epidemiology) from School of Health Sciences, University of Pune. He has taught Biostatistics in many Medical Colleges/institutions & Universities (including National Institute of Mental Health and Nuero Sciences, Bangalore and Institute for Research in Medical Statistics [presently known as National Institute of Medical Statistics], New Delhi, as permanent faculty member for 3 years each & at Pune University as guest faculty for 5 years). On special invitation, he also lectured at Lawrence Education Center of Borgess Medical Center (arranged by Department of Community Health) of University of Michigan, USA in their 'grand rounds'. His presentation during International Biometric Conference (16-21 July 2006) held at McGill University, Montreal, Quebec, Canada and at many places of academic interest in London, UK (on "Need For and Modifications in Clinical Trial Methodology for Evaluation of Indigenous Treatments/Medicinal Drugs") was appreciated. He also participated on invitation in symposium on "Controversies in Clinical Trials" (2-3 August, 2010) and in 11th International Congress of Behavioral Medicine (4–7 August, 2010) organized by NIH's National Heart Lung and Blood Institute with Rush University Medical Center at Grand Hyatt Washington, USA.

He has more than 50 scientific publications in peer reviewed international journals (including Lancet, i, 14th January 1989, pp85-88; Acta Psychiatrica Scandinavica, vol.74, 1986, pp183-186; Annals of Natl. Acad. Med. Sci., vol.32, 1996, pp177-184; Medical Teacher, vol. 23, No. 1, 2001, pp 102-103; Soc. Psychiatry and Psychiatric Epidemiology, vol. 42, 2007, pp 561-572; The British Journal of Psychiatry, vol.195, 2009, pp368–371; Transcultural Psychiatry, vol. 48, No. 3, 2011, pp 257-283; Oxford Journal for the British Society for Rheumatology, vol. 299, 2013, pp 258-268; Clinical Rheumatology, doi: 10.1007/s10067-011-1809-z; Evidence-Based Complementary and Alternative Medicine, doi:10.1155/2011/724291; etc) and has more than 80 total number of publications/conference presentations to his credit. He reviewed articles for journals like Bulletin of WHO, Int. Jr. eCAM, AJP, IJMR, Clinical Rheumatology, PLOS ONE etc. A co-authored book entitled "Medical Biostatistics" is published in 2001 by a New York (USA) based publisher Marcel Dekker, Inc. He also has authored "Fundamentals of Biostatistics", "Biostatistics - Simplified", and "Clinical Biostatistics" A new book "Methods & Controversies in Clinical Trials" is under publication.

He recently (31.07.2015) retired from Maharashtra Institute of Mental Health, B.J.Medical College & Sassoon Hospital Campus, Pune, India.

Editorial Board – (*Continued…*)

N.C. Jain

E-mail ID: *drencejain@gmail.com*

Joined the Indian Council of Medical Research (ICMR), New Delhi in 1982, Dr N.C. Jain is currently Editor (Production) of the *Indian Journal of Medical Research (*IJMR*)* and Scientist 'G' in the Division of Publication & Information, ICMR. Dr Jain has been associated with the IJMR, research evaluation, modernization of the library and information network of ICMR institutes, IPR issues for a short time and human resource development programmes, among others.

He has studied in the University of Delhi, Delhi for various degrees *viz.* B.Sc (Hons)[Hans Raj College], M.Sc, M.Phil and Ph.D. in Botany. Dr Jain has also done a Post Graduate Diploma in Journalism from Bhartiya Vidya Bhavan, New Delhi and General Course on Intellectual Property from World Intellectual Property Organization, Geneva.

He has over 50 papers on S&T communication, book reviews, *etc*. in research journals / books / proceedings, two handbooks on communication, over 100 articles in newspapers, popular science magazines and over 60 AIR programmes to his credit.

Dr Jain is Member, Editorial Board, ***Indian Journal of Community Medicine***, Indian Association of Preventive and Social Medicine and Member, Editorial Advisory Board, ***Indian Journal of Research in Homoeopathy***, Central Council for Research in Homoeopathy, New Delhi.

Dr Jain is Fellow of the Society for Information Science, New Delhi for 2012. He is recipient of the B.C. Deb Memorial Award for Popularization of Science for 1997-98 of the Indian Science Congress Association, Kolkata and Raizada Memorial Award 1992 for Young Information Scientist of the Society for Information Science, New Delhi.

Traveled widely in India and abroad like PR China, France for his professional development, he is member of various professional societies like COLLNET, European Association of Science Editors, Indian Science Writers Association, Society for Information Science, Indian Science Congress Association, *etc*.

K. Thannarasu

E-mail ID: *kthenna@gmail.com*

Dr. K Thennarasu did his MSc and PhD from IASRI, New Delhi with Junior research Fellowship (JRF) and Senior Research Fellowship (SRF) respectively. He started his professional carrier as Lecturer and later as Assistant Professor Biostatistics & HOD at PGIMER, Chandigarh from 1994 to 2000. Later he joined at NIMHANS and currently he is Professor and Head of the Department of Biostatistics, NIMHANS, Bangalore.

In the field of genetical statistics his work has provided a new method of non-parametric G x E stability measures and these have been named as Thennarasu's measures of stability and these algorithms are available in R package. He has many research project as PI and Co-PI funded by different national and international agencies. He have guided many PhD Scholar for their PhD Degree and currently guiding four Ph.D candidates as guide and one Ph.D. candidate as joint guide in Biostatistics apart from many others for their clinical PhD. He has been joint guide to DM Neurology, Ph.D (Clinical Neuro science) and Ph.D(Clinical Psychology). He has published more than 170 papers as author/co-author, which were published in National\International journals. He has published one book titled "NIMHANS Neuropsychological Battery for Children" as joint-author. As on September 2015 his Google scholar (Thennarasu Google scholar link) citation index is 2235 with h-index of 27and i10 index 62.

Dr. K. Thennarasu has one patent "Top2a Inhibition By Temozolomide Useful For Predicting GBM Patient's Survival" (patent no: WO2012038812) as co-patentee. This invention provides a TOP2A inhibition by temozolomide useful for predicting glioblastoma patient's survival. He is course leader for various teaching programs at NIMHANS and coordinator of Ph.D. course work for research methodology, ethics, computer application and quantitative software modules. He is conducting regular workshop on Biostatistical methods and Research methodology.

He has received Fogarty INDO-US Fellowship 2005-2006 and did my PDF from Washington School of Medicine, St. Louis, US. He has been awarded FSMS for the year 2014. He is a life member of ISMS for the last twenty years. I also served ISMS as a member of various committees including executive and award committees. He is also recipient of SWAN-NaThaM award for best field based article published Journal of School Social work for the year 2008-2009 and active member of number of other professional bodies. He was the "Fellow of The Royal Statistical Society (FRSS) London, UK. He is reviewer and editorial board member of various journals.

Original Articles and Technical Notes

Understanding Basics of Impact Factor and Good Publication Practices in the Field of Statistics and Probability

N.C. Jain
Scientist 'G'
Division of Publication and Information,
Indian Council of Medical Research,
V. Ramalingaswami Bhawan,
Ansari Nagar, New Delhi 110 029, India
E-mail ID: *drencejain@gmail.com*

It is now increasingly being realized that communicating the outcome of research as rapidly as possible to the global scientific community is as important as doing research itself. Besides establishing the claim of the investigators, quick publication ensures other researchers to know of and utilize the results of these investigations. Equally important is the need to publish in journals that are widely indexed internationally and read by a majority of scientists. Even while new journals are being started regularly, they are just not able to cope with the new information generated by the ever increasing scientific community. Publication in such widely indexed journals with international outreach continues to be competitive and difficult. Fortunately, so many biomedical journals now exist that the chances of not having the work published are small.

Even in the field of statistics, the options for choosing an appropriate journal are many. For example, *StatSci.org* continues to be a dependable portal for statistical science [1], which provides among others, 197 journals in the field of statistical science. The SCImago Journal & Country Rank is another important portal that includes the journals and country scientific indicators developed from the information contained in the Scopus® database (Elsevier B.V.)[2]. Specifically, its subject category, statistics and probability for 2014 lists 171 journals with several journal metrics. However, a researcher working in the field of medical statistics, invariably looks at the PubMed database for selecting an appropriate journal for publication of a research paper. As many as 32 journals are indexed in PubMed under category, statistics [3] and another 20 under vital statistics [4]. All said and done, still a researcher wishes to publish in those journals which are indexed in Thomson Reuters' Web of Science's principal component, the *Science Citation Index-Expanded* database, with impact factor (IF) [see Box for definition] available since 1975 in its sister annual publication, the *Journal Citation Reports*, popularly known as JCR.

With this backdrop, the latest science edition of JCR contains 2014 IF of 8618 journals in all fields of science & technology (S&T) including statistics and probability[6]. The IF ranged from 115.840 [*CA-Cancer Journal for Clinicians*] – no IF; top 20 journals in terms of 2014 IF are given in Table I. The JCR category, statistics and probability has 122 journals; 2014 IF ranged from 4.472[*Statistical Methods in Medical Research*] – NA[*Annual Review of Statistics and its Application*]. Of these 122 journals, 52(42.62%) are having IF >= 1.00[IF >=4.000, 1 journal, IF>=3.000,3 journals, IF >=2.000, 8 journals and IF >=1.000, 40 journals].Table II lists 20 journals of statistics and probability (n=122) with 2014 IF of interest to those working in the field of medical statistics. However, important journals which are not included under the category of statistics and probability like *International Journal of Epidemiology* has 2014 IF as 9.176, *Journal of Health Economics* as 2.579 and *Health Economics* as 2.227 .Table III provides a list of 20 important Indian S&T journals with 2014 IF of interest to researchers working in the field of medical statistics; total Indian journals covered in the JCR being 104.

Just a couple of more points on IF. All the *Annual Reviews* have high IF as denominator is low because of just one issue a year for *Annual Reviews* and citations are many thereby resulting in generally high IF for all *Annual Reviews* (Table IV). Similarly, a review article attracts more citations; the classical case being that of mutation in IF [7] in the journal *Acta Crystallographica Section A* wherein with the publication of just one review article, its IF jumped from 2.051in 2008 to 49.926 in 2009 and even increased to 54.333 in 2010; now its 2014 IF is 2.307. *h*-index is another metrics which is fast emerging for evaluation of research output[8].

With severe competition for publishing in journals with IF with extremely high rejection rate, one has to be very cautious in selecting a journal for publication of a research paper by prudently avoiding predatory journals [those that exist solely to collect publishing fees with no consideration of academic integrity] [9]. It is therefore suggested to verify the journal IF with JCR and cross check its identity with its unique International Standard Serial Number, popularly known as ISSN (online/print). Equally important is to follow the ethical issues in publication practices as laid down by especially the well known global agency, the COPE [Committee on Publication Ethics][10]. Lastly, one should not forget to thoroughly check the paper through an appropriate plagiarism software to ensure hassle free publishing in an indexed journal with its IF (?) serving as icing on the cake!

Acknowledgment

The author acknowledges Dr Ajit Mukherjee, Scientist F, Indian Council of Medical Research (ICMR), New Delhi for constructive suggestions, especially in short-listing 20 relevant journals for medical statistics professionals as also Dr G.Mahesh, Editor, *Annals of Library and Information Studies* & Principal Scientist, CSIR-NISCAIR (Council of Scientific & Industrial Research- National Institute of Science Communication and Information Resources), New Delhi for valuable inputs.

Disclaimer

The author's views expressed in this communication do not necessarily reflect the views of the ICMR, New Delhi.

References

1. StatSci.org. Available from: *http://www.statsci.org/jourlist.html*,accessed on August 6, 2015.
2. SCImago Journal and Country Rank. Available from: *http://www.scimagojr.com/journalrank.php?category=2613*, accessed on August 6, 2015.
3. NLM Catlogue [currentlyindexed[all] + statistics]. Available from: *c http://www.ncbi.nlm.nih.gov/nlmcatalog/?term=currentlyindexed%5Ball%5D+%2B+statistics*, accessed on August 6, 2015.
4. NLM Catlogue [currentlyindexed[all] + vital statistics]. Available from: *http://www.ncbi.nlm.nih.gov/nlmcatalog/?term=currentlyindexed%5Ball%5D+%2B+vital+statistics*, accessed on August 6, 2015.
5. Garfield E. Citation indexex for science: a new dimension in documentation through association of ideas. *Science* 1955; *122*: 108-11.
6. *Journal Citation Reports 2015*, s**cience edition** (web-based). Philadelphia: Thomson Reuters; 2015.
7. Jain NC. Mutation in impact factor. *Curr Sci* 2012*; 103:* 1266.
8. Hirsch JE. An index to quantify an individual's scientific research output. *arXiv: physics/0508025v5[physics.soc-ph]* 29 Sep 2005.
9. Jain NC. Commentary: Predatory journals. *Indian J Med Microbiol* 2015; *33*: 426.
10. Committee on Publication Ethics (COPE). Available from: *http://publicationethics.org/*, accessed on August 6, 2015.

Box: Impact Factor (IF)

A measure of the frequency with which the 'average article' in a journal has been cited in a particular year. The IF is available in the *Journal Citation Reports*, popularly known as JCR. IF is basically a ratio between citations and recent citable items published. Thus, the IF of Journal X would be calculated by dividing the number of all current citations of source items published in Journal X during the previous two years by the number of articles Journal X published in those two years. Eugene Garfield first mentioned the idea of an

Impact Factor in *Science* in 1955 [5].

Impact Factor Calculation: The IF for a particular journal is calculated by dividing the total number of citations received by the articles published in the journal during the two previous years, by the number of articles published in the journal the same period.

Mathematically,

$$IF(J_Y) = \frac{X_1 + X_2}{Y_1 + Y_2}$$

where,

$IF(J_Y)$ = Impact factor of the journal J for the year Y

X_1 = Number of citations received by Y_1 source items in Y year

X_2 = Number of citations received by Y_2 source items in Y year

Y_1 = Number of source items published in journal J in (Y-1) year

Y_2 = Number of source items published in journal J in (Y-2) year

Thus, the impact factor is the average number of citations received per year by the articles published in the journal during the previous two years, and is therefore an estimate of the number of citations an average article published in that journal the present year will receive.

Impact factors are calculated and published yearly in the *Journal Citation Reports* (JCR), brought out by the Thomson Reuters, USA. The latest edition released in June this year provides 2014 IF of journals.

Impact Factor Calculation of ***the Indian Journal of Medical Research for 2014****:*

2014 IF: 1.396

Cites in 2014 to articles published in:

2013 = 230

2012 = 355

Sum = 230+355 = 585

Number of articles published in:

2013 = 205

2014 = 214

Sum= 205+214 = 419

Calculation:

Cites to recent articles = 585= **1.396**

Number of recent articles 419

Table I. Top 20 journals in terms of 2014 IF (n= 8618)

Sl. No.	Abbreviated Journal Title	ISSN	2014 IF
1	CA-CANCER J CLIN	0007-9235	115.84
2	NEW ENGL J MED	0028-4793	55.873
3	CHEM REV	0009-2665	46.568
4	LANCET	0140-6736	45.217
5	NAT REV DRUG DISCOV	1474-1776	41.908
6	NAT BIOTECHNOL	1087-0156	41.514
7	NATURE	0028-0836	41.456
8	ANNU REV IMMUNOL	0732-0582	39.327
9	NAT REV MOL CELL BIO	1471-0072	37.806
10	NAT REV CANCER	1474-175X	37.4
11	NAT REV GENET	1471-0056	36.978
12	NAT MATER	1476-1122	36.503
13	JAMA-J AM MED ASSOC	0098-7484	35.289
14	NAT REV IMMUNOL	1474-1733	34.985
15	NAT NANOTECHNOL	1748-3387	34.048
16	SCIENCE	0036-8075	33.611

17	CHEM SOC REV	0306-0012	33.383
18	ANNU REV ASTRON ASTR	0066-4146	33.346
19	NAT PHOTONICS	1749-4885	32.386
20	CELL	0092-8674	32.242

Table II. Twenty journals of statistics and probability (n=122) with 2014 IF of interest to those working in the field of medical statistics

Sl. No.	Full Journal Title	2014 IF
1	STATISTICAL METHODS IN MEDICAL RESEARCH	4.472
2	JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY	3.515
3	BIOSTATISTICS	2.649
4	STATISTICS IN MEDICINE	1.825
5	JOURNAL OF COMPUTATIONAL BIOLOGY	1.737
6	BIOMETRICS	1.568
7	JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS	1.494
8	ANNALS OF APPLIED STATISTICS	1.464
9	ANNALS OF APPLIED PROBABILITY	1.454
10	IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS	1.438
11	BIOMETRIKA	1.418
12	STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY	1.127
13	STATISTICAL MODELLING	0.977
14	BIOMETRICAL JOURNAL	0.945
15	JOURNAL OF AGRICULTURAL BIOLOGICAL AND ENVIRONMENTAL STATISTICS	0.91
16	INTERNATIONAL JOURNAL OF BIOSTATISTICS	0.741
17	STATISTICS IN BIOPHARMACEUTICAL RESEARCH	0.618
18	JOURNAL OF BIOPHARMACEUTICAL STATISTICS	0.587
19	METRIKA	0.517
20	JOURNAL OF APPLIED STATISTICS	0.417

Table III. Twenty Indian S&T journals (n=104) of interest to researchers of medical statistics with 2014 IF

Sl. No.	Full Journal Title	ISSN	2014 IF
1	JOURNAL OF FOOD SCIENCE AND TECHNOLOGY-MYSORE	0022-1155	**2.203**
2	JOURNAL OF BIOSCIENCES	0250-5991	**2.064**
3	ANNALS OF THORACIC MEDICINE	1817-1737	**1.803**
4	INDIAN JOURNAL OF MEDICAL RESEARCH	0971-5916	**1.396**
5	INDIAN JOURNAL OF DERMATOLOGY VENEREOLOGY & LEPROLOGY	0378-6323	**1.387**

6	PHARMACOGNOSY MAGAZINE	0973-1296	**1.256**
7	NEUROLOGY INDIA	0028-3886	**1.232**
8	JOURNAL OF CHEMICAL SCIENCES	0974-3626	**1.191**
9	JOURNAL OF PLANT BIOCHEMISTRY AND BIOTECHNOLOGY	0971-7811	**1.094**
10	JOURNAL OF GENETICS	0022-1333	**1.093**
11	INDIAN PEDIATRICS	0019-6061	**1.04**
12	CURRENT SCIENCE	0011-3891	**0.926**
13	INDIAN JOURNAL OF OPHTHALMOLOGY	0301-4738	**0.9**
14	INDIAN JOURNAL OF MICROBIOLOGY	0046-8991	**0.899**
15	INDIAN JOURNAL OF MEDICAL MICROBIOLOGY	0255-0857	**0.882**
16	INDIAN JOURNAL OF BIOCHEMISTRY & BIOPHYSICS	0301-1208	**0.871**
17	INDIAN JOURNAL OF PEDIATRICS	0019-5456	**0.867**
18	JOURNAL OF POSTGRADUATE MEDICINE	0022-3859	**0.855**
19	INDIAN JOURNAL OF EXPERIMENTAL BIOLOGY	0019-5189	**0.835**
20	JOURNAL OF VECTOR BORNE DISEASES	0972-9062	**0.806**

Table IV. 2014 IF of Annual Reviews

Sl. No.	Annual Reviews	ISSN	2014 IF
1	ANNU REV ANAL CHEM	1936-1327	8.833
2	ANNU REV ANIM BIOSCI	2165-8102	3.857
3	ANNU REV ASTRON ASTR	0066-4146	33.346
4	ANNU REV BIOCHEM	0066-4154	30.283
5	ANNU REV BIOMED ENG	1523-9829	14.211
6	ANNU REV BIOPHYS	1936-122X	15.436
7	ANNU REV CELL DEV BI	1081-0706	16.66
8	ANNU REV CHEM BIOMOL	1947-5438	8.676
9	ANNU REV CLIN PSYCHO	1548-5943	12.674
10	ANNU REV CONDEN MA P	1947-5454	14.786
11	ANNU REV CONTROL	1367-5788	2.518
12	ANNU REV EARTH PL SC	0084-6597	8.582
13	ANNU REV ECOL EVOL S	1543-592X	10.562
14	ANNU REV ENTOMOL	0066-4170	13.731
15	ANNU REV ENV RESOUR	1543-5938	5.892
16	ANNU REV FLUID MECH	0066-4189	11.163
17	ANNU REV FOOD SCI T	1941-1413	6.289
18	ANNU REV GENET	0066-4197	15.724
19	ANNU REV GENOM HUM G	1527-8204	8.957
20	ANNU REV IMMUNOL	0732-0582	39.327
21	ANNU REV MAR SCI	1941-1405	14.356
22	ANNU REV MATER RES	1531-7331	11.854

23	ANNU REV MED	0066-4219	12.928
24	ANNU REV MICROBIOL	0066-4227	12.182
25	ANNU REV NEUROSCI	0147-006X	19.32
26	ANNU REV NUCL PART S	0163-8998	11.256
27	ANNU REV NUTR	0199-9885	8.359
28	ANNU REV PATHOL-MECH	1553-4006	18.75
29	ANNU REV PHARMACOL	0362-1642	18.365
30	ANNU REV PHYS CHEM	0066-426X	16.842
31	ANNU REV PHYSIOL	0066-4278	18.51
32	ANNU REV PHYTOPATHOL	0066-4286	9.62
33	ANNU REV PLANT BIOL	1543-5008	23.3
34	ANNU REV PSYCHOL	0066-4308	21.81
35	ANNU REV PUBL HEALTH	0163-7525	6.469
36	ANNU REV RESOUR ECON	1941-1340	1
37	ANNU REV STAT APPL	2326-8298	

On Ascertainment Correction in Rates

S. B. Sarmukaddam
25/4, 'Sangeet Sadhana' Krishna Colony, 11th Lane
Paramhans Nagar, Paud Road, Pune – 411 038, India
E-mail ID: *sanjeev.sarmukaddam@gmail.com*

Abstract
Any prevalence study using many different sources, each of which captures only a fraction of affected cases in the population, is likely to be "incomplete", i.e. sum of all distinct individuals captured by these sources is highly likely to be an underestimate of the true population prevalence. Precise prevalence / incidence rates are central to epidemiology, but they are also very difficult to achieve. A critical component of disease monitoring is the degree of undercount. One way to improve rates is by correcting them for level of ascertainment. Ascertainment level can be considered a primary determinant in the calculation of the rate.

Chandrashekar and Deming's method of estimating number of events missed by both the sources of Sample Registration System proved to be very useful in evaluating vital registration (in India) in many cases where it is suspected to be inferior. Capture-Recapture methodology, which is popular in wildlife literature, can be applied for the same purpose. The primary assumption of both these methods is that of independence of sources, which is unrealistic to be achieved in most of the instances, for example, physicians tend to refer patients to specific hospitals. The technique that addresses dependency between sources is use of log-linear models. Log-linear models are fitted to contingency table, which use the observed cells to predict the missing cell. By predicting the value of missing cell, estimates of the size of population of interest may be obtained. This helps to give an ascertainment correction factor.

Key words: Ascertainment correction, Capture-Recapture methodology, Chandrasekar-Demming Method, Log-linear models

1. CHANDRASHEKAR AND DEMMING'S METHOD

This procedure[1] is generally used (for example by Sample Registration System which is dual system based on continuous vital registration and annual survey in selected Sample Registration Areas) to estimate the coverage of two independent systems for collecting information about demographic or other events, based on assumption that the probability of an event recorded by one system is the same regardless of whether the event is recorded by the other system. When the results of the two systems - such as sample survey and sample vital registration - are matched on such a level, it is possible to obtain a numerical estimate of the degree of completeness of both the systems and hence to estimate the true total number of events.

Suppose that births are recorded for a given year in a sample vital registration system and in a corresponding sample survey (conducted at the end of the year) in which a question on births during the 12 months period preceding the survey is asked. Suppose further that the two sets of birth records so obtained are matched event-by-event. From the matching procedure, the following classification of these events may be obtained.

C = Number of Events Recorded in Both Registration & Survey
N_1 = Events Recorded Only By Registration
N_2 = Events Recorded by Survey
X = Events Missed by Both Systems

	In Registration System	**Not In Registration System**	**Total**
In Sample Survey	C	N_2	S
Not In Sample Survey	N_1	X*	
Total	R		N*

* Estimate.

Thus $(C + N_1)$ is the total number of events registered and $(C + N_2)$ is the total number of events collected by the sample survey. Naturally, there is a possibility that a few events could have been missed by both registration and the sample survey, and X is

that unknown number.

If constant proportion of undercount is assumed for both the agencies, i.e. in statistical terminology, if the chance of detecting an event is independent of not detecting that event for both the agencies, then we have,

$C / (C + N_1) = N_2 / (N_2 + X) =$
$(C + N_2) / (C+N_1+N_2+X) = (C + N_2) / N$

On simplification we get,

$$X^* = (N_1N_2) / C$$

where X* is an estimate of X.

We have thus,

$$N^* = (C+N_1+N_2+N_1N_2) / C.$$

This may be written as:

$$N^* = (C+N_1)(C+N_2) / C$$
$$= RS / C.$$

where R denotes the number of events recorded by registration and S denotes the number of events recorded by the survey.

The completeness of the coverage of the registration system is the matching rate of the survey ($R/N \times 100$ % or $C/S \times 100$ %) and the completeness of the survey is the matching rate of the registration ($S/N \times 100$ % or $C/R \times 100$ %).

The error in estimation of N* is $\sqrt{(N^*q_1q_2 / p_1p_2)}$

where

p_1 = The chance of R (registration) detecting an event.

p_2 = The chance of S (survey) detecting an event.

$q_1 = 1 - p_1$.

$q_2 = 1 - p_2$.

It follows that the better the performance in either R or S, the higher be p_1 or p_2; the smaller be q_1 or q_2 and more precise is the estimate N* of the total of events (as then smaller will be its S.E.).

The precision of N* expressed as a proportion namely as a coefficient of variation is $\sqrt{[(q_1q_2 / N^*p_1p_2).\ p_1]}$ is a measure of performance of the registration system (i.e. registrar) an estimate for which is $p_1 = C / (C+N_2)$. This estimate p_1 is subject to a coefficient of variation of

$$\sqrt{[\{q_1 / (C+N_2)\ p_1\}\ \{N^* - C - N_2 / (N^*-1)\}]}$$

This error decreases as $C+N_2$ increases. For perfect performance on the part of the survey interviewer, $C+N_2 = N$ and there is then no error in estimating the performance of the registrar.

2. CAPTURE - RECAPTURE METHOD

Capture-Recapture methodology[2] has a long history and is very popular in wild life literature of population censuses. The following wild life example briefly illustrates the simplest form of a two-sample capture-recapture model.

A sample of animals (M) is captured, marked and then released. Subsequently, a second sample of animals (n) is captured. An estimate of the population size (N) can then be calculated based on the number of marked animals captured in the second sample (m).

$$N_{(C-R)} = \{[\ (M+1)(n+1) / (m+1)\ ] - 1\}.$$

The variance in estimate of $N_{(C-R)}$ can be estimated by

$$\text{Var}\ (N) = \frac{[(M+1)(n+1)(M-m)(n-M)]}{[(m+1)^2(m+2)]}.$$

and the confidence (95%) interval is constructed by

$$N \pm 1.96\ [\sqrt{(\text{Var}(N))}].$$

In epidemiology, capture-recapture techniques allow the number of cases of disease in a defined population to be estimated using two (or more) sources of cases. Potential sources often include hospital records, private practice physicians, disease registries, death certificates, or other patient's lists. Taken alone, each of these sources may considerably undercount the actual number of cases. However, by using information provided by duplicate cases (cases found in more than one source), an estimate of the number of individuals not identified can be calculated. From the estimates of the number of "uncounted" cases, an "ascertainment corrected" rate can be derived.

3. GENERAL FRAMEWORK

If two sources X and Y are independent and the structure of the population is as shown in the following figure:

Fig: Data Structure By Source of Ascertainment:-

		Cases reported in Source Y: Yes	Cases reported in Source Y: No	
Cases reported in Source X	Yes	a	b	a+b
	No	c	d=?	
		a+c		

where a: Cases reported in both X and Y
b: Cases only in source X and
c: Cases only in source Y

Then the maximum likelihood estimate (MLE) of cases in the population but not in X and in Y is:

$$d_{MLE} = (bc) / a$$

The corresponding estimate of the total population N is

$$N_{MLE} = (a)+(b)+(c)+([bc] / a)$$

or equivalently $N_{MLE} = (a+b)(a+c) / a$

While these are MLEs, they are biased for small samples and a preferable Nearly Unbiased Estimator (NUE) for unascertained cases is

$$d_{NUE} = (bc) / (a+1)$$

or total number in population is

$$N_{NUE} = (a)+(b)+(c)+[(bc) / (a+1)]$$

or equivalently

$$N_{NUE} = \{[(a+b+1)(a+c+1)] / [(a+1)] \} - 1$$

It is shown to be unbiased for wide ranges of parameter values[3]. It may be noted that d_{MLE} and Chandrashekar - Demming's X* are same. And so N_{MLE} and N* are same. N_{NUE} is same as N_{C-R}.

If two sources are used, designated as A and B, there are three possible ways in which a case can be ascertained: only in A, only in B, or in both A and B (denote this as AB). If there are three sources, A, B and C, then there are seven nonoverlapping possible modes of ascertainment: A only, B only, C only, AB not C, AC not B, BC not A and ABC. In general, if there are K possible sources, the number of different possible combinations of non-overlapping sources of ascertainment is $[(2^K)-1]$.

Occasionally, there is an advantage to pooling various types of sources and treating their union as a single source[4]. If investigators keep track of the nature of the ascertainment of cases by source and collect & report data of number of cases by source intersection, then they (or at least others), may derive estimates of missing cases and of the total population affected[3].

Both the above methods require that the population be closed to additions or deletions, that individual identifiers are not lost or overlooked, and that for a given source each case is equally likely to be listed by that source (i.e. independence). The first two assumptions are not usually problematic as these populations are relatively stable and unique identifiers such as names are rarely lost. The most difficult assumption, however, is that of independence, as truly independent sources rarely occur. For instance, physicians tend to refer patients to specific hospitals.

If a list from, say, a group of general practitioner and from medical specialists is used, and cases seen by the first tend to be referred to the second, as is quite likely, these sources are positively dependent and the resulting estimate is too low. A negative dependence would result in an overestimate[5]. Therefore, a positive association or dependence of two sources will result in an underestimate of the true population prevalence and a negative dependence in an overestimate[6].

Although such estimates generated from two sources are not necessarily accurate (because of the independence assumption) they may be useful. If one has previous knowledge of likely dependence of sources used, one may at least put a likely upper or lower boundary on the number of cases affected. For instance, spina bifida (in live births) on birth or death certificates are likely to be positively dependent[6]. The capture-recapture estimates of those affected from these sources is thus biased to be low and provides a minimum estimate of the total affected, a figure that may be valuable for several purposes - e.g., economic justification of preventive programmes.

Therefore, even if the sources are not independent, the derived values of d and N (defined earlier) may be very useful in evaluating estimates of prevalence rates for the population. One may have some independent knowledge of the likely direction of the independence of any of the two sources that have been used. Thus, if one knows that two particular sources are likely to be positively dependent, then the estimates derived from the equations above and the boundaries of their confidence intervals may be regarded as plausible lower limits of the true values. If there is likely to be a negative dependence, then the estimates and the boundaries of their confidence intervals become plausible upper limits of true values.

The "efficiency" or "completeness" of any source may be defined as the estimated proportion of cases in the population listed in that source[7]. A simple approach to determine the validity of the

assumption of the independence is accomplished through pair wise comparison of different sources of ascertainment. If pair wise estimate provides a similar estimate of the number of cases, then it is unlikely that there is any major bias[8].

Any prevalence study using many different sources, each of which captures only a fraction of affected cases in the population, is likely to be incomplete i.e. the sum of all distinct individuals captured by these sources is highly likely to be an underestimate of the true population prevalence. Therefore, an alleged complete enumeration census, which is generated from pooling lists of cases from different sources, should be regarded with scepticism unless capture-recapture analysis is done to confirm the claim of completeness. For example, a recent analysis of data from a prevalence study that had claimed 100% ascertainment, (whose exhaustive methods, suggested complete ascertainment) estimated that about 25-40 % of cases were missed[3].

In some articles[6] for estimates derived assuming independence of three sources, the following equation is solved for N.

$$\prod_{(i = 1\text{-}3)} (N - n_t) = M^2 (N - n_t)$$

where n_t is the total number of distinct names in all three sources.

The variances were estimated by the formula:

$$\text{Var } (N) = N [(1 - p)^{-1} + (k - 1) - (1 - pi)^{-1}]^{-1}$$

where $pi = ni / N$ and $p = 1 - \Pi (1 - pi)$

and k is the number of distinct lists (2 or 3) used in reaching estimate. Useful over-view of this methodology is given in books[9,10].

4. OTHER METHODS

The simplest method is use of the truncated binomial (similar to truncated Poisson). This method requires information only on the total number of sources in which each distinct name appears. The parameters are estimated from the observations in other cells and the estimate of the missing cell is made from the estimated parameters. This approach assumes that reporting to each source is independent, and that each individual has the same, constant probability of being captured by each source.

The second method is in essence a generalization of the first. Here the underlying distribution of the number of sources reporting a given case is assumed to be a truncated β-binomial distribution. In this model, it is assumed again that sources are independent, but the probability of being captured by any specific source, while constant for any individual, varies among individuals according to a β distribution.

The third method[11] known as "Bernoulli census approach" is analogous to capture-recapture methods. If many different sources are used and if all are mutually independent then a single maximum likelihood estimate of the total prevalence can be reached which provides the optimal estimate[4]. This type of Bernoulli census estimate may be termed as "independent sources".

The fourth approach is, use of log-linear methods for estimating the value of a "missing cell" in 2^k cross-classification, where k is the number of sources. The approach, an analogue of analysis of variance applied to categorical data, provides an alternative method for analysis of and adjustment for dependencies ("interactions") of the data sources.

Each of the methods makes certain assumptions about ascertainment. The truncated binomial model makes the most rigid assumptions:

(i) for any specific source, each case in the population has the same probability of being listed by that source;

(ii) there is no variation between sources in these probabilities i.e. the ascertainment probability is the same for all sources; and

(iii) the sources are independent.

The β-binomial model makes the second and third assumptions, the independent sources model makes the first and third assumptions and the log-linear model makes only the first assumption.

A study which compares all these above four methods/approaches[12] concludes that, in general, the log-linear model appears preferable for prevalence estimation. In the section below, log-linear models method is described with example.

5. USE OF LOG-LINEAR MODELS

The technique that addresses (rather adjusts) dependency between sources is the use of log-linear models for incomplete multiway contingency tables. Using this technique involves placing data in an

incomplete 2^k contingency table, where k represents the number of sources selected for case ascertainment. This table displays the recapture history of all cases in the community, with the exception of one missing cell that corresponds to the number of unascertained cases in the population. Log linear models are then fitted to this contingency table that uses the observed cells to predict the missing cell. This can be easily achieved by defining this missing cell as "structural zero". The term "structural zero" is used to refer to a cell whose probability of containing an outcome is known to be zero. Note that we are not referring to cells with zero frequencies (because very few observations were collected) but rather to cells that are naturally empty (for example, number of males suffering from a particular gynaecological disorder).

By predicting the value of the missing cell, estimates of the size of the population of interest may be obtained. Source dependencies may be modelled using interaction terms that are adapted to the simplest model (independent sources) until no significant improvement of fit occurs.

To illustrate the application, one published example where this method is used[2] is given here. To find the incidence of ALS in Harris county, Texas for the years 1985-1988, cases were identified from three sources: Neurologist, Hospital and death certificates. The data obtained can be put in table like:

		Death Certificates	
Neurologist	Hospital	Pr.	Ab.
Pr.	Pr.	23	28
	Ab.	2	9
Ab.	Pr.	14	11
	Ab.	10	*

As can be seen, the cell corresponding to the number of cases not found by the neurologist, hospital or the death certificates is missing and denoted in the table by *. That means * represents the number of cases not ascertained in the population.

A model with the main effects and two interaction terms, neurologist & hospital, hospital & death certificates provided a good fit of the data (G^2 = 0.80, P = 0.37). This model, which controlled for dependencies between the sources, estimated that 44 cases had not been ascertained from the population. Adding the estimate of 44 uncounted cases to the 97 ascertained cases provide a total estimate of 141 cases in the population.

REFERENCES

1. Chandrasekar C, Deming WE. On A method of Estimating Birth and Death Rates And the Extent of Registration. *J Am Stat Asso (JASA)* 1949; 44: 101-115.
2. McCarty DJ. Ascertainment Corrected Rates: Applications of Capture-Recapture Methods. *International J Epi* 1993; 22: 559-565.
3. Hook EB, Regal RR. The Value of Capture-Recapture Methods Even for Apparent Exhaustive Surveys. *Am J Epi* 1992; 135: 1060 -1067.
4. Wittes JT. Applications of a Multinomial Capture-recapture Model to Epidemiological data. *J Am Stat Asso (JASA)* 1974; 69: 93 - 97.
5. Hook EB, Regal RR. Capture-Recapture Methods: Letter to Editor. *Lancet* 1992; 339: 742.
6. Hook EB. Use of Bernaulli Census And Log-Linear Methods for Estimating the Prevalance of Spina-bifida in Livebirths and the completeness of Vital Record Reports in New York State. *Am J Epi* 1980; 112: 751-758.
7. Wittes JT. Capture-Recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis* 1974; 27: 25-36.
8. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with application to epidemiological research. *J. Chronic Dis* 1968; 21 : 287 - 307.
9. Sarmukaddam S.B. (2006), *Fundamentals of Biostatistics*. Jaypee Brothers Medical Publishers Ltd., New Delhi, India.
10. Indrayan A. and Sarmukaddam S.B. (2001), *Medical Biostatistics*. Marcel Dekker, Inc., New York, USA.
11. Wittes JT. Estimation of population size: Bernoulli census. Ph.D. Dissertation, 1970, Harward University, Cambridge, MA, USA.
12. Hook EB, Regal RR. Validity of Bernoulli census, Log-linear and Trucated Binomial Models for correcting Underestimates in Prevalance Studies. *Am J Epi* 1982; 116: 168-176.

Changing Practice in Statistical Methods:

In Logistic Regression and Validation of Randomization using P values in Clinical Trials

Visalakshi J and L. Jeyaseelan

Department of Biostatistics, Christian Medical College, Vellore – 632002, India

E-mail ID: *visali_pv@hotmail.com* and *ljey@hotmail.com*

Abstract:

Background: Statistical data analysis is fundamental to any research.Statisticians have provided scientists with tools of enormous power to find order and meaning in the most complex of datasets, and scientists have embraced them with glee. They have not, however, embraced statistics *education* and thereby still lack appropriate training in research. Therefore, we have attempted to present two of the inappropriately used statistics in research due to lack of update in statistical techniques: (1) Islogit link still appropriate to obtain correct RR in cohort and cross sectional studies?(2) Validation of randomization using p values in clinical trials.

Methods: Log binomial model or robust Poisson models are reasonably good models for estimating RRs when the prevalence is over 10% and sample sizes are reasonably large. However, recent updates on these models suggest that Robust Poisson models are better than Log Binomial models in terms of convergence and also when outliers are present. Baseline differences are usually reported with p values in clinical trials. Many studies have pointed out how these p values can be misleading in clinical trials to report baseline differences. Instead studies recommend presenting the distribution of baseline characteristics and check for differences if any clinically rather statistical.

Discussion and Conclusion: There should be ways to update ourselves through good training and the developments in methods must be emphasized to editors, researchers, and statisticians.It is also important to encourage young scientists to read good articles which have been constantly promoting good research through recent developments.

Background:

Statistical data analysis is fundamental to any research. A good research focuses on all components of statistical procedures including data collection, analyses, interpretation and reporting. Open a random page in your favorite medical journal and you'll be deluged with statistics: *t* tests, *p* values, proportional hazards models, risk ratios, logistic regressions, least-squares fits, and confidence intervals. Statisticians have provided scientists with tools of enormous power to find order and meaning in the most complex of datasets, and scientists have embraced them with glee. They have not, however, embraced statistics *education*, and many graduate programs in the sciences still lack statistical training. Since the 1980s, researchers have described numerous statistical fallacies and misconceptions in the popular peer-reviewed scientific literature, and have found that many scientific papers – perhaps more than half – fall prey to these errors. Various clinical reviews have been consistently finding many statistical flaws and reported that nearly 50% medical articles contain statistical errors and most of them are in selection of wrong statistical or sampling methods(1). The misuse of statistics has been discussed extensively and this fact has been made clear that inappropriate usage is both unethical and constitutes serious clinical hazards(2). Altman(3)has written by quoting the medical researcher mistake to the medical person as "Is it ethical when a doctor uses a wrong treatment either wilfully or through ignorance or who uses the right treatment but with the wrong dose?" The other important concern is that journal editors too have very limited updates on the latest developments to direct researchers towards good science.

Huge sums of money are spent annually on research that is seriously flawed through the use of inappropriate designs, unrepresentative samples, small samples, incorrect methods of analysis and faulty interpretation. "Publish or Perish" concept compelled researchers to carry out research with poor quality. Studies following these type of concept take this as good example and start designing or publishing bad studies. Thus few bad examples serve as model for forth coming bad examples. Therefore, responsible medical journals invest considerable effort in getting papers refereed by statisticians. There are very few papers that get

rejected on statistical grounds (4). Altman et al(5)reported that of the papers published in BMJ and Annals of Internal Medicine from May through August 2000, 85% of the papers received input from Biostatisticians or epidemiologists, 71% of papers were rejected without statistical assistance were rejected without review as compared to 57%, which received assistance (p<.001).

The question is how do we break the cycle and train researchers with good reporting designs and appropriate techniques to address the issues correctly. Therefore, we have attempted to present two of the inappropriately used statistics in research due to lack of update in statistical techniques. I) Islogit link still appropriate to obtain correct RR in cohort and cross sectional studies? and II) Validation of randomization using p values in Clinical trials.

Methods and Results:

Odds Ratios –Can we still rely on logit link in logistic regression models to obtain OR when the prevalence is more than 10%? Is there any better link function or models?

Current Practice:

When the outcome of any study is binary, the most common method of analysis is to calculate an Odds Ratio (OR) to find the strength of association with exposure for any study design. To estimate Odds Ratio adjusting for confounders in Cohort studies or any baseline differences in Clinical trials, logistic regression method is applied. This estimated OR in cohort studies is then corrected using Zhang and Yu to obtain Risk ratios (RR). However, when the outcome prevalence is >10%, OR is no longer an acceptable estimate of RR as it always overstates and sometimes dramatically(6,7). There are good amount of literature available to educate a researcher that they should apply log link function which provided risk ratio. This is more appropriate when we use cross sectional studies and this is termed as "Prevalence ratio" instead of "Risk ratio". The prevalence of poor mental health from the IndiaSAFE study was 38.5 (35.03 - 35.97). Therefore ideally we should use log binomial method. If someone uses logit link based method then the researchers would overestimate the risk. In the following table the OR for any physical violence was 2.79 by logit link function method, however this should be ideally 1.9. The prevalence ratio is 1.9, based on log link function based analyses. Thus OR overestimated this estimate. This is because the prevalence of mental health in this study (outcome rate) was 38.5%. However there are many occasions when log binomial models had convergence issues.

Change in practice: Log binomial model or robust Poisson models are reasonably good models for estimating RRs when sample sizes are reasonably large. However, of the two models, robust Poisson models are much preferred. Robust Poisson regression is not prone to convergence problems unlike log-binomial regression models(8).For example, consider the data presented in Table 2. Fitting a log-binomial model to this data ends in failed convergence in R (9), STATA (10)SPSS(11)after100 iterations. However, simulation studies have also shown that Robust Poisson regression models are quite robust even in the presence of outliers when estimating RRs. Robust Poisson regression model consistently outperformed even when there were high order terms included in the linear and non-linear models as well. Data generated was contaminated as follows: 0% - original dataset; 2% - flipping outcome of records with p at bottom or top 1%; 5% - flipping outcome of records with p at bottom or top 5% .

The flipping approach produced outliers that are more likely to be leverage points therefore likely to affect the estimates. For above scenario, simulation process was repeated 1000 times and relative bias for log binomial and robust Poisson model is shown below:

When the data was not contaminated (0%) log binomial and robust Poisson models produced accurate estimates of regression coefficients. However, when contamination increased by just 2% or 5%, robust Poisson models produced less bias than log binomial models for linear, non-linear models for moderate and strong levels of association(12).

Validation of randomization using P values in clinical trials – Wrong practice:

The main 'ingredient' that adds flavour in any Clinical Trial design is the concept of Random allocation of treatments to the subjects which places Clinical trial on the top of the design triage in epidemiological designs. Randomization is defined as a process of allocating treatments to subjects in a

random manner. In other words, all subjects have an equal and known chance of receiving any of the treatments and in the process of doing this all the known and unknown confounders (baseline and unstudied variables) get well balanced across the groups, especially in a reasonably large sample study. However there is no guarantee that the subjects allocated to the different treatment groups will be similar with respect to all important characteristics, such as age, severity of disease, type of disease, and so on (13). Therefore the challenge is how do we make sure that there is no imbalance in the baseline variables, while the imbalance of baseline variables is clearly undesirable.

Purpose of table 1: It has been prevailing in practice that p values are provided to compare baseline variables (socio demographic and baseline outcome variables), usually presented as table 1 in any trial related paper. The purpose of table 1 is twofold. First is regarding generalizability of this information. The treatment could be used for subjects whose age is in this range, and whether this treatment could be prescribed for both males and females and so on. The second and foremost reason is to study whether these variables between the study groups are well balanced or not. In order to know whether they are well balanced or not, researchers use testing hypothesis (doing t test or chi-square test etc) rather than assessing the difference between the two or more groups clinically.

Absurd: The philosophy of a controlled trial is to compare groups that differ only with respect to their treatment, while the baseline variables are balanced. Also that, we need to be concerned only when the variable in question is related to the outcome variable, that is, when it is of prognostic importance. An obvious common example of such a variable is age which is often related to outcome, especially in survival studies (13). As presented by Altman (1985) (13), "although the intention behind performing a statistical test of significance may well be to see whether an observed difference is a real or important one, the test actually assesses the probability (the 'P value') that the observed difference, or a greater one, could have occurred by chance when in reality there was no difference. In a clinical trial in which treatment allocation was properly randomized, a difference of any sort between the two groups at the time of entry to the trial will necessarily be due to chance, since randomization precludes any external influences (biases) on which subjects receive which treatments. Putting these two ideas together, performing a significance test to compare baseline variables is to assess the probability of something having occurred by chance when we know that it did occur by chance. Such a procedure is clearly absurd".

Absence of Evidence is not an Evidence of absence of difference/efficacy: When we do a test, we need make sure that we have a hypothesis and there is a research question. Is it our research questionto test age or gender or Body Mass Index of the two groups' baseline variables? Have we decided the sample size based on the age or gender data and therefore we are comfortable in interpreting the p value? If no difference in mean (sd) of age between the two groups, it does not mean that these are similar. It might mean that the sample size may not be enough or power is low for that variable. If we start interpreting equivalence because there is no statistical significant difference, every researcher would take a convenient number and do testing hypothesis, and if no difference then they would conclude equivalence. Testing equivalence is totally new and different concept. For an example, (Fentiman et al; 1983) (14) compared the control of pleural effusions in patients with breast cancer given either mustine or talc. Four baseline characteristics of the two groups are shown in table (3). All these variables have shown larger differences that are clinically important. The age difference was nearly 5 years, and the stage of disease was over 20% different, yet these variables differences were not statistically significant and therefore no further analyses were done to control these differences. Sometimes the researchers wonder whether these differences were due to interference in the randomization?.

Discussion:

Researcher's failure to understand ORs and when do we obtain good estimates of RR is not known clearly. This is due to lack of training and updating themselves through reading the research methods, epidemiological and statistical journals. Researchers have been using logistic regression in cohort studies, cross sectional studies till recently but there have been studies from early 2000

emphasizing the need to change to better models like log binomial model to calculate RRs instead of logistic regression(6).There have been recent updates on log binomial models which have led to Robust Poisson models due to convergence issues and in the presence of outliers that one may commonly face in any regression methods(15).Another good example is comparison of baseline characteristics in a trial which was in practice until early 1990 (1,16,17). Then series of articles have been written in British Medical Journal and Lancet to educate researchers regarding the methodological issues involved in it (18–21). These kinds of challenges are expected to grow more as statistical methods have been evolving. The current decade has witnessed a steady growth in some techniques like multilevel modelling in analysing cluster design or repeated measures (longitudinal) data. The reviewers ought to know the statistical methods which are better for estimation and for the CIs. Bootstrap methods are also some of the other techniques that could be applied whenever the violations of any of the parametric assumptions are present which may be prominent especially in a small sample size. The design related issues are expected to be more in journals from developing countries as there is no good infrastructure and human resource in the country to train researchers. The universities which train statisticians often fail to provide them hands on with real time data to validate the assumptions and demonstrate the good and bad use of methods for a particular research question. Lack of qualified Biostatisticians is the main challenge in developing countries. The self-made epidemiologists many times act as biostatisticians which thereby leads authors to wrong application of statistical techniques. What is not transparent when the articles get published is whether the wrong method was used because of authors or because of reviewer. There must be a system to mention that this method was used as requested by the editorial board. In the country there must be a system to update the reviewers and writers on recent developments and therefore, asking for correct methods, for example, when computing risk in cross sectional studies whether to reporting ORs or RRs especially when the prevalence is more than 10%.

In summary, as the statistical science is evolving rapidly, there is a need for each of us to update recent developments to avoid mistakes which have been existing all through the year. There must be a network for researchers to work in groups updating latest developments as it may be tedious for a single person to do so. Also, individuals must also be encouraged to participate in forums which are being grouped by some software companies like R, STATA, SAS as some researchers pose queries which becomes a foundation for a better approach. At the minimum there must be some training in the country, until then this may not be minimized. The journals should also invest money to hire qualified Biostatisticians to improve their standards.

References:

1. Curran-Everett D, Benos DJ, American Physiological Society. Guidelines for reporting statistics in journals published by the American Physiological Society. Am J Physiol Endocrinol Metab. 2004 Aug; 287(2):E189–91.
2. Gardenier JS, Resnik DB. The misuse of statistics: concepts, tools, and a research agenda. Account Res. 2002 Jun;9(2):65–74.
3. Altman DG. The scandal of poor medical research. BMJ. 1994 Jan 29;308(6924):283–4.
4. Bailar JC., Mosteller F. Bailar JC, Mosteller F, eds. Medical uses of statistics. Communicating with scientific audience In.
5. Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. JAMA. 2002 Jun 5;287(21):2817–20.
6. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol. 1987 May;125(5):761–8.
7. Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP, Groenwold RHH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. CMAJ Can Med Assoc J J Assoc Medicale Can. 2012 May 15;184(8):895–9.
8. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. Am J Epidemiol. 2011 Oct 15;174(8):984–92.
9. R Development Core Team (2008). R: A language and environment for statistical

computing. R Foundation for Statistical Computing, [Internet]. Vienna, Austria; Available from: http://www.R-project.org

10. StataCorp. 2013. Stata Statistical Software. Texas; 2011.
11. IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp. NY; 2010.
12. Chen W, Shi J, Qian L, Azen SP. Comparison of robustness to outliers between robust poisson models and log-binomial models when estimating relative risks for common binary outcomes: a simulation study. BMC Med Res Methodol. 2014;14:82.
13. Altman DG. Comparability of randomized groups. The Statistician 1985; 34: 125-136.
14. Fentiman IS, Rubens RD, Hayward JL. Control of pleural effusions in patients with breast cancer. A randomized trial. Cancer. 1983 Aug 15;52(4):737–9.
15. Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003 Oct 20;3:21.
16. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002 Oct 15;21(19):2917–30.
17. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet Lond Engl. 1990 Jan 20;335(8682):149–53.
18. Altman DG. Statistics and ethics in medical research. Misuse of statistics is unethical. Br Med J. 1980 Nov 1;281(6249):1182–4.
19. Altman DG. Statistics and ethics in medical research. VII--Interpreting results. Br Med J. 1980 Dec 13;281(6255):1612–4.
20. Altman DG. Statistics and ethics in medical research. VI--Presentation of results. Br Med J. 1980 Dec 6;281(6254):1542–4.
21. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. Br Med J. 1977 Jan 8;1(6053):85–7.

Table 1: Comparison between Logistic regression and Log binomial model result using Poor mental health and important risk variable

Variables	Logistic Regression			Log Binomial Model		
	Odds Ratio	Std Error	P value	Risk ratio	Std Error	P value
Dowry harassment	0.61	0.03	0	0.87	0.01	0
HPP during childhood	1.3	0.06	0	1.12	0.03	0
Witnessed father beating mother	1.34	0.07	0	1.09	0.03	0
Drunk	1.27	0.03	0	1.1	0.01	0
Any physical violence	2.79	0.14	0	1.9	0.06	0

Table 2 Relative bias (%) in log scale (n=1500)Chen W et al. 2014.

RR	Prob (Y=1)	Contamination rate	Association bet Zand Y: linear		AssociationBet Z andY: linear		AssociationBet Z andY: Non-linear		AssociationBet Z andY: Non-linear	
			Level of association bet Z and X,Z and Y: Moderate		Level of association bet Z and X, Z and Y: Strong		Level of association bet Z and X,Z and Y: Moderate		Level of association bet Z and X,Z and Y: Strong	
			LB	RP	LB	RP	LB	RP	LB	RP
1.5	10%	0%	1.8	1.8	2.0	2.0	0.0	0.0	-0.6	-0.8
		2%	-17.8	-16.7	-16.0	-16.0	-30.0	-19.5	-28.5	-20.4
		5%	-43.4	-36.7	-41.8	-36.9	-51.7	-41.3	-49.1	-43.0
	25%	0%	-0.4	-0.4	0.4	0.3	0.5	0.5	-0.8	-0.8
		2%	-10.4	-10.9	-10.1	-11.4	-14.1	-11.8	-18.6	-16.2
		5%	-25.3	-24.4	-26.6	-27.3	-34.6	-28.2	-43.4	-36.2
	40%	0%	0.2	0.2	-0.7	-0.7	-0.0	-0.1	0.9	0.7
		2%	-7.5	-8.0	-9.8	-10.8	-11.0	-10.5	-14.9	-13.1
		5%	-19.2	-19.4	-23.4	-24.6	-26.9	-24.7	-36.9	-32.2
2.0	10%	0%	0.4	0.4	1.5	1.5	-0.2	-0.2	0.6	0.7
		2%	-18.9	-18.1	-17.1	-16.8	-26.3	-19.0	-25.0	-18.0
		5%	-42.0	-37.5	-40.2	-36.7	-47.9	-39.5	-47.4	-39.6
	25%	0%	0.5	0.5	0.3	0.3	0.3	0.3	-0.3	-0.3
		2%	-9.0	-9.2	-9.2	-9.9	-12.3	-10.4	-13.7	-11.7
		5%	-23.3	-22.2	-23.7	-23.6	-30.0	-24.7	-34.5	-28.5
	40%	0%	-0.1	-0.1	0.1	0.1	0.3	0.3	-0.3	-0.3
		2%	-6.7	-7.0	-7.2	-7.8	-8.4	-8.0	-10.9	-10.2
		5%	-16.9	-16.6	-18.5	-19.0	-21.2	-19.5	-26.6	-24.5

Note: Association between Z and X always linear
LB: Log Binomial Model
RP: Robust Poisson
Relative Bias was defined as the average of 1000 estimated RR in log scale minus log of the true RR divided by log of true RR

Table 3: Some baseline variables of patients in controlled trial of Fentiman et al (1983)

	Treatment	
	Mustine (n=23)	**Talc (n=23)**
Mean age (SE)	50.3 (1.5)	55.3 (2.2)
Stage 1 or 2	52%	74%
3 or 4	48%	26%
Mean interval in months (SE)Between breast cancer diagnosis and effusion diagnosis	33.1 (6.2)	60.4 (13.1)
Postmenopausal	43%	74%

Assessment of association in contingency tables with structural zeroes

Subha, P.T. [a], Thennarasu, K [b], Janardhan Reddy Y.C. [c] , Subbakrishna, D.K [b]

NSSO (DPC), MOSPI, Bangalore [a], Department of Biostatistics [b],

Department of Psychiatry [c], NIMHANS, Bangalore.

E-mail ID: *kthenna@gmail.com*

Introduction

Chi-square method is often employed for a test of independence between two variables presented in a two-way contingency table. In fact, this assumption of independence may be the very base for model building and advanced analysis of the data at next stages. Clogg & Shihadeh (1994) suggested that when the contingency table produces sparse data, it is always better to consider alternative methods for analysis. The meaning of sparse data is several or many cells in the contingency table may have small frequencies and possibly sampling and structural zeroes as well.

Limitation of Chi-square test

The empty or zero cells in a contingency table can be classified as structural zeroes or sampling zeroes (Everitt, 1977). Sampling zeroes arise owing to sampling variation when a relatively small sample is collected for a table having a large number of cells. An obvious way to eliminate these is to increase the sample size which may not be feasible always. A sampling zero cell may have nonzero expected value. When zero arises owing to certain combinations of variables being impossible, it is known as structural zero. A cell with a structural zero has an expected value of zero, which means that not only did no observations in the data set at hand fall into that cell, but also no observation *could* fall into that cell. Structural zeroes are not part of the data. Therefore, they should not contribute to the Chi-square or model fitting. A contingency table containing structural zeroes is, in some sense, an incomplete table. Hence the usual χ^2 test cannot be applied directly in such cases.

Quasi Independent (QI) Model

Suppose we have a square table under consideration and the two variables are commensurate with one another (Clogg and Shihadeh, 1994). We can partition the cells in the table into two sets P and Q, with P U Q equal to the entire set of cells. Here P may consist of the cells on the main diagonal, or perhaps just some of the diagonal cells or perhaps cells on or close to the main diagonal or structural zeroes. Then the QI model is defined as

$$\log(F_{ij}) = \lambda + \lambda_{A(i)} + \lambda_{B(j)}, (i, j) \in Q \qquad (1)$$

where F_{ij} is frequency of i[th] row and j[th] column, A denote rows, B columns and λ's are used to denote either log N or log $\prod$. ($\prod$ is the marginal probability of corresponding column/row)

As the Chi-square test fails to give an exact picture of association between the variables when dealing with sparse data and the presence of sampling/structural zeroes further complicates the assessment. We attempt to study the application of loglinear modelling, in which the cells with zero entries due to sampling zeroes can have nonzero expected value. The QI model represents a way to deal with contingency tables with structural zeroes with a simple modification.

The data

The live data on Obsessive Compulsive Disorder (OCD) from OCD clinic register, NIMHANS, which was collected during the month July 2006, was used for the analysis. The sample size was 543. The objective of the study was to assess any association between two variables 'age at onset' and 'birth cohorts'. The speciality of the data set at hand is the presence of structural zeros in the table. Such data arise in the field of genetics more often and deserve a special mention. Even a seasoned statistician may overlook the structural zero and apply statistical techniques in the usual manner leading to wrong conclusions.

Due to the development of modern science, we may believe that the possibility of detection of the disease in a particular age group (say, 10-20), for one birth cohort (say, 1950-60), may be different from that of another birth cohort (say for 1990-2000) or in other words the age at onset and the birth cohort are associated with each other. We want to check the null hypothesis that the birth cohort and age at onset are independent.

Plan of analysis

For the two way contingency table, the association between variables age at onset and birth cohorts was tested using chi-square test of independence and based on the chi-square statistic (χ^2), the conclusion was arrived. The association was then rechecked employing loglinear analysis. At the first level, in order to tackle the structural zeroes present in the table, quasi-independent loglinear model method was used. For this a weight of zero was given to the cells with structural zeroes and weight of one was given to the remaining cells. Using these weights the statistic χ^2 was computed. While checking the contribution of each cell towards chi-square value, it was found that it is more from the cells closer to structural zeroes. Hence a need to incorporate the possibility of detecting the case with given age of onset during the period of study was felt. Then appropriate weights to different cells were calculated and QI model with these smoothed weights considering possibility of detection assigned to the cells was formed. Finally $\chi^2$was calculated using smoothed weights. The χ^2 values under the three models were examined to reach at the final conclusion on the hypothesis.

For the analysis of the above QI model with different weights for the data and the model fit by Pearson's χ^2, a computer program was written in R (R Development Core Team, 2011)) The specific library used were 'gnm'(Turner and Firth, 2011), 'stats' and 'MASS' (Venables and Ripley, 2002).

Assessment of association in the presence of structural zeroes

As said earlier, age at onset and birth cohorts were two variables of interest associated with the study on Obsessive Compulsive Disorder (OCD). One of the basic objectives of the study was to find association between the birth cohort and age at onset i.e., the birth cohort influence the age at onset of the disease. Table 1 gives the birth cohort-age at onset classification table.

Table: 1 Distribution of the age at onset and birth cohort

Age at onset (years)	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-'00	2000-'10
0-10	0	0	1	1	8	19	1	0
10-20	0	0	4	17	73	185	3	***0***
20-30	0	2	4	17	99	29	***0***	***0***
30-40	2	4	5	28	14	***0***	***0***	***0***
40-50	1	3	15	4	***0***	***0***	***0***	***0***
50-60	0	4	0	***0***	***0***	***0***	***0***	***0***
60-70	1	1	***0***	***0***	***0***	***0***	***0***	***0***

This table had many cell with very small frequencies and zeroes in ***bold italics*** represents the structural zeroes which arise due to impossible combination of the variables. At first, ordinary chi-square test was performed for testing the hypothesis of independence. Here each of the entries in the table was given equal weight of 1 (Table 2).

The '1' in italics denotes the cells where structural zeroes were present and still they were taken into account in the analysis. The result of this analysis was equivalent to the independent loglinear model analysis (Table 3). The result indicated that there is a strong association between cohort and the age of onset (χ^2 = 381.57, p<0.0001). However, it has been established that in the case of small sample sizes and sparse tables, the chi-square approximation will generally be poor (Vermunt, 2005).

Table: 2 Initial weight allocations for chi-square test

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-2000	2000-2010
0-10	1	1	1	1	1	1	1	1
10-20	1	1	1	1	1	1	1	***1***
20-30	1	1	1	1	1	1	***1***	***1***
30-40	1	1	1	1	1	***1***	***1***	***1***
40-50	1	1	1	1	***1***	***1***	***1***	***1***
50-60	1	1	1	***1***	***1***	***1***	***1***	***1***
60-70	1	1	***1***	***1***	***1***	***1***	***1***	***1***

From the Table 3, it is clear that, the above sparse table had many cells with small frequency and expected values of some of the structural zeros cells were positive quantities. This resulted in inflated chi-square value leading to the wrong conclusion that the variables under consideration are associated.

Table: 3 Expected values under independence model

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-‘00	2001-‘10
0-10	0 (0.22)	0 (0.77)	1 (1.6)	1 (3.69)	8 (10.68)	19 (12.83)	1 (0.22)	0 (0)
10-20	0 (2.07)	0 (7.24)	4 (15.01)	17 (34.67)	73(100.38)	185(120.56)	3 (2.07)	***0 (0)***
20-30	0 (1.11)	2 (3.88)	4 (8.04)	17 (18.56)	99 (53.75)	29 (64.56)	***0 (1.11)***	***0 (0)***
30-40	2 (0.39)	4 (1.36)	5 (2.82)	28 (6.52)	14 (18.87)	***0 (22.66)***	***0 (0.39)***	***0 (0)***
40-50	1 (0.17)	3 (0.59)	15 (1.22)	4 (2.83)	***0 (8.82)***	***0 (9.83)***	***0 (0.17)***	***0 (0)***
50-60	0 (0.03)	4 (0.10)	0 (0.21)	***0 (0.49)***	***0 (1.42)***	***0 (1.71)***	***0 (0.03)***	***0 (0)***
60-70	1 (0.02)	1 (0.05)	***0 (0.11)***	***0 (0.25)***	***0 (0.71)***	***0 (0.86)***	***0 (0.02)***	***0 (0)***

The usual method of analysis of association between attributes can’t be used in this situation because of structural zeros. For example, in table 1, in the right hand lower triangle cells, a person born between 1980-90 couldn’t have age at onset of more than 29 years at the time of data collection. This situation of structural zero should be considered while analyzing the data. The solution lies in fitting a quasi-independent loglinear model by assigning appropriate weight matrix. The earlier weight matrix (Table 2) was modified to assign zero weights to the cells with structural zeroes and weight of 1 to all non-structural zero cells. The quasi-independent model was fitted using loglinear model with weight option under ‘gnm’ package in R. Table 4 gives the new weight allocation for the quasi-independent model.

Table: 4 Weight allocation - zero for structural zeros and one for sampling zeroes and other observed values

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-‘00	2000-‘10
0-10	1	1	1	1	1	1	1	1
10-20	1	1	1	1	1	1	1	***0***
20-30	1	1	1	1	1	1	***0***	***0***
30-40	1	1	1	1	1	***0***	***0***	***0***
40-50	1	1	1	1	***0***	***0***	***0***	***0***
50-60	1	1	1	***0***	***0***	***0***	***0***	***0***
60-70	1	1	***0***	***0***	***0***	***0***	***0***	***0***

The output of the quasi-independence model which was restricted to nonzero cell frequencies only is given in Table 5.

Table: 5 Observed and Expected values under QI model

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950- 60	1960- 70	1970- 80	1980-90	1990-‘00	2001-‘10
0-10	0 (0.13)	0 (0.47)	1 (1.09)	1 (2.76)	8 (10.14)	19 (15.03)	1 (0.39)	0 (0)
10-20	0 (1.25)	0 (4.38)	4 (10.22)	17 (25.91)	73 (95.31)	185 (141.32)	3 (3.62)	***0 (0)***
20-30	0 (0.68)	2 (2.34)	4 (5.54)	17 (14.05)	99 (51.70)	29 (76.65)	***0 (0)***	***0 (0)***
30-40	2 (0.48)	4 (1.70)	5 (3.95)	28 (10.02)	14 (36.85)	***0 (0)***	***0 (0)***	***0 (0)***
40-50	1 (0.69)	3 (2.42)	15 (5.63)	4 (14.27)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***
50-60	0 (0.32)	4 (1.11)	0 (2.58)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***
60-70	1 (0.44)	1 (1.56)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***

The results of the analysis indicated strong association between the cohort and the age at onset (χ^2 = 195.72, p<0.0001). On a closer examination, it was found that the cells closer to structural zeroes contributed more to the value of overall chi-square. For example, subjects of the birth cohort 1980-90 do not have equal probability to fall in the age at onset group 10-20 or 20-30. The people born in the first half of 1980-90 contribute more to the age at onset group 10-20 and those born in the second half of 1980-90 contribute more to the age at onset group 20-30. Hence the weights assigned to these two cells should not be exactly 1 but only a fraction of it. This suggests that, if the possibility of detecting the case with a given age of onset during the period of study is accommodated in the weight matrix and a QI model is fitted with these weights, it should give a better assessment of the relationship between the variables under consideration.

Now for a birth cohort of *k* years starting at year b_0 and an age at onset interval of length l starting at a_0, the weight contributed is given by the formula

$$W(b_0,k,a_0,l,y)=\frac{1}{k.l}\sum_{b=b_0}^{b_0+k-1}\sum_{a=a_0}^{a_0+l-1}w(b,a,y) \quad (2)$$

where w(b,a,y) on the basis of smoothed truncation is

$$w(b,a,y)=\begin{cases}1.0 & b+a<y\\ 0.5 & b+a=y\\ 0.0 & b+a>y\end{cases} \qquad (3)$$

('b' is the year of birth; 'a' the age at onset and 'y' the year in which the observation ends) (Thennarasu, 2002).

The weight matrix arrived at by using the smoothening technique described above and the expected values given by the corresponding model are given in the following tables (Table 6 and 7).

Table: 6 Smoothened weights

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-'00	2000-'10
0-10	1	1	1	1	1	1	0.875	0.125
10-20	1	1	1	1	1	0.875	0.125	***0***
20-30	1	1	1	1	0.875	0.125	***0***	***0***
30-40	1	1	1	0.875	0.125	***0***	***0***	***0***
40-50	1	1	0.875	0.125	***0***	***0***	***0***	***0***
50-60	1	0.875	0.125	***0***	***0***	***0***	***0***	***0***
60-70	0.875	0.125	***0***	***0***	***0***	***0***	***0***	***0***

Table: 7 Observed and Expected values under QI model with smoothened weights

Age at Onset	Birth Cohort							
	1930-40	1940-50	1950-60	1960-70	1970-80	1980-90	1990-'00	2001-'10
0-10	0 (0.03)	0 (0.13)	1 (0.37)	1 (1.41)	8 (7.32)	19 (19.17)	1 (1.58)	***0 (0)***
10-20	0 (0.3)	0 (1.43)	4 (3.96)	17 (15.16)	73 (78.58)	185 (180.15)	3 (2.42)	***0 (0)***
20-30	0 (0.39)	2 (1.88)	4 (5.18)	17 (19.85)	99 (90.01)	29 (33.69)	***0 (0)***	***0 (0)***
30-40	2 (0.55)	4 (2.64)	5 (7.29)	28 (24.43)	14 (18.09)	***0 (0)***	***0 (0)***	***0 (0)***
40-50	1 (0.97)	3 (4.65)	15 (11.23)	4 (6.15)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***
50-60	0 (0.58)	4 (2.45)	0 (0.97)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***
60-70	1 (1.19)	1 (0.81)	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***	***0 (0)***

The analysis gave $\chi^2 = 29.60983$, p= 0.925 indicating the independence of the variables under consideration.

The results of the analysis are summarised in the Table 8 given below.

Table: 8 Summary table

Method	χ^2 value	P-value	Inference
Chi-square test of independence	381.57	< 0.0001	Age at Onset and Birth Cohort are associated
Quasi-independent loglinear model	195.72	<0.0001	Age at Onset and Birth Cohort are associated
Quasi-independent loglinear model with smoothed truncation	29.61	0.93	Age at Onset and Birth Cohort are not associated

Observations and results

The first two models suggested significant association between the age at onset and the birth cohort. But the results were not reliable due to the presence of structural zeroes and assumption of discrete weights. The smoothened weight model gave the correct conclusion as it took into account the structural zeroes as well as the probability of detection of the case given the age at onset. The independence of the variables is established and further advanced analysis can be taken up with the data set.

Thennarasu (2002) had commented that in the case of psychiatric illness, majority of the studies have the age at onset and birth cohort as two variables associated with the studies. It is essential to verify the results by reanalyzing the data on the basis of the method explained above before drawing final conclusions. The present study on OCD patients corroborates this view. However, the results with respect to earlier studies reported by different authors couldn't be validated in the absence of base data.

Bibliography

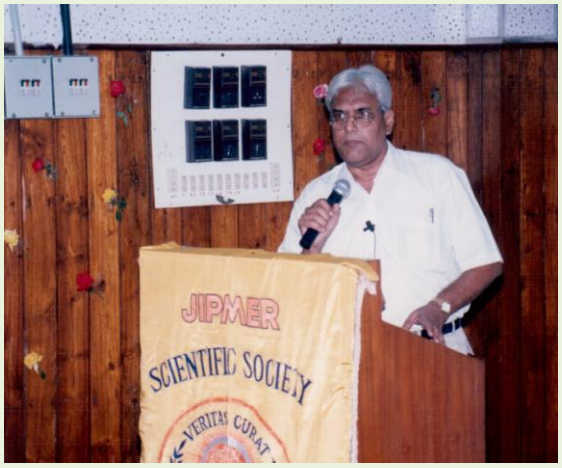
1. Clogg, C. C., and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, London: Sage publications.
2. Everitt, B. S. (1977). *The Analysis of Contingency Tables*. London: Chapman & Hall.
3. R Development Core Team (2011). R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, ISBN 3-900051-07-0, URL http://www.R-project.org.
4. Thennarasu, K. (2002). Loglinear models for shift in age at onset in different birth cohorts in psychiatric patients, *Biostatistics of Health*, 244-249.
5. Turner, H. and Firth, D. (2011). Generalized nonlinear models in R: An overview of the gnm package, R package version 1.0-1, http://CRAN.R-project.org/package=gnm.
6. Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*, fourth edition, New York, Springer, ISBN 0-387-95457-0.
7. Vermunt, J. K. (2005). Loglinear models, *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Ltd

Biostatistics Departments in Medical Institutions in India – Part III (*Continued…*)

Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER) Pondicherry

The 15th Institute Advisory Committee meeting held in March 1975 had recommended pooling all the Statisticians in the institute and bringing them under an independent Department with Professor of Biometrics as the Head. Subsequently, in the year 1987, a faculty in the discipline had joined as Associate Professor of Biometrics and was temporarily placed in the Department of Preventive & Social Medicine till an independent Division of Biometrics became functional in the year 2001.

Since March 1987 the Biometrics-Division through the Department of Preventive and Social Medicine had been actively teaching Biostatistics and Demography to all Undergraduates; Biostatistics and Research Methodology to Medical & Paramedical Postgraduates; apart from regular consultations to Postgraduates, Post-Doctoral programmes, PhD scholars, Faculty and Staff to help them designing research protocols-projects, analyzing data for interpreting results and writing communications.



The Standing Academic Committee (SAC) of the institute in its first meeting held in July, 2009, had agreed in principle and subsequently, in the third meeting held in November 2010, approved the initiation of PhD programme and Post-Graduate courses in Medical Biometrics (Biostatistics) at the earliest after the creation of essential faculty - Assistant Professors and Technical staff positions, apart from establishing the Department with adequate infrastructure and facilities. During the past years the load of teaching and research consultations has increased immensely. Therefore, since January 2011, an independent full-fledged Department of Medical Biometrics and Informatics has become operational along with initiation of the process of upgrading the existing infrastructure and facilities.

During the recent past the Professor of Biometrics, had been supported by the services of Lecturer in Statistics & Demography under Post-Partum Programme (PPP) and the institute administration had recently made arrangements to recruit two Assistant Professors and adequate Technical staff on contractual basis along with essential administrative assistance as had already been recommended by the Standing Academic Committee (SAC) of the Institute.

Educational Courses: The department is currently involved in teaching Biostatistics to the following regular courses;

Undergraduate Courses:

a. MBBS (III and VII Semesters);
b. B.Sc. MLT (3rd Year);
c. B.Sc. Nursing (4th Year)

Postgraduate Courses:

a. MD and MS (1st and 3rd Year) – Two batches each – Jan and July
b. M.Sc. Nursing (1st Year)
c. M.Sc. Medical Biochemistry (2nd Year)
d. M.Sc. Medical Pathology (2nd Year)
e. M.Sc. Medical Physiology (2nd Year)
f. M.Sc. Medical Physics (2nd Year)
g. Post Basic Diploma in Nursing.

Higher Specialty Courses:

a. DM and M.Ch – (1st Year) – Two batches each – Jan and July.

Research Courses:

a. PhD (1st Year)

Certificate Courses:

a. Medical Records Officers
b. Medical Records Technician

Continuing Medical Education Programs

The Department is actively involved as Core-faculty through the Department of Medical Education in conduct of regular CMEs (Orientation Programmes & Workshops) on Research Methodology and Biostatistics organized frequently for MD & MS; DM & MCh and PhD scholars and had coordinated a series of young faculty orientation programmes in research methodology and biostatistics applied to epidemiological studies; laboratory experiments & animal studies; and clinical research studies through the office of Dean Research.

In continuity with several seminars organized during the past, a full-day Seminar was conducted coinciding with the availability of the visiting Professor to the Department; Dr. Abhaya Indrayan, Ex. Senior Professor and Head, Department of Biostatistics and Medical Informatics, University College of Medical Sciences (UCMS), Delhi.

Prof. Indrayan has acted as the key resource person apart from the faculty from JIPMER. So far, a total of around 200 Faculty and Researchers have participated in the Seminars undertaken by the Department till date.

The interactive Seminars on "Writing Research Proposals for Jipmer Scientific Advisory Committee (JSAC)" had four Sessions focused on - Research Methodology Basic Concepts for Research Proposals, Questionnaire and Qualitative Research, talk on 'Intramural Funding'; Study Designs and Sample Size followed by detailed discussions; Writing Research Proposals along with Group Task & a Plenary Session and Testing of Hypothesis & Tests of Significance & Philosophy of Statistical tests.

It is proposed to conduct regular training programmes, seminars & workshops on 'Research Methodology and Biostatistics' for medical Postgraduates, PhD Scholars, Junior & Senior Faculty of the Institute and also to hold periodic regional & national level workshops and seminars on topics of current interest.

Research Consultations:

The Department while providing Biostatistical guidance to various research projects (around 500) and postgraduate dissertations (around 2500) through regular consultation-sessions had also taken the role of co-investigator to some 50 research projects and as co-guide to some 100 postgraduate dissertations. On an average since 1987, a minimum of 300 such consultations are being provided regularly every year. Consultation slots for faculty, staff, resident doctors and students had already been

worked out; and subsequent interactions do take place mostly in the afternoons considering the mutual convenience of time.

In addition to those biostatistical consultations at least 300 hours of teaching had been devoted for the subjects of biostatistics and research methodology regularly for various courses each year. Earlier through the department of P&SM attending as well coordinating conduct of journal clubs, seminars, CMEs & research workshops and training under various national health programmes was part of regular responsibility of the division of Biometrics

SPSS Computer Lab Facility:

Since July 2012, a Computer lab-cum seminar room facility with stand-alone SPSS software (version -19) terminals or nodes is open to various categories of researchers of this institution. The Department had already organized four workshops, soon after the inauguration of SPSS Computer Lab-cum-Seminar room by the Director. The Department conducted a three-day's Workshop on Statistical Package SPSS – Hands on Training (Part I), during the period 30th, 31st August and 1st September 2012. Part II of the SPSS-Workshop was held subsequently. A number of PhD Scholars and Faculty have participated.

Considering the requests and demands' of the participants again a SPSS-Workshop (Part III) was held. A separate full-day 'SPSS- hands on training' Workshop for College of Nursing was conducted for M.Sc. Nursing students. The Department through its SPSS Computer lab facility is committed to hold periodically the workshops on various aspects of SPSS usage.

Population based Monitoring, Evaluation and Epidemiological Studies:

The Division of Biostatistics since its inception in 1987 through the Department of P&SM had coordinated a number of epidemiological studies apart from conducting and supporting a variety of monitoring and evaluations; training and continuing education programmes; including the research components of several national health programmes such as; Universal Immunization Programme (UIP); Integrated Child Development Services Programme (ICDS); Pulse Polio Immunization Programme (PPI & IPPI). Some salient evaluations are listed hereunder;

- Annual "Evaluation studies on immunization coverage, antenatal care, KAP levels of mothers and lameness surveys in Pondicherry". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1988, 1989 and 1990, 1991, 1992.
- "ICDS Evaluation studies at Karaikal, Pondicherry". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1989 and 1991.

- "Report on the training of medical and health personal for the Universal Immunization Programme in Pondicherry". Department of Preventive and Social Medicine, JIPMER, Puducherry, 1991.
- "Study on health and social status of adolescents (11–18 years) in Pondicherry". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1991–92.
- A WHO sponsored study - "A case study on utilization of facilities and services provided by community health centers". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1994–98.
- A Ministry of HRD, Dept of WCD –ICDS-CTC sponsored study - "Impact of unified approach of implementation of Integrated Child Development Services Programme in the Union Territory of Pondicherry". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1995–99.
- "Pulse Polio Immunization (PPI) Evaluation in Pondicherry". Department of Preventive and Social Medicine, JIPMER, Pondicherry, 1996, 1997 & 1998.
- Multiple Indicator Survey (MICS–2000): India Summary Report. Department of Women and Child Development (Government of India) &

United Nations children's Fund (India Country Office), December, 2001.

The Department had also been actively involved in various research decision making committees at national level and within the institution. Indian Society for Medical Statistics (ISMS) was led by the senior faculty in Biometrics as its President (2007-08) and the Department is also recognized as the Office of the Editor of the Bulletin of ISMS (2013-17) providing academic inputs to more than 850 life members of the Society. The annual national conference of ISMS was held in January 2005 at JIPMER Pondicherry. A total of 250 national as well international delegates have participated in deliberations. Also, for the first time four pre-conference workshops / seminars were organised.

Local chapters of Indian Public Health Association (IPHA) and Indian Association of Preventive and Social Medicine (IAPSM) were established and coordinated by the Division of Biostatistics as President, since 2001 and continued through till the year 2006; and still busy providing guidance to the office bearers. Also, the first ever efforts in Spiritual health awareness and research among medical fraternity were initiated during the year 2001-02 by the department.

Computerized Information and Administrative inputs:

The evolution of the institute level computerized services for patients care (HIS); Administration & Accounts (AIS); Hindi Cell; Academics & Research (CRAFT); entrance examinations and related activities had continuous involvements and key challenging roles for the department as Coordinator (JICMIS -1998-2000), Consultant and Faculty-in-Charge, throughout the period (1987-2010). Also for a long period of seven years (2004-2010) entire Central Library activities were coordinated by the Department including establishment of computer lab and long term developmental planning and programmes of the Library.

The Department had been frequently invited by the institute administration to participate in policy decision making and preparation of various reports and documents as well to officially take –up the responsibilities to support and help conduct institute administration on certain occasions; including serving as Faculty Administration and OSD to Director.

More than twenty institutional level core-committees were headed by senior faculty of the Department. The Oversight committee (Medical Group) final report and DPRs' preparation to enhance undergraduate and postgraduate seats in centrally governed medical institutions (around 12 Medical; 6 nursing; Dental and AYUSH) was helped and coordinated by Professor of Biometrics during the year 2006. The Department had recently during the year -2013, had helped preparation of final reports related to Housekeeping and Security requirements of the institute.

Documentations & Information Disseminations:

Apart from co-authoring a number of research papers presented in conferences and published (around 200) in journals of repute; a series of documents; information booklets and policy reports were prepared for JIPMER.

Have been serving as Editorial Adviser / Consultant and Reviewer for Journals & Bulletins of various academic societies such as; IJMR, ISMS, IJCM & IJPH etc. (2002-15).

Also, had been on the Editorial Board, as professional advisor, to some of the Indian Journals published through JIPMER such as;

- Indian journal of Urology (of USI)
- Indian Journal of Pharmacology (of IPS)
- Biomedicine – an International Journal of Biomedical Sciences (of IABMS)
- Indian journal of Dermatology, Venereology and Leprology (IJDVL)
- Journal of Pharmacology & Pharmaco-therapeutics – continuing (2002-15)
- International Journal of Clinical and Experimental Physiology (IJCEP)- continuing (2013-15).

Dr. Ajit Sahai,
Professor & Head,
Department of Medical Biometrics and Informatics (Biostatistics),
JIPMER,
Puducherry – 605 006.
India.

E-mail Id: *ajit.sahai@gmail.com*

Website:
http://jipmer.edu.in/departments/pre-para-clinical/medical-biometrics-informatics-biostatistics/general-info-medical-biometrics-informatics-biostatistics/

Workshop & Seminar Reports

Short Course on
Advanced Design and Statistical Methods in Clinical Trials- A Report

December 9-11, 2014

L. Jeyaseelan,
Dept. of Biostatistics, Christian Medical College,
Vellore, India
E-mail ID: *ljey@hotmail.com*

India has become the focus for clinical trials and drug discovery & development because of its huge population; its vast ethnic variability, biodiversity, different disease prevalence patterns, practice of different systems of medicines, and widely varying socioeconomic groups. Academic institutions and government research agencies are the major contributors for conducting clinical trials in India. As on February 2011, nearly one –fifth of the total trials registered in the Clinical Trials Registry- India (CTRI) were from academia. Biostatistics constitutes the backbone of any clinical trial but currently the capacity of academic institutions in India to undertake this task is still in its infancy. Most of such institutions where majority of the trials are conducted are working without any credible capacity for Biostatistics (designs and methods). Consequently, the clinical trials undertaken by these institutions rely on pharmaceutical industries or Contract Research Organizations (CRO) for managing these activities. This has resulted in poor capacity among most academic institutions in both planning and executing clinical trials. Added to this, is the exorbitant cost involved in outsourcing the conduct, monitoring, management, and analyses of such trials to a CRO. As medical science advances rapidly, clinical trial designers and biostatisticians need to keep abreast of current methodological developments. As the literature is so vast and journals are published so frequently, it is difficult to keep up with recent developments. Many of these come from the pharmaceutical industry and not so much from academic institutions. The goal of this short course was to summarize and impart the recent developments in clinical trial methodology to participants who already have some knowledge on the basic principles of clinical trials and of the statistical methods for their analyses. The focus was Adaptive Designs in Phase II and III studies. This workshop provided an opportunity for the faculty at Vellore to work with University of North Carolina faculty to develop course materials for Advanced Level Clinical Trials course.

This short course provided a comprehensive summarization of recent developments regarding methodologies in clinical trial design and analyses. It covered important topics in early-phases of clinical development such as Adaptive Designs for phase II and III designs:

Day 1
- Adaptive methods: adaptive population, adaptive randomization, adaptive interventions, adaptive sample size, adaptive protocol, adaptive follow-up, and adaptive outcome methodologies
- Alternative randomization schemes: Wei's urn method, Efron's biased coin design, minimization and optimized randomization methods

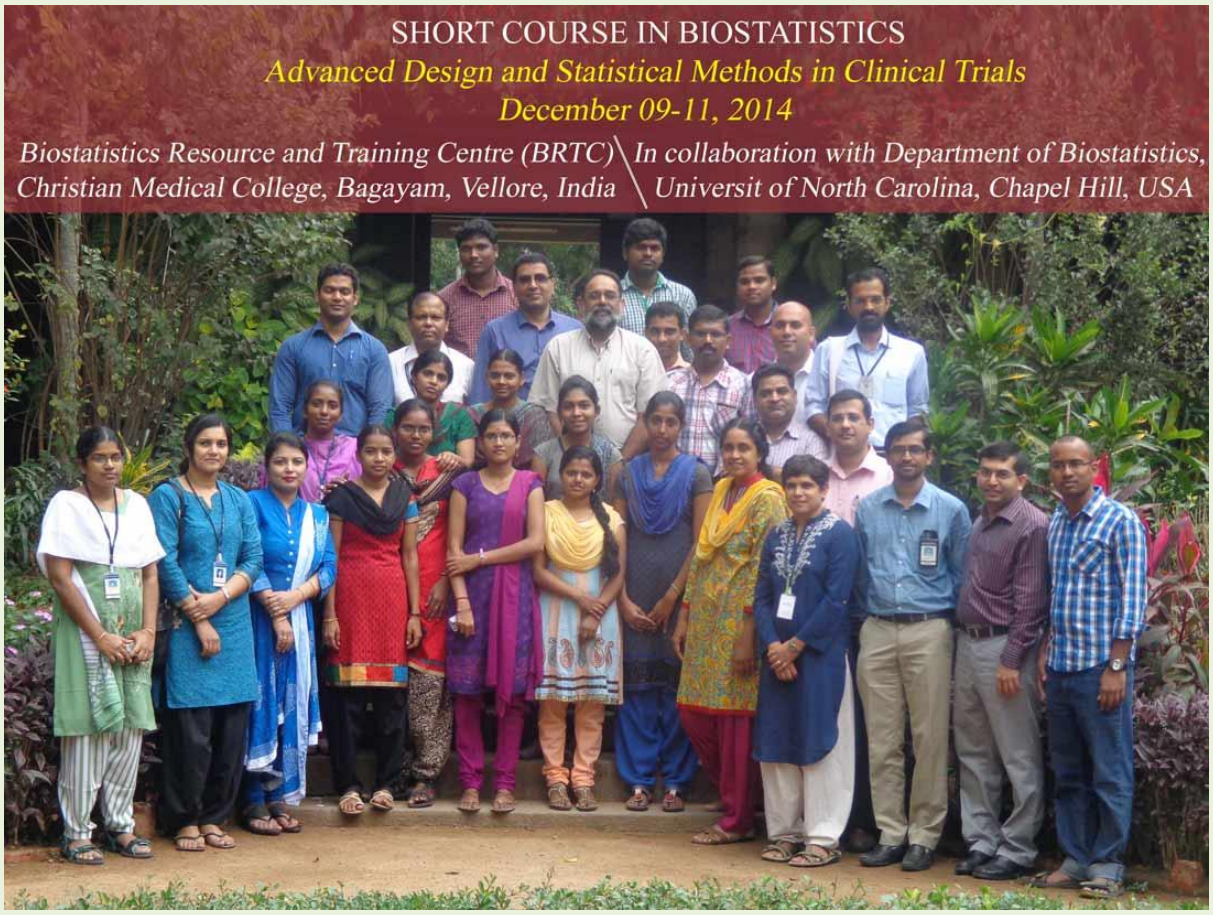
Day 2
- Advanced conduct-related topics: tests to check on randomization, tests to check on adherence/compliance, masking/blinding [including practicals]
- Interim analysis statistical methods: group sequential, conditional power, stochastic curtailment, alpha spending functions

Day 3
- Longitudinal data analysis: Generalized Estimating Equation

(GEE) models, mixed models for repeated measures

- Multilevel regression models beyond the Normality assumption: Gamma and log-normal distributions for continuous outcomes, Poisson and negative binomial distributions for count data
- Handling missing data: imputation methods, multiple imputation

Faculty: The faculties were Dr.Shrikant Bangdiwala, Department of Biostatistics, University of North Carolina, Chapel Hill, USA, Dr.L.Jeyaseelan, Dr.Visalakshi from the Department of Biostatistics, Christian Medical College, Vellore. 33 participants have participated in the workshop.



Biostatistics and Epidemiology Training Courses (BETC) at ICPO, Noida & Outreach Training programmes (National & International) by our division of ICPO in 2015

Smita Asthana
Scientist D
ICPO (Noida)
E-mail ID: *smitasanjay97@yahoo.com*

ICPO is Training courses/workshops are being conducted on regular basis - two to three times a year at ICPO as well at other Institutions on Invitation.

The courses are targeted for Medical faculty, Medical Under graduate and post-graduate students & Ph.D. students from Basic Sciences from various Medical Institutions.

Training Courses (BETC) 2015
National:
1) In house: Ongoing Series of workshop/ Training courses are being conducted at Institute of Cytology and Preventive Oncology (ICPO) Noida .

For UG & PG (Bio-medical) courses Dr L Satyanarayana as Course Director and Dr Smita Asthana as course coordinator and these are the two main faculties. The other faculty from ICPO are Dr Shashi S, Dr R. Suresh Kumar, Ms Uma Kailash and Ms Sarita S. Invited speakers from ICMR head quarters are Dr Ajit Mukherjee in workshop on Biostatistical analysis and Dr NC Jain for Research Paper writing workshop.

1. In-house Training courses:

1. For Under Graduate Medical: **- Project Protocol writing and Study report writing** (Two day workshop),
2. For PG Bio-Medical: **- Biostatistical Analysis and design of medical studies** (Four day workshop)
3. For PG Bio-Medical: (Four day workshop) **- Protocol development & Research Paper writing** (Four day workshop)
4. For PG Bio-Medical/Statistics:– **- Orientation course on Research methodology & Biostatistical analysis** (1 month duration)

Total about 100 students trained from January till August 2015

2. Outreach Training (On invitation) National: **Hands on Workshop on Thesis writing** was organized at SRMS medical college Bareilly in collaboration with SRMS medical college Bareilly and Indian Association of Preventive & Social Medicine (IAPSM) in August 2015.

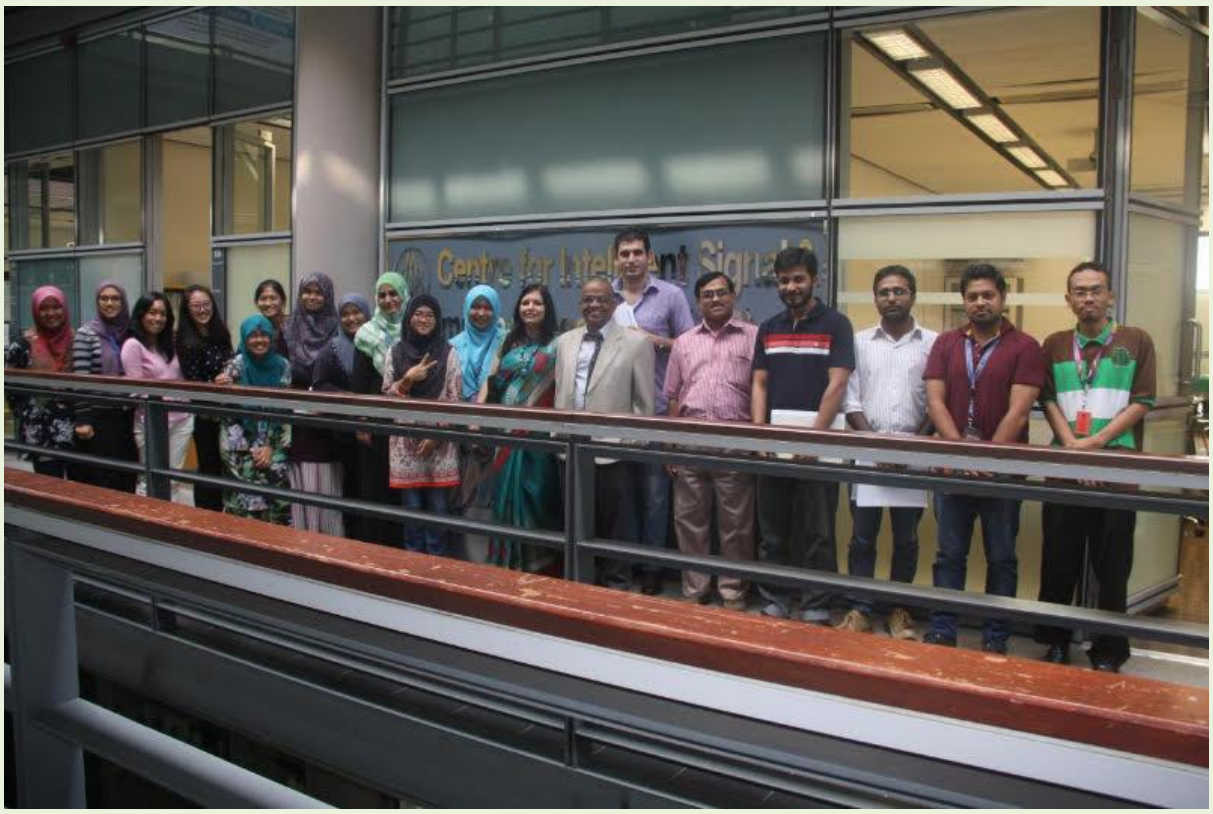
ICPOs scientist Dr L Satyanarayana & Dr Smita Asthana were recognized as resource faculty. Dr B.L. Verma, from MLB Jhansi was also a resourse faculty in this workshop.

3. Outreach Training (On invitation) International:
University technologies, PETRONAS, Malaysia invited as main faculty

Brief Report on Foreign visit for the Meeting/ Workshop. An invitation received from Prof. IR Dr Ahmad Fadzil Mohamad Hani, Director, Centre for intelligent Signal and Imaging Research (CISIR) and deputy Vice Chancellor (Academic) **by Dr Satyanarayana Labani**, Scientist G, ICPO (ICMR), **Dr Smita Asthana** Scientist D, ICPO (ICMR and **Dr Ajit Mukherji** Scientist F, ICPO (ICMR) as Speakers to deliver talks on Biomedical Research-Experimental, Biostatistical Design and Analysis during 2-5 June 2015. The scientists contributed in delivering the talks from some of the following topics as desired by the host institution Various topics of the talks in the workshop were - Steps in protocol preparation- Development of Research protocol, Writing Research Report, ethics in medical research. Biostatistics and Epidemiology Requirements of Research Protocol such as: Types of studies, sampling, data collection methods, approach of statistical evaluation (Descriptive statistics, Significance Tests, concept of p-values). Specific course wise details may be obtained on enrolment. Ethics in medical research, research paper writing and critical evaluation of research paper/ reports. Types of studies (descriptive, analytical & clinical trials. Descriptive statistics, concept of p-values, confidence intervals, principles of statistical tests, tests on compare means & proportions, correlation, concepts of linear / logistic regressions, methods in epidemiology such as assess diagnostic tests, measures of disease frequency, relative risks & odds ratios, etc.,. Hands on training of statistical analysis are given using statistical

software for most bio-statistical methods used in clinical research.

The participants of the workshop are 30 bio-Engineering students who were pursing at the University technology PETRONAS for their Ph D degrees. The scientist received a certificate of Appreciation in the appreciation of his contribution as speaker for 4 days workshop on biomedical research – experimental, bio-statistical design and analysis held during 2-5 June 2015.



Executive Council

President (2015-16)

Dr. A Indrayan,
Noida.
a.indrayan@gmail.com

President Elect (2015-16)

Dr. C M Pandey,
SGPGIMS, Lucknow
cmpandey@sgpgi.ac.in

General Secretary (2014-16)

Dr. Anil C Mathew,
PSGIMSR, Coimbatore.
dranilmathew@rediffmail.com

Treasurer (2014-16)

Dr. A K Bansal,
UCMS, Delhi.
drakbansal2011@gmail.com

Editor (2013-17)

Dr. Ajit Sahai,
JIPMER, Puducherry.
ajit.sahai@gmail.com

Member (2015-16)

Dr. D K Subbakrishna,
NIMHANS, Bangalore.
drsubbakrishna50@yahoo.com

Member (2014-16)

Dr. R J Yadav,
NIMS, New Delhi.
rjyadav@hotmail.com

Member (2014-16)

Mr. Sharad Mathur,
NIMS, New Delhi.
sharadkmathur@yahoo.co.in

Member (2013-15)

Dr. R M Pandey,
AIIMS, New Delhi.
rmpandey@yahoo.com

Member (2013-15)

Dr. Dilip C Nath,
Gauhati University,
dcnath@rediffmail.com

Member (2013-15)

Dr. N K Tyagi,
KLE University, Belgaum.
nareshktyagi@gmail.com

Member (2014-16)

Dr. L Satyanarayana,
ICPO, Noida.
satyanarayanalabani@yahoo.com

Awards Committee (2014-16)

Chairman

Dr. K R Sundaram,
AIMS, Ernakulam
krsundaram@aims.amrita.edu

Member

Dr. R M Pandey,
AIIMS, New Delhi
rmpandey@yahoo.com

Member

Dr. B Antonisamy,
Web Coordinator,
CMC, Vellore
b.antonisamy@gmail.com

Chairman, Smt. Ramrati Lalima Sahai Award Committee

Member

Dr. C M Pandey,
SGPGIMS, Lucknow
cmpandey@sgpgi.ac.in

Member

Dr. L Satyanarayana,
ICPO, Noida
satyanarayanalabani@yahoo.com

Dr. T Krishnan,
Bangalore.
krishnant001@gmail.com

Nomination Committee (2014-16)

Chairman

Dr. Arvind Pandey,
NIMS, New Delhi
arvindp55@hotmail.com

Member

Dr. B L Verma,
MLBMC, Jhansi
blverma49@gmail.com

Member

Dr. Ajit Sahai,
JIPMER, Puducherry
ajit.sahai@gmail.com

C.R. Rao - Prestigious Awards

***Shanti Swarup Bhatnagar Award**, Council of Scientific and Industrial Research, India, for "notable and outstanding research in statistics" from Pandit Jawaharlal Nehru - 1963*

***Samuel S. Wilks Medal**, awarded by the American Statistical Association for the great influence he has had on the application of statistical thinking in different disciplines, embodying over a career of more than 40 years in the spirit and ideals of Samuel S. Wilks -1989*

***Padma Vibhushan**, second highest civilian award, from the Government of India for "outstanding contributions to Science and Engineering/Statistics"- 2001*

***National Science Medal, USA** honored as "prophet of a better age" from President Bush - 2002*

***Mahalanobis Prize**, awarded by International Statistical Institute at the 54th Session held in Berlin for "lifetime achievement" - 2003*

***India Science Award**, "for major contributions of a path-breaking nature based on work done in India" from Prime Minister Manmohan Singh - 2009*

***Guy Medal in Gold** of the Royal Statistical Society, UK "for those who are judged to have merited a significant mark of distinction by reason of their innovative contribution to theory or application of statistics" from the president of the Royal Statistical Society - 2011*

Honored as a "statistician of international repute" by the President of India, Pranab Mukherjee, at the World Telugu Conference in Tirupathi, India - 2012

.

***Jerzy Splawa-Neyman Medal**, from the Polish Statistical Association in recognition of "his outstanding contributions to the theory, applications, and teaching of statistics" - 2014*

Statistics in India

Development of Statistics in India: C.R. Rao worked at the Indian Statistical Institute for a period of 40 years in various capacities as Professor, Director, Jawaharlal Nehru Professor and National Professor. During most of this period there was no separate minister for statistics. All problems relating to the development of statistics in India were under direct administrative control of the Prime Minister Pandit Nehru as he was greatly interested in the development of statistics in India. He visited the ISI a number of times at the invitation of Professor Mahalanobis, the founder of ISI. Rao had the opportunity to discuss the national statistical system and training of statisticians to work in statistical bureaus with Pandit Nehru. His successor Prime Minister Smt. Indira Gandhi was also interested in the development of statistics. Rao was made a member of Committee on Science and Technology (COST) in 1962 and chairman of National Committee on Statistics in 1969. He had several opportunities to discuss the development of research activities in India with Smt. Indira Gandhi and Prime Minister Morarji Desai**.**

Rao and Bhargavi Rao with Professor Mahalanobis (founder of ISI) and Rani Mahalanobis at the convocation of ISI in 1967 where the election of Rao to Royal Society was announced. Rao worked with the Professor for a period of 30 years.

Rao with Prime Minister Nehru at a meeting to discuss promotion of science research in India .

Rao met Prime Minister Smt. Indira Gandhi after he received FRS (Fellowship of Royal Society) in 1967 to discuss the future of science research in India as she was very concerned about lack of good scientists In India. The photo was taken in the residence of the Prime Minister.

Rao with Prime Minister Morarji Desai at the conference of the International Statistical Institute held at New Delhi in 1977. During the discussion on promoting research in statistics in developing countries, Dr. Desai referred to lack of good scientists in India and made valuable suggestions on what universities and research institutions could do to promote science research in India. The discussion was continued during his visit to the Indian Statistical Institute to receive an Honorary Doctorate degree.

A Place in the History of Statistics

Rao's contributions to the development of statistics as an independent discipline earned for him a place in the history of statistics as one of the founders of statistics. Rao is the only Asian mentioned in all reports (websites on history of statistics) on main contributors to the development of statistics.

- Rao's major contributions to statistics leading to several technical terms incorporated into textbooks on statistics were made in the 1940's while working at the Indian Statistical Institute.
- Figures from the history of probability and statistics by Professor John Aldrich, University of Southampton, UK describing the work of 35 major contributors to probability and statistics since 1650 includes Rao.
- Statisticians in History by American Statistical Association has Rao's name mentioned in a list of 50 contributors.
- Chronology of Probabilists and Statisticians: A list of 57 major contributors to probability and statistics from 16th century to 20th century prepared by Statistics Department, University of Texas at ElPasso, USA, has Rao's name.
- Rao is one of the 77 contemporary scientists from all over the world in all areas of science, selected by Gerard Piel, editor of Scientific American, covered in the book "Faces of Science" by Mariana Cook. The book contains portraits of scientists taken by the author with a short description of contributions made by each. The portraits were exhibited at

James D. Watson (Nobel Laureate) and Rao in front of portrait of Rao exhibited at the "Faces of Science" display at the Gallery of Arts and Sciences, New York Academy of Sciences - 2005

CR Rao Advanced Institute of Mathematics, Statistics, and Computer Science, in University of Hyderabad Campus, Prof. C. R. Rao Road, established to promote basic research in science and technology - 2011

The C.R. Rao gallery capturing Rao's life in statistics over 65 years (40 years of work in ISI , India, and 25 years in USA at The Pennsylvania State University and University of Pittsburgh) was inaugurated by Nobel Laureate V. Ramakrishnan on December 22, 2013 at the C.R. Rao Advanced Institute of Mathematics, Statistics, and Computer Science, Hyderabad. The Gallery is open to the public.