Analysis of hospital based ayurvedic clinical practice to gain real world data knowledge

Vinay Mahajan

PhD thesis

Guides: Dr.Ashwini Godbole,

Dr.Girish Tillu,

Dr.Ashwini Mathur

Contents

[Introduction 4](#_Toc42079224)

[Origin of word doctor 4](#_Toc42079225)

[Origins of pharmaceutical industry 5](#_Toc42079226)

[Elaboration of clinical trials: Origin of RCT blinded trials 5](#_Toc42079227)

[Role of statistics, analyst, programmer 7](#_Toc42079228)

[Modern hospitals, every day clinical practice and healthcare environment 8](#_Toc42079229)

[Practice of Ayurveda, reverse pharmacology, experts’ view 9](#_Toc42079230)

[Dr. Ashok Vaidya 9](#_Toc42079231)

[Prof R H Singh 10](#_Toc42079232)

[Dr. Baghel 10](#_Toc42079233)

[Dr. Ram Manohar 10](#_Toc42079234)

[Prof Bhushan Patwardhan 10](#_Toc42079235)

[Prof. Darshan Shankar 11](#_Toc42079236)

[Observational studies, Real world evidence, level of evidence 11](#_Toc42079237)

[Efforts at national level 12](#_Toc42079238)

[AYUSH 12](#_Toc42079239)

[Digital India movement 13](#_Toc42079240)

[National AYUSH Morbidities and Standardized Terminologies E- Portal (NAMASTE Portal) 14](#_Toc42079241)

[AYUSH Research Portal 14](#_Toc42079242)

[AYUSH Hospital Management System 14](#_Toc42079243)

[Electronic Medical Record, medical IT infrastructure 14](#_Toc42079244)

[Need of this study and methodological framework 15](#_Toc42079245)

[The proposed study in brief 18](#_Toc42079246)

[Chapter outline 19](#_Toc42079247)

[Summary of the chapter: 24](#_Toc42079248)

[Converting real life clinical data into analyzable format 25](#_Toc42079249)

[Background 25](#_Toc42079250)

[Material and Methods: 26](#_Toc42079251)

[Real actions adopted on acquiring data 31](#_Toc42079252)

[Data access 31](#_Toc42079253)

[Part 1 31](#_Toc42079254)

[Details of the database 31](#_Toc42079255)

[Understanding Source Databases 34](#_Toc42079256)

[Part 2 36](#_Toc42079257)

[Data extraction for downstream activities 36](#_Toc42079258)

[Data extracted from Hospital database 36](#_Toc42079259)

[Part 3 37](#_Toc42079260)

[Data preparation 37](#_Toc42079261)

[1. Merging, joining: 37](#_Toc42079262)

[2. Appending 37](#_Toc42079263)

[3. Filtering 37](#_Toc42079264)

[4. Deduping 37](#_Toc42079265)

[5. Cleansing 37](#_Toc42079266)

[6. Transforming 38](#_Toc42079267)

[7. Aggregating 38](#_Toc42079268)

[8. Format revision 38](#_Toc42079269)

[9. Key Restructuring 38](#_Toc42079270)

[10. Bucketing/Binning 38](#_Toc42079271)

[11. Z-Score Normalization and Max-Min Scaling 38](#_Toc42079272)

[12. Imputation 38](#_Toc42079273)

[13. Numerical Imputation 39](#_Toc42079274)

[14. Handling Outliers 39](#_Toc42079275)

[Part 4 39](#_Toc42079276)

[What is this data used for? 39](#_Toc42079277)

[1. Who is doing the analysis? 39](#_Toc42079278)

[2. What type of questions need to be asked or answered? 39](#_Toc42079279)

[Part 5 39](#_Toc42079280)

[How to do it? 39](#_Toc42079281)

[1. Prepare the way you think 40](#_Toc42079282)

[2. Compartmentalize each step 40](#_Toc42079283)

[3. Modular approach to make maintenance easy 40](#_Toc42079284)

[Exploration of the database 40](#_Toc42079285)

[Flowcharts of the algorithms 41](#_Toc42079286)

[Details of the reference dataset “01adsl\_met\_rmsd” 47](#_Toc42079287)

[Other visit wise data 54](#_Toc42079288)

[Results and discussions: 56](#_Toc42079289)

[Summary of the chapter: 57](#_Toc42079290)

[Clinical data understanding 58](#_Toc42079291)

[Challenges 61](#_Toc42079292)

## Introduction

## Origin of word doctor

The word doctor is a Latin word and later a French one, meaning anyone who is a teacher - usually of law, theology, philosophy, as well as medicine for a learned profession. Physician is a more accurate way to call a medical doctor, which comes from the Latin word physicum and later the French word physic, coined in 1212.Around 1400 AD, the third edition of the Shorter Oxford English Dictionary was the first to use the word “physician”. It isremarkable to note that, as per this edition, physicians and surgeons were members of entirelydistinct professions, and were thought of as rivals.  
  
The work done by Szasz and Hollender (1956)defined three basic models of the doctor-patient relationship. These are (a) active-passivity, (b) guidance-co-operation and (c) mutual participation. The activity-passivity and guidance-co-operation models are completely paternalistic and largely doctor centered. The final, mutual participation has a bigger importance on patient centered treatment.  
  
Early age doctor patient relation

In older Egyptian medicine, the activity-passivity type relationship existed and remained unaltered throughout.The Greeks depended more on real-life observation, boosted by practical trial and error experience, discarding magical and religious explanations of human bodily dysfunction.In the medieval Europe, the doctor, packed with magical powers, was in a high-ranking position in society and his patients were regarded as powerlesssubordinates, similar to the activity-passivity model.  
  
During the early 18th Century, doctors were few in number and their patients mainly upper class and noble. This status inequality confirmed the dominance of the patient and doctors had to compete with each other in order to satisfy the patient. During the late 18th Century, hospitals came up as places to treat patients who were deprived. Doctors found themselves providing medical treatment for those who were traditionally regarded as more passive, i.e. an activity-passivity (paternalistic) model.This hospital model involved the checkup of the patient's body, the expert clinical and anatomical knowledge possessed by the doctor to make a diagnosis, making the patient dependent as a result. The association was between a dominant doctor and a passive patient, i.e. an activity-passivity (paternalistic) model.  
  
Balint (1964) maintained that “the most potent therapeutic tool the doctor possessed was himself or herself”. However, Balintrecognized that very little was known about the ‘pharmacological’ attributes of this drug, such as the right ‘dosages’ (frequency of visits), any addictive properties (whereby the patient becomes gradually reliant on the doctor), and side effects (i.e. what harm the doctor could do), the doctor-patient relationship was a joint investment which over time would benefit both parties.  
  
The doctor-patient relationship in the two oldest civilizations, those of India and China, has remained far more constant than in Western societies. A patriarchalstyle still dominates, and doctors have a high status in society. In the ancient India, doctors were enjoying the highest level of respect in the society because of the outlook and interest towards the patient. They were deemedas “Demi Gods”. In that era medicine tradition was seen from social and philanthropic standpoint rather than taking it as a business standpoint. Through the 18th century, general physicians mostly practiced medicine in their own village or nearby area. They established a respectable relationship with the community, as they were always available. They commonly discussed with the patients regarding diverse topics of family rather than usual doctor- patient interaction. They knew the health status of all family members of the patient. As a result, a good interpersonal connection was developed between doctor and patient. Patient had utmostfaith that doctor could do any harm. At times, any undesirable condition aroused then patient thought that it was only God’s will.

The vaidyas and the hakims from the ayurvedic and Unani systems of medicine complemented and used skills and concepts from each other as well as from the Arabs and Chinese. There is almost no report of enmity between them. However, the relations between the doctors of indigenous and Western medicine was anything but honest.There was a clash of cultures, the East was perceived as weak vs. the powerful knowledge of the West. In the East, medicine was mostly pluralistic. There was recognition and acceptance of alternative traditions. Medicine was not regarded only as a biological phenomenon and importance was given to a patient's societal standing, environment, and relation with the therapist. As Imperial rule intensified, declarations of the Western scientific superiority increased. Allopathic practitioners saw themselves as modernizers and often treated their indigenous peers with disrespect for their “inferior knowledge”. Local knowledge was labeled unscientific or irrational. While Western medicine was given the status of official medicine, the state turned biased and hostile toward the other systems [17]. The rising nationalism also posed to be an obstacle in a beneficial give and take of ideas [18].

Origins of pharmaceutical industry  
The roots of the pharmaceutical industry lie back with the apothecaries and pharmacies that offered traditional remedies as far back as the middle ages, offering a hit-and-miss range of treatments based on centuries of folk knowledge.The modern medicines regulation began only after revolutionary development in the 19th century life sciences, in chemistry, physiology and pharmacology, which put a robust foundation for the modern drug research and development.  
  
Unfortunate events like deaths due to diethylene glycol poisoning in the US in 1930s, the thalidomide disaster in late 1950s, have catalyzed the development of medicines regulation more than the evolution of a knowledge base throughout the world. The formation ICH consortium is one such a byproduct. The use of statistics to support R&D of new medicines has grown multifold since the Kefauver-Harris Amendments (1962). These clearly stated that the Food and Drug Administration (FDA) would require “substantial evidence” of the impact of a drug in a clinical trial setting and that new drug approvals would not be based only on proof of safety. In the USA and all over world, since 1970s, the value of medicine has been clearly exhibited by a longer life expectancy, a lower infant mortality rate, and the higher quality of life many of our senior citizens have been enjoying.

## Elaboration of clinical trials: Origin of RCT blinded trials

Randomized Controlled Trial (RCT) is a classicalresearch design, in which the participants are randomly allocated to one or other treatment conditions under thestudy. RCTs help in reliabledescription of causality.Randomization is the fundamental characteristic of an RCT andit describes the random distribution of subjects to thestudy arms. RCTs provide the researcher the promise that the difference in the outcomes among subjects instudy arms was exclusively caused by the intervention, asrandomization equalizes the study group in all otherfactors. Thus, RCTs set the standard of excellence inhealth sciences research.

At the beginning of the study, randomization reduces the biasin assigning subjects to the intervention and control group. However, it does not rule out the possibilities of differentialtreatment of groups or biased adjudication of outcome variables. Blinding assists in monitoring several types of biasesthat might unintentionally creep into the study. The two major biases, namely performance bias and the ascertainment bias that can be controlled using blinding. The four groups of people blinded in the trial are the study subjects, theinvestigator(s), the outcome assessor(s), and the data analyst(s). Based on the number of people blinded, trialsare classified as open label trial, single blinded trial, double blinded trial, triple blinded trial, and quadruple blindedtrial.

<http://ejournal.manipal.edu/mjnhs/docs/Volume%203_Issue%201/Full%20text/9.%20Vishnu.pdf>

Physicians and clinical researchers have remained confident that RCTsdeliver the most rigorous test of preventive, diagnostic, and therapeutic interventions. They are universally denoted as the “gold standard” of experimental medical analysis, as aundisputable starting point in diagnostic or therapeutic evaluation.When did RCTs become the “gold standard”? The first occurrence the authors of the referred paper have found of the phrase “gold standard” to refer to RCTs came in the pages of The New England Journal of Medicine (NEJM) in December 1982, in an article written by Alvan Feinstein and Ralph Horwitz. This date surprised the authors as a very late date for the first usage.Despite extensive searching, they have found no earlier occurrence of “gold standard” in reference to RCTs. They quote “We are eager to be proven wrong, but until all textbooks, conference proceedings, journals, and archival collections have been digitized and made full-text searchable, the gold standard of historical research itself remains elusive”.

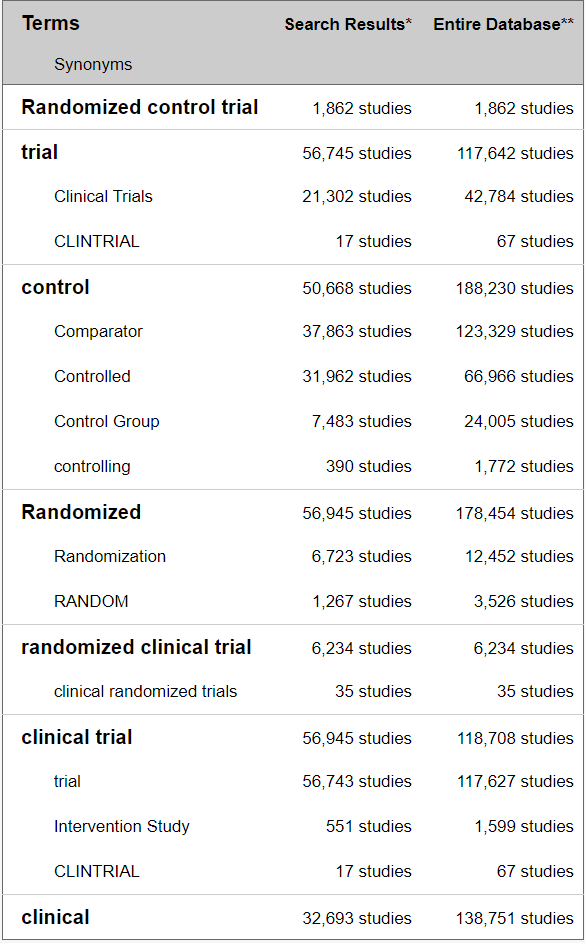
<https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)60742-5/fulltext>

While probing on the details regarding RCTs on the ClinTrial.gov website the following data was discovered on 24th April 2020:

* Total trials in the ClinTrial.gov database = 336905
* The estimated % of RCTs based on the results range between 2.40% (only considering “Randomized control trial” and “randomized clinical trial” 1862 + 6234) to 19.31% (considering “Randomized control trial”, “Randomized” and “randomized clinical trial” 1862 + 56945 + 6234), remaining 80% to 97% of the registered trials are not RCTs.

Screenshot from the Search Details: “Randomized control trial OR randomized clinical trial”.

<https://clinicaltrials.gov/ct2/results/details?cond=&term=Randomized+control+trial+OR+randomized+clinical+trial&cntry=&state=&city=&dist=&Search=Search>



As of April 2020, Clintrial.gov database had 19% RCTs of all the studies registered in the database. This provides us with enough evidence that a lot of research workis carriedoutside of the RCT framework.

## Role of statistics, analyst, programmer

The practice of biostatistics, or statistics in biomedical research, has developed by quantitative methods from other areas such as mathematics, operations research, and economics. Statistics is regularly applied to genomics, population genetics, epidemiology, ecology, nutrition, etc. The 1990s presented explosive increase in applications of computationally demanding methods originated on statistical principles, due to the focus on translational research, personalized medicine, and the relevance of real-world data. Statistical modeling has become more complex due to the size of data, computational requirements and varied data sources. These developments presented additional roles like statistical programmer to statistical analyst to clinical programmer to clinical analyst.  
  
A clinical data analyst (or clinical informatics analyst) is a healthcare professional responsible for confirming the validity of scientific experiments and data gathered. They apply their knowledge of data acquisition, management, analysis, and interpretation to healthcare data, providing actionable insights that physicians, clinical researchers, decision-makers, and others can use. They may be responsible for automating internal and external reports, creating executive-level dashboards, and presenting information to help hospital executives and others understand the operational impact of the data. These data analysts ensure that processes and protocols are followed, thereby improving the quality and efficiency of care.

They help organizations improve the quality of care, lower the cost of care, and enhance the patient experience. They provide data insights that drive clinical process improvement, such as reducing readmissions and hospital-acquired conditions. In addition, healthcare analysts help insurers, vendors, and others synthesize data that guides decision-making, population health management, cost containment, and quality improvement.

Clinical data analysts report the results of clinical business intelligence to management, stakeholders, and other interested parties. They coordinate with other relevant departments such (e.g., clinical strategy, clinical operations) to determine the areas to be analyzed as well as the appropriate measures that should be taken to ensure data analysis proves useful.  
  
This role has the ability to combine strengths from multiple areas to build a powerful storyline.By providing greater insight to patients, providers, and policy makers into the appropriate application of interventions, and quality and costs of care, these data offer the opportunity to accelerate progress on the six dimensions of quality care—safe, effective, patient centered, timely, efficient, and equitable (Chaudhry, 2006; IOM, 2001, 2009; Safran et al., 2007).

## Modern hospitals, every dayclinical practice and healthcare environment

One trend all modern hospitals have in common is that the amount of processed information generated per patient is constantly increasing EMR, patient charts, CT scans, X-rays, ECGs, pathology reports, wearable devices, sensor data, etc. When such information is properly compiled and aggregated, it provides data for efficient hospital administration. How each patient is treated can yield important statistical data, while the number of patients handled, grouped by sex, age, and diagnosis, provides basic information required by administrators.  
  
Health information provided by the hospital is vital forpublic health administrators at the district level. Such data, accumulated from all the district's hospitals, are essential in formulating health planning for the district, by matching with population statistics, other demographic data and the health resources of the district.This medical information, containing not only data about the patient's illness but also the success or failure of therapy, is a precious source of clinical training for medical students, nurses and related health care work force.This is a basis of new medicine and therapies. The development of new therapeutic drugs is built upon the careful observations of experienced physicians, and the comparison and analyses of the effects of previously administered drugs, gathered from the patient's data. This in turn leads to continued medical progress in, say, uncovering a new entity of a disease or a new method of diagnosis.This information is processedmanually or partially automatically. This inefficient method results in delays in reaching crucial treatment decisions, for instance when test results fail to get back to the physicians in time, or in delayed action by administrations when a prompt response is needed.  
  
Internet or google era  
In the last decade of the 20th century and starting of 21st century, a chronic displeasure began to settle over the relationships between doctors and patients. Due to multiple advances in medical, chemical, bio-chemical, IT industries, patients were doing better but feeling worse, the personal, solid and faithful relationship is failing.

Earlier, patients were most often considered medically illiterate to make decisions on their own behalf. Thus, informing patients about the uncertainties and shortcomings of medical interventions worked only to damage the faith that was so important to the therapeutic success. Doctors felt secure in making decisions on behalf of their patients. Later on, the distance between the doctor and patient widened. Little social mixing continued, and the doctor-patient relationship became impersonal and isolated, based upon negotiation and financial transaction.

Modern-day effect on the doctor-patient association has been the exponential growth in the use of the Internet by patients. This has meant that patients are largely better informed. In an effort to understand more about the cause of their symptoms through Internet searching, patients assume that they are suffering from a serious disease, which is often not the case, say doctors. They have warned that self-diagnosis through Internet might be misleading for patients.

## Practice of Ayurveda, reverse pharmacology, experts’ view

The western medicines are developed using a method called as hierarchical method where it tries answering the questions with limited scope e.g. what is the efficacy of a particular drug, what is the safety profile of a drug? This method assumes a step wise approach and deals with the problem in successively conducted clinical trials of various types in a specific sequence. The pharmacology of the molecule is ascertained first at the very beginning. These studies are followed by cohort studies, Open-label randomized studies. The process ends with the blinded, randomized, placebo controlled trials (RCT). The RCTs offer most internal validity and reduce the bias. These studies could be complemented by then moving onto case studies, case series. This “one step at a time” approach has worked very well in the western medicine framework.

There are some other models proposed by various authors to handle complex and tricky situations arising in defining and understanding the action of mechanism of Ayurvedic intervention.

As described by Dr. Girish Tillu, huge observational data for ayurvedic medicines covers: more than 1,00,000 books and manuscripts, 57 authentic books (Drug and cosmetic act 1940), > 4500 diseases including subtypes and conditions (Ayusoft database), > 81,000 formulations (TKDL database), > 4,00,000 Practitioners (Planning Commission - 11th Plan) in India, infinite documents, references, experiential data, living tradition and knowledge in public domain. Dravyaguna (Pharmacology), BhaisajyaKalpana (Pharmaceutics), Nidana (Diagnosis) and Chikitsa (Management principles).This data points to a validated knowledge base and it is acceptable that Ayurveda is an evidence based knowledge system. Let us take a look at points of views of some of the leading figures in the field:

### Dr. Ashok Vaidya

Dr. D. B. Vaidya has explained the concept of reverse pharmacology to understand the action mechanism of Ayurvedic intervention. Reverse pharmacology is the science of integrating documented clinical/experiential hits, into leads by trans-disciplinary exploratory studies and further developing these into drug candidates by experimental and clinical research. It comprises of three stages - experiential, exploratory and experimental.

1. Experiential robust documentation of clinical observations of the biodynamic effects of standardized ayurvedic drugs by meticulous record keeping.
2. Exploratory studies for tolerability, drug interactions, dose range finding in ambulant patients of defined subsets of the disease and para-clinical studies in relevant in vitro and in vivo models to evaluate the target activity.
3. Experimental studies, basic and clinical, at several levels of biological organization, to identify and validate the reverse pharmacological correlate of ayurvedic drug safety and efficacy.

Based on the huge observational data and the relatively low rate of side effects, it is easy to test Ayurvedic intervention in larger clinical trials. This would help build the required safety and efficacy information relevant to the Ayurvedic intervention under question. Once these key parameters are established, the pharmacokinetic properties can be understood. This process is economical and may take lesser amount of time when compared with the hierarchical model used in western medicine.

### Prof R H Singh

Research experiments within Ayurvedic area undertaken during the last 50 years have not been very rewarding, except for the extremely useful exercise of literary research.These have at least made a few of the classical Ayurvedic texts accessible to contemporary readers and researchers. A number of literary reviews published in recent years have helped create a conceptual interface between Ayurveda and modern science. However real laboratory-based new research is still awaited. Such a scientific deadlockdeserves exploiting newer strategies,while keeping with the fundamental principles of Ayurveda-as-it-is, without distorting it to suit the application of modern research technology. In any research, the goal of research should not be compromised to suit the convenience of research methods. But unfortunately in Ayurvedic research, there has always been a reverse compromise, and in my perception this attitude is the main reason for failures in this otherwise potentially most fruitful field of contemporary medical research. The so-called scientific research of several decades has helped neither Ayurveda nor modern medicine to any significant extent except in creating awareness.

<http://europepmc.org/article/PMC/3151394>

### Dr. Baghel

Research should be a process that converts data into information, information into knowledge and knowledge into wisdom. It should be balanced, comprehensive, and equally emphasizing in the literary field, experimental and clinical research. It should be able to impact the fields of academics, pharmacy and practice in a profound way.The researches done in the last 60 years on Herbal Pharmacology have led confirmation of few concepts like Reverse Pharmacology and use of whole crude drugs in place of isolation of fractions for clinical trials.Various researchers started to feel that conventional clinical trial regimen is not fit for Ayurveda. Ayurveda is a pure science based on strict logical explanation, which is called *Darshana*. Ayurveda was always in the developmental phase like all the medical systems should be.

### Dr. Ram Manohar

He has opined that Ayurveda is based on 5000 years of clinical practice. Hence, in place of conventional evidence-based medicine (EBM) clinical trials, practice-based clinical trials should be organized for Ayurveda.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3215413/>

### Prof BhushanPatwardhan

The holistic concepts of Ayurveda give emphasis to health promotion, disease prevention, early diagnosis and personalized treatment. There seem to be substantial similarities between the traditional systems like Ayurveda and the innovative approach of predictive, preventive and personalized medicine (PPPM). Continuous research on safety, quality and efficacy of Ayurvedic drugs and procedures is needed. Systematic documentation and critical analysis of clinical practice are necessary.It is very important to review available evidence in the right perspective. In case of Ayurveda, the evidence can be drawn from two main sources.

1. Source of evidence may be based on historical, classical and present nature of clinical practice. Here, the documentation of practice to support various claims is very crucial. Mere reference to classical texts is not sufficient as evidence for practice.
2. Source of evidence may be based on scientific research to support various theories, medicines and procedures used in Ayurvedic medicine. A critical situation analysis of present status of clinical practice and scientific research on Ayurvedic medicine may be necessary at this stage.

In another paper, he remarks “We need contemporary practices and models for healthcare that bring confluence of Ayurveda and modern science. Becoming modern is not a crime; it does not prevent us from maintaining our own cultural identity. For instance, Charaka would not have ignored technologies like electron microscope if they had been available during his time”.

<https://link.springer.com/article/10.1186/1878-5085-5-19>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4061590/>

### Prof. Darshan Shankar

As of 2018, at the national level, Ayurveda receives a small outlay of around 2.5% of the Central health Budget and at the State level, it receives a minuscule part of the state health budgets.Despite its strengths, Ayurveda definitely has some limitations in current scenario.Today, Ayurveda needs to interface in an epistemologically informed manner, with molecular biology in order to discover its own mode of actions at the structural level and embrace tools of information technology to organize its enormous multifaceted data, in searchable formats. Rigorously documented clinical experiences interpreted through Ayurveda-biology will deepen modernization of healthcare in India.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6314269/>

## Observational studies, Real world evidence, level of evidence

Real-world data are data captured in an observational manner, in a natural, uncontrolled setting – outside of traditional clinical trials. Ordinary clinical practice, producing a never-ending flow of results from every day practice, can be viewed as infinite sequence of unsystematic observational studies. There must be a rational use of outcome data from patients, far more representative of the general population than those included in formal clinical trials. Practice-based medicine is an important way to advance science. For this to happen, the treating doctor should be a “clinician researcher investigator”. Doctors who, even in their busy practice, are able to see patterns in their practice, then formulate a research hypothesis, and then proceed to test it. Thus, a doctor can be good at both good clinical practice (GCP) and good clinical research practice (GCRP). One can feed the other and vice versa, and this way medicine advances.

Investigation of the literature across timepoints from year 2000 to year 2020 presentsthat well designed observational studies could be producing treatment effects similar to the RCTs:

**Year 2000:** For the five clinical topics and 99 reports evaluated, the average results of the observational studies were remarkably similar to those of the randomized, controlled trials. The results of well-designed observational studies do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic.

John Concato, Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs, N Engl J Med 2000;342:1887-92

**Year 2007:**“When done correctly, epidemiologic studies of drug effects can be both more conceptually demanding and more powerful than the average RCT, especially in assessing drug safety. RCTs offer one kind of knowledge but prevent us from seeing other properties of a drug. Epidemiologic studies can help elucidate those properties but may introduce new blind spots. We must do both kinds of research with rigor and with humility.”

Jerry Avorn, In Defense of Pharmacoepidemiology — Embracing the Yin and Yang of Drug Research, N Engl J Med.2007:357;22

**Year 2014:** Our results across all pooled reviews (pooled ROR: 1.08) are very similar to the results reported by similarly conducted reviews. As such, we have reached similar conclusions, on average, there is little evidence for significant effect estimate differences between observational studies and RCTs regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmacological interventions. Factors other than study design per se need to be considered when exploring reasons for a lack of agreement between results of RCTs and observational studies. Our results underscore that it is important for review authors to consider not only study design, but the level of heterogeneity in meta-analysis of RCTs or observational studies. A better understanding of how these factors influence study effects might yield estimates reflective of true effectiveness.

Plain language summary: healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials (review) 2014, The Cochrane Collaboration.

**Year 2020:**Accumulating evidence suggests that appropriately conducted RWD studies have the potential to support regulatory decisions in the absence of RCT data. Further work may be needed to better illustrate the settings in which RWD analyses can robustly and consistently match the results of RCTs and, more importantly, the settings in which they cannot match them. After careful consideration of the potential for bias, regulators can then determine when they would unequivocally accept RWD in place of an RCT. If studies based on RWD are ever to replace RCTs, regulators may need to accept that the cost of accelerating patient access to treatment carries a higher level of decision-making uncertainty than that with which they are familiar.

Ramagopalan, S.V., Simpson, A. &Sammon, C. Can real-world data really replace randomised clinical trials?. BMC Med 18, 13 (2020). https://doi.org/10.1186/s12916-019-1481-8

Practice uses knowledge, focuses on individual patient, has a short action span, the reward is immediate, it regards authority, follows custom, earns income, and encourages research. On the other hand, Research creates knowledge, focuses on groups, has a long action span, the reward is delayed, it questions authority, challenges custom, earns reputation, and enriches practice. There have been many attempts to generate clinical evidence from primary health care by systematic utilization of patient records. Many examples have been cited for the (1) deficient clinical data, (2) in-accurate input giving "garbage in garbage out", (3) insufficient follow-up, (4) very few fully completed case records with risk factors, co-morbidities, etc.With standardized protocols, checklists and assessment scales constituting the hard core of a renewed medical record in primary care, good clinical practice would be better supported and even integrated with research on a regular basis. That is how future clinical research, including observational studies and experimental trials, will be based on the practice and principles of primary care.

## Efforts at national level

### AYUSH

The Ministry of AYUSH is an initiative by the Government of India to promote the propagation and development of AYUSH systems of health care and medicine in India (Ayurveda, Yoga & Naturopathy, Unani, Siddha and Homoeopathy).To administer the different medicine systems encompassed by the Ministry of AYUSH, it has five research councils or departments, affiliated courses, and affiliated national institutes.

### Digital India movement

The Digital India Movement initiated by Govt. of India has made significant changes in IT sector. We are witnessing tremendous improvement in application of Information Technology (IT) in every Ministry and its organizations. Health is a major concern in India. Implementation of IT in healthcare is being accelerated to provide effective and better care for citizens of India.

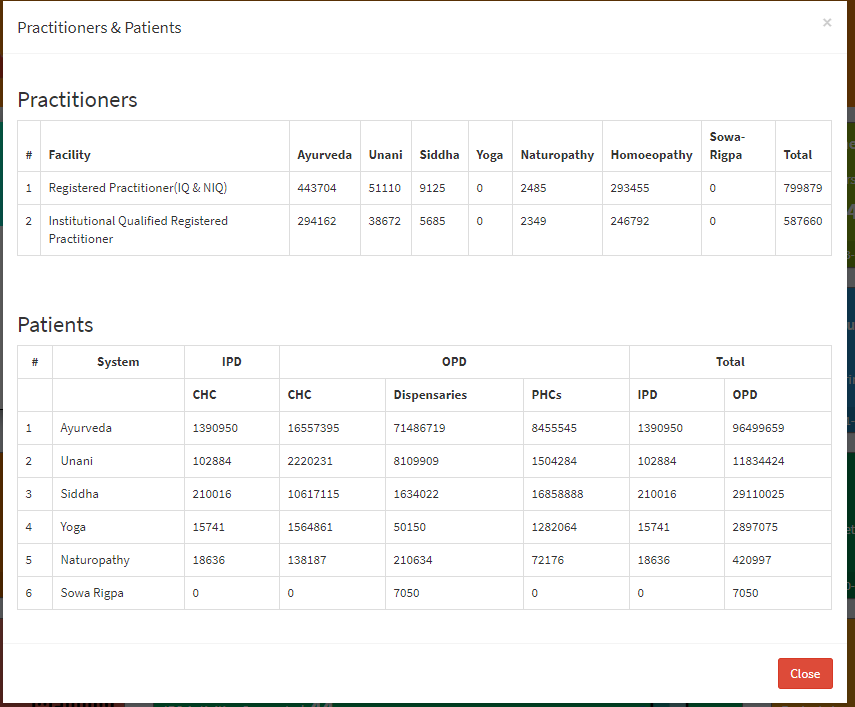
Ministry of AYUSH has taken initiative to get onboard in healthcare Information technology implementation projects. In this regard, Ministry of AYUSH has created AYUSH GRID to get all the IT projects under one umbrella. It is integration of IT projects exclusively meant for improvement and facilitation of AYUSH pan India.

The initiative is a poised to emerge as a game changer as it is the first of its kind citizen centric service from Government of India to provide Electronic Health Record and Personal Health Records facility to the recipients of AYUSH Health Services catered through Research Councils, National Institutes and other related agencies under Ministry of AYUSH, Government of India.

The dashboard available on AYUSH homepage (accessed on 25th May 2020), reveals great facts on various components.



Details of the number of ayurvedic practitioners and total number of patients treated in the picture below show almost 10 crore patients treated at some or the other point. There are approximately 140+ countries with less than 10 crore population, which provides a perspective on the size of data available at the AYUSH level. It remains to be seen how to convert “data into information, information into knowledge and knowledge into wisdom”.



### National AYUSH Morbidities and Standardized Terminologies E- Portal (NAMASTE Portal)

CCRAS along with other stakeholders has developed the portal. It provides information about Standardized terminologies and Morbidity codes along with dedicated data entry module for updating morbidity statistics in consolidated form as well as on real time basis.

<http://namstp.ayush.gov.in/#/index>

### AYUSH Research Portal

To disseminate the merits of AYUSH systems across the globe, a web-based portal for Research publications in AYUSH was launched in 2011, which is being maintained by NIIMH Hyderabad. The portalis successfully continuing and the information is being updated periodically.

<http://ayushportal.nic.in/>

### AYUSH Hospital Management System

A-HMIS is a comprehensive IT platform to effectively manage all functions of health care delivery systems and patient care in AYUSH facilities. THERAN (THE Research Application Nexus) - Hospital Management Information System developed by Siddha Central Research Institute, Chennai under Central Council for Research in Siddha, Ministry of AYUSH was near to the requirement to give complete solution of HMIS amongst various HMIS present in AYUSH Institutions of Govt. Of India, hence it was decided to upscale and customize it as per systems in AYUSH.

## Electronic Medical Record, medical IT infrastructure

In early 2000s countries like Canada, UK, New Zealand, Estonia, etc. started collecting data at national level. The key challenges experienced while implementing EMR were: (1) infrastructure creation, (2) Policy & regulations, (3) Standards & interoperability, and (4) Research, development & education.

India has a mixed system of healthcare consisting of a large number of government run hospitals, the private hospitals, family doctors and private medical practices. We see this trend reflected in the actual health seeking behaviour of communities where people tend to combine medicine systems like Allopathy, Ayurveda, Siddha, Sowa Rigpa, Unani, Homeopathy and Yoga depending on the nature of the disease [Darshan 1].

Prevention, wellness and treatment of Non communicable Diseases (NCDs) are a few of the core strengths of Ayurveda. The Ayurvedic strategies for immunity (*vyadishamta*) and tissue regeneration (*rasayanatantra*) can lead to new research. Its systemic pharmacology (*dravyagunashastra*), its thousands of brilliantly designed food and drug formulations (*bhaishajkalpana*) with undiscovered pharmacokinetics and dynamics can feed a new paradigm of drug discovery for syndromes and personalized nutrition science (nutrigenomics) for decades to come. Rigorously documented clinical experiences interpreted through Ayurveda-biology will deepen modernization of healthcare in India [Darshan 1].

Ayurvedic vaidyasusually use paper based case report to record a patient’s ayurvedic parameters along with other details of medical consultation. These are not typically exchangedwith other vaidyas. Increased use and interoperability with electronic medical records of digital Ayurvedic patient management systems is required.Based on a report published by AYUSH above, there are 4.5 lacs registered Ayurvedic practitioners. Even if 5% of doctors start using EMRs, i.e. 22500 doctors and if data for 2 new patients is entered every day (~225 working days) for the whole year, 50 lacs unique patient data can be generated in a single year. Currently, this gold mine of data is not yet built.

In general, the level of use of IT in the healthcare sector in India has been lower in comparison to other countries. The use is further lower in case of ayurvedic practice. The increased penetration of mobile phones and the availability of high-speed Internet offer the possibilities to provide healthcare services across the whole country. The Ministry of Health & Family Welfare and Department of AYUSH have taken several steps to create standards. It has issued guidelines for EMR standards and meta-standards. The implementation and penetration are yet to increase.

EMR data has been widely used for analysis and many papers have been published. These have generated supportive data for a variety of clinical outcomes, evaluation methods, and implementation of new technology or intervention along with awareness of unintended consequences; thus supporting the clinical decisions and aiding to improve the healthcare process or clinical outcomes.

## Need of this study and methodological framework

All of the thought leaders cited above point to the strengths of Ayurveda as well as the immediate needs.

* They have written about the research impact within the Ayurvedic paradigm
* They have pointed out that the research mustbe of high quality and it must be impactful
* They have indicated the need for experimental as well as experiential research
* They have already provided a few new solutions and have urged to the research community for new ways of tackling problems

The researchmust be directedat increasing patient benefit, while keeping it as cost-effectively as possible to permit efficient and viable service. These discussions give rise to the methodological framework covering short, mid and long term impacts, framework categories are as follows:

1. Policy makers – AYUSH and other relevant ministries,
2. Healthcare providers - Ayurveda Healthcare systems, General healthcare systems,
3. Hospital managements and individual clinicians,
4. Patients and
5. Universities and learning institutes – clinical communication, researchers to build vital evidence-base.

Thefollowing section defines a few questions falling in various categoriescovering frameworks listed above.

* Evidence base generation (EBM): EBM aims to improve quality of care through the integration of best research evidence with clinical expertise, patient and parents’ preferences.

|  |
| --- |
| 1. Is it possible to formulate an answerable clinical question (PICO format)? Patient, Intervention, Comparison, Outcome 2. Find the evidence 3. Understand the evidence 4. Apply the evidence 5. Evaluate the performance |

* Creation of a cohort and generating robust information

|  |
| --- |
| 1. Understand the natural history 2. Cohort identification 3. How to create a cohort of similar patients? What is the meaning of cohort in ayurvedic area? 4. Once the cohort has been designed then would it be possible to predict any questions, trajectory of diseases, etc. 5. Predicting the next complication or next disease, when and what? 6. Quantifying the effect of intervention 7. Construct evidence based guidelines |

* Description of the hospital database / operational insights

|  |
| --- |
| 1. How many patients are available in the database? Are there some simple summaries available summarizing the underlying data? 2. What are the characteristics of these patients? 3. What is % of males and females? 4. What is the age group? 5. Which countries, states, cities do they come from 6. How many times do they visit hospital? 7. What is the visit pattern? 8. What is the number of In Patients &Out Patients? 9. What kind of assessments are done at each visit? 10. What is the duration of visits overall for a patient? 11. What kind of diseases are reported? 12. How many diseases do they have? 13. What operational insights could be drawn? |

* Disease variations as seen in the database – does it point to epidemiology of Bengaluru?

|  |
| --- |
| 1. How many times the same disease is reported by a patient? 2. Are there differences seen in the diseases by gender? 3. Are there differences seen in the diseases by season? 4. Are there differences seen in the diseases by age group? 5. "How many disease - disease combinations can be identified,    1. some could be clinically meaningful,    2. some may be clinically meaningless    3. some could be rare combinations 6. Which diseases are reported a lot more than other diseases? 7. What kind of diseases are reported before an onset of a disease? 8. What kind of diseases are reported after an onset of a disease? 9. What is the duration of visits for each disease (episode, different episodes) for a patient? 10. How many doctors treat one patient through the course of their diseases at the hospital? 11. Which are the rare diseases identified in the database and any clinically meaningful document could be written? |

* Treatments prescribed and their variations as reported in the database

|  |
| --- |
| 1. How many different treatments are prescribed? 2. "How many different ayurvedic services are offered to these patients? 3. Which ayurvedic procedures are ordered for which diseases?" 4. What kind of treatments prescribed? 5. Any description of treatment types? -herbo mineral, classical, rasaushadhi, etc. 6. "Summary statistics for treatments    1. frequency    2. how many days a treatment is prescribed for    3. total duration of treatment prescription" 7. What treatment and treatment combination appears the most? 8. What is the total duration per treatment found, and does it provide any clinical meaning? 9. "Medicines based on metal formulations are under attack from non ayurvedic community, 10. What insights can be drawn about the rasa aushadhis?     1. which are the medicines?     2. for what diseases are they given?     3. for what duration?     4. Before providing the metal based treatment and after providing the metal based treatment, is there any difference in duration seen in treatments?     5. % of patients provided these medicines?     6. % of duration of all the duration of treatment given to these patients |

* Regulatory / insurance / education point of view

|  |
| --- |
| 1. How can anyone use these data as "secondary use"? 2. Insurance companies related aspects of data which has not been thought of so far? 3. Do the approved labels of medicines and prescriptions in the database match each other? 4. How can this framework be used as a supplementary material for any MD student? 5. Can TDU initiate registry for ayurvedic data from other hospitals, private ayurvedic doctors? 6. In absence of a true “hard clinical endpoint” what kind of “endpoints” be created?    1. short term outcomes (intent adeherence to the treatment)    2. intermediate term outcomes (actual adhere)    3. long term outcomes (overall health) |

* Research impact / operational impact

|  |
| --- |
| 1. Hospitals in India or any part of the world focus on treatment and not on research publications -- can a team be put together for the medical communication who are publishing papers as their primary job? 2. If there is no known profile of patients visiting the ayurvedic hospital, if this data is represented in the right form then will it provide novel information? 3. How can we measure the strengths and weaknesses of the practice? 4. How can we innovate and improve? 5. How to evaluate changes in results over time? 6. Is there a way to execute any longitudinal analysis? 7. Is it possible to build new hypothesis? 8. Is prescribed treatment truly personalized or are there enough overlaps? 9. Is it possible trace back the treatment regimen followed as per the classical fundamentals or not? 10. Is there a way to compare the demographics of patients from our hospital against the main stream western hospital? 11. How to corroborate information coming from     1. Patient level unit     2. Visit level unit     3. Disease level unit     4. Medical level unit |

* Information system challenges on a day-to-day basis

|  |
| --- |
| 1. How can our experience of doctors using the Health Information system be viewed?    1. do doctors like data entry?    2. do they see this as an additional burdern or an integral part of the work?    3. do they see the data entered by them reviewed at a later point by someone else "intrusion into their work |

## The proposed study in brief

Ayurveda has been practiced over many centuries in India. It will be safe to assume that the conceptual developments in ayurvedic knowledge base have taken place through every day observations and basic laws of nature. These fundamentals have been adjusted to the relevant times as per the passage of time, which is quite evident from vast literary history of Ayurveda, which covers subjects like pharmacology, principle of diagnosis and treatment for all branches of medicine and surgery, philosophical framework and logic, pharmacy and numerous pharmacopeias. Traditional texts enumerate more than described for each disease condition.

Generating credible evidence for such a large pool only through modern experimental means such as trials is very challenging. Methodologists are challenging current hierarchical evidence model and the circular model comprising observational research methods are proposed for CAM research. Ayurveda like any other system of medicine, is practiced more in clinics than in clinical research setting, where there are no artificial restrictions on usage of medicines, duration of treatment or type of patients to treat, which is next to impossible in a protocol driven clinical trial setting.

The TDU IAIM Ayurvedic hospital database is a pioneer in capturing electronic health or medical records (EHR, EMR) since year 2011. Till October 2017, it contained data for >51,000 patients, >1,50,000 visits, > 900 variations of disease types, >3,000 variations of medical procedures. This study will be one attempt in plugging the gap of missing empirical evidence, systematic analysis of observational clinical data. It is focused on IAIM clinical data for study of efficacy and safety trends.Given the availability of the data, we have an opportunity to unearth wealth of knowledge. This kind of study has never been carried out within ayurvedic area, providing us with the pioneering chance.

### Chapter outline

The subsequent chapters of the thesis are defined as follows:

1. Converting clinical life data into analyzable format
2. Clinical data understanding
3. Studying demographics and patient specific factors
4. Diagnostics and treatment data
5. Outcome and effect
6. Prediction / NLP

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Objective s as Chapters | Hospital management | Hospital physicians | Patients | Policy makers | Ayurvedic Healthcare (insurance) | General healthcare (insurance) | Clinical communication | Research related impact | Researcher (move after patients column) |
| 01 Converting clinical life data into analyzable format | x | x |  |  | x | x | x | x | x |
| 02 Clinical data understanding |  |  |  | x | x |  |  | x | x |
| 03 Studying demographics and patient specific factors | x |  | x |  |  |  |  | x | x |
| 04 Diagnostics and treatment data |  | x | x | x | x | x |  | x | x |
| 05 Prediction / NLP |  |  |  |  |  |  |  | x | x |
| 06 Outcome and effect |  |  |  |  |  |  | x | x | x |

(research related impact may be removed – will continue thinking)

**Potential opportunities within ayurveda:**

1. People are realizing the importance of Ayurveda and are turning towards it across world
2. Sustainable treatment solutions for non-communicable diseases readily available
3. Every day-to-day practice generates huge amounts of data, which is fragmented in nature; most of the data is stored on paper in individual practice:
   1. The size, quantity and depth are unknown, make people aware of the hidden potential
   2. Simple user friendly and effective data review tools for improved patient and doctor interaction
   3. Enhance the existing evidence further from day to day clinical practice data
4. Over the years, teaching methodology has not changed substantially and quite a lot of modern methodology, learning objectives of different kinds:
   1. Make complex relationships easily visually available
   2. Very few large scale studies like this have taken place in Ayurveda – explore the data and describe the findings, textually as well as diagrammatically
   3. Accumulate data for experienced vaidyas as well as new ones
5. Data collection and its usage at different units
   1. Doctor's unit of assessment: Whole patient
   2. Researcher unit of assessment: Disease condition, health condition
   3. Hospital Management Information System (HMIS) unit of documentation: Patient Visit
6. Do two practitioners follow a similar pattern of regimen in treating patients?

**Needs of transdisciplinary approaches:**

1. Create Transdisciplinary evidence to increase the scientific understanding outside of the community, then increase the confidence and thereby widening the user base
2. Contribution to Public health data creation based on large data at our disposal which is not marred by artificial boundaries imposed on patient disease conditions, treatments prescribed
3. Generate Real World Evidence (RWE) to supplement classical understanding which is rooted within the community
4. Make recommendations to the practitioners for standardized way of data collection, analysis and reporting which will support future RWE studies
5. Understand the hidden wealth of data for Transdisciplinary expansion of thoughts

**Why do we need transdisciplinary approaches: how strengths of multiple approaches should be used – how can multiple disciplines can cover the shortcomings – how 1 + 1 can become 5 and not 2? How to compensate for shortcomings and gain benefits.**

**Untapped potential of ayurveda (to make the communication more powerful, the communication is not global very localized, but if it has a potential of global application, this will make it wide open for a bigger audience, it is under-utilized as a public health tool, due to insufficient data, these solutions may help us bring it to the fore in public health, promotion of health, prevention of diseases):**

1. Shlokas to diagrams – generating supplementary material to classical texts
2. Numerical evidence for the shlokas (shastratatva, vyavahar)
3. Thought provoking work to generate new needs, unconventional use of the data
4. Expand the use of modern IT solutions like IT infrastructure, electronic health records, cloud, etc. within ayurvedic area where appropriate – Ayur IT solutions.
5. Take advantage of freely available cutting edge software(s) to create new approaches
6. Introduce statistical programming as a tool to ayurvedic area
7. Create a frame work of data analysis – may be available to MD students, replace their individual studies by this retrospective work
8. Generate viable financial model for making data available for insurance

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5950611/>

<https://www.tandfonline.com/doi/pdf/10.1080/028134300453278>

<https://bestpractice.bmj.com/info/toolkit/practise-ebm/why-do-we-need-ebm/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5299804/>  
<https://www.sas.com/storefront/aux/en/sppharmastat/60>

<https://www.sas.com/storefront/aux/en/sppharmastat/60622_excerpt.pdf>  
https://www.npr.org/2011/01/28/133306341/Science-Diction-The-Origin-<https://www.npr.org/2011/01/28/133306341/Science-Diction-The-Origin-Of-Physician>  
https://www.bartonassociates.com/history-of-physicians<https://www.bartonassociates.com/history-of-physicians>  
https://www.google.co.in/amp/s/m.hindustantimes.com/pu<https://www.google.co.in/amp/s/m.hindustantimes.com/pune-news/if-you-search-for-symptoms-online-when-sick-you-might-be-suffering-from-google-syndrome/story-lkbwf7x67gDsslS2yOfWbJ_amp.html>  
https://www.nepjol.info/index.php/AJMS/article/download/13929/12040<https://www.nepjol.info/index.php/AJMS/article/download/13929/12040>  
https://pharmaphorum.com/articles/a\_history\_of\_the\_pharmaceutical\_i<https://pharmaphorum.com/articles/a_history_of_the_pharmaceutical_industry/>  
https://www.who.int/medicines/technical\_briefing/tbs/Drug\_Regulation\_Histor<https://www.who.int/medicines/technical_briefing/tbs/Drug_Regulation_History_Present_Future.pdf>  
https://www.ncbi.nlm.nih.gov/books/NBK54296/<https://www.ncbi.nlm.nih.gov/books/NBK54296/>  
https://apps.who.int/iris/bitstream/handle/1<https://apps.who.int/iris/bitstream/handle/10665/46981/WH-1989-Aug-Sep-p6-8-eng.pdf?sequence=1&isAllowed=y>  
https://searchbusinessanalytics.techtarget.com/definition/data-exploration<https://searchbusinessanalytics.techtarget.com/definition/data-exploration>  
https://core.ac.uk/download/pdf/34578672.pdf<https://core.ac.uk/download/pdf/34578672.pdf>  
https://www.hpc2n.umu.se/sites/default/files<https://www.hpc2n.umu.se/sites/default/files/events/para06/papers/paper_213.pdf>

Diagrams needed, (flowchart):

Differences in the diagnosis (ayurvedic practice)

Modern medicine: with same disease, biomarker (surgery or medicine)

Ayurveda:

Put an emphasis on why RCT is a challenge for ayurveda.

RCTs may not be applied due to the complexity, so ayurveda suffers.

If we use the latest analytical tools it may uncover authentic information, unbiased, in absence of RCTs, using these tools, we may be able to get placebos are difficult to develop,

Do not use the reductionist approach – we wouldreductionist approach.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Hospital management | Hospital physicians | Patients | Policy makers |

|  |
| --- |
| Summary of the chapter:  * This chapter outlines the need to undertake the study, by providing the historic perspectives on medicine, pharmacy, development of hospitals throughout the world, internet era, and the Indian context relating to modern medicine as well as ayurveda * The thoughts from the thought leaders in the field of ayurveda are profound and they call for the modern methods, new approaches, innovative strategies to be attempted to take the science forward * There is reflection on the type of evidence generated through different controlled experiments (RCTs) and life experiments (Observational, Experiential) – some evidence about how these multiple approaches could yield similar results * This chapter has section on a few questions from diverse points of views and seeks answers – some of these answers are hidden in the every dayayurvedic clinical practice – which is still largely untapped * How can data analysis of electronically captured data help in advancing collective understanding by at 1% is defined in the following chapters |

## Converting real life clinical data into analyzable format

## Background

During the passage of time, revolutions in technology have continually increased the creation of information and its exchange. When humans advanced from spoken communication to written records they could steadily produce documents, papers, texts, books that some could read and learn from. Growth of writing was compilation of articles, tables and other records, which grew to become libraries. The ability to effortlessly widen accumulated data had to wait until the 15th century. Around 1439, Johannes Gutenberg developed the printing press, causing an astonishing growth in the sharing of information at an economical cost.

The 20th century generated a remarkable growth in the publication of scientific journals and monographs, most of which were not critically reviewed, as most physicians had no way to access to the existing medical information. Towards the late 20th century, the spread of computers and the internet providing immediate virtual access to diverse information has entirely changed the way knowledge is collected, stored, and circulated. The flow of information has been increasing at almost exponential levels. Today, data sets are measured in zettabytes (10^21 bytes).Cost-effectively collected and stored data allows researchers across the world to successfully advance understanding of science and medicine.

International Data Corporation (**IDC**) is one of the premier global providers of market intelligence, information technology, and a host of other areas. They predicted in a report released in Dec 2018 that the cumulative sum of the world’s data will grow from 33 zettabytes this year to a 175ZB by 2025, for a compounded annual growth rate of 61%. A zettabyte is a trillion gigabytes multiplied that by 175 times. This expansion of data has been seen in every industry, in every corner of the world. The unending increase in the quantity and flood of information denotes an important professional opportunity and a challenge at the same time for those in medicine and science.

As indicated by Toby Cosgrove MD, from Cleveland, medical information will soon double every 73 days, in year 2020, which used to take approximately 3.5 years in 2010. An estimated 800,000 papers are published yearly in 5,600 medical journals. It is projected that 12,000 new articles and 300 randomized controlled trials are added to Medline each week, and that new medical articles appear at a rate of one every 26 seconds. To be able to generate any kind of analysis and make accurate predictions, there is a need to access, connect various sources, collate, and consume all of the data. Data can be produced, obtained, and stored in a numerous number of structures.

Raw data, like unrefined gold buried deep in a mine is a precious resource. It is often:

* Inconsistent. It contains both relevant and irrelevant data
* Imprecise. It contains incorrectly entered information or missing values
* Repetitive. It contains duplicate data

Before anyone can benefit from raw data, it needs to be extracted, filtered through, understood, and transformed into something that could be analyzed. Understanding the scope of data being analyzed and seeing the changes made to the data can accelerate the entire process of going from “information to building wisdom”.

Data access, extraction, cleansing, transforming and consistent data are a few steps of data preparation or data wrangling. One of the surveys carried out by Forbes estimates that data cleaning accounts for up to 80% of the development time and cost in data warehousing projects. The subsequent sections will provide us with detailed information to go from an observation to information from a database point of view.

## Material and Methods:

During an appointment at a hospital, diverse types of data are collected. Raw data are observations about individual patients created by the treating doctor at a hospital. These data may be in the form of measurements of patient’s characteristics such as age, gender, height, weight, blood pressure, heart rate, etc. Raw data may also include description of the medical history, physical exam information, clinical laboratory results (e.g., serum lipid values, hemoglobin levels), whole exome or genome sequences, imaging results (e.g., X-ray, magnetic resonance imaging [MRI]), procedure results (e.g., electroencephalogram [EKG], endoscopy), or self-reported data (e.g., symptoms, quality of life).

I would present a flow diagram of how the hospital visit unfolds creating data through patient and doctor interaction.

|  |  |  |
| --- | --- | --- |
| Step | Description | Pictorial representation |
| 1 | A patient visits a hospital | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9E73FB80.tmp |
| 2 | Medical data/ patient history is populated, either on paper or electronically | Electronic Health Record Stock Illustrations – 823 Electronic ... |
| 3 | Medical assessments are done if prescribed   1. Pathology reports 2. ECGs 3. CT scans 4. Ayurvedic tests 5. Panchakarma | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\840CC8A6.tmp  Heartbeat Line Icon. Heart Rhytm. ECG. Cardiogram Stock Vector ...  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\23EF0755.tmp  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\DD1FC92B.tmp  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\44C8A771.tmp |
| 4 | All the above combined together generates the Electronic health record / Electronic Medical record is generated  This can be called as “raw data” or “source data” | Clipart Electronic Health Record |
| 5 | The data from pathology labs, ECGs, CT scans come from multiple labs, hence there is a need of “data transfer” from different locations to the source database | Paper Sheets Having Indicating Arrows Line Icon Data Transform ... |
| 6 | As seen in step 3: Source data can come from many sources, hence there is a need to integrate data sources into 1 source called as any one of the following:   1. source database 2. data ware-house 3. data lake 4. data mart | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\ACE814D8.tmp  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\E79DFCBA.tmp |
| 7 | Once the source data is electronically made available, it has to be made accessible through a system | Monitor Screen Check Mark Symbol Padlock Data Access Icon — Stock ... |
| 8 | This point onwards, the data preparation steps begin   1. Data extraction from source to a staging area 2. Data is transformed as per the needs 3. Data is loaded into areas for the “end users” to use per needs | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\FC837A54.tmp |
| 9 | Data extraction from the source systems: Source to analysis area | Data Transfer Icon - Structured |
| 10 | Data is transformed by performing various steps:   1. Data filtering | Vector black filter data icon set. ... | Stock vector | Colourbox |
| 10 | 1. Data cleansing | The Dirty on Data Cleansing & Appending |
| 10 | 1. Data deduplication | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7DA8F17.tmp |
| 10 | 1. Data merging | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3628E2E0.tmp |
| 10 | 1. Data transposing | Questions from Alteryx Training | InterWorks |
| 10 | 1. Data mapping | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1F9097E3.tmp |
| 10 | 1. Data outlier detection | Local outlier factor - Wikipedia |
| 10 | 1. Data imputation, as necessary | Chapter 4 Multiple Imputation | Book_MI.utf8.md |
| 10 | 1. Data aggregation | Download Free png Free Aggregation Icon 203770 | Download ... |
| 11 | Finally getting raw data to structured format.  Some of these steps are repeated on a periodic basis to ensure that the database is maintained at the expected levels of performance. | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\413B5782.tmp |
| 12 | Some of the data transformations are performed specific for each analysis so that the data is fit for purpose of that specific analysis |  |
| 12 | 1. Hospital management | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\62C2BC8D.tmp |
| 12 | 1. Individual researcher | Analytics - Free people icons |
| 12 | 1. Health authorities | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\553C02C8.tmp |
| 12 | 1. Data mining | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\69771C16.tmp |
| 12 | 1. Various publications | Modern Outline Style Data Analytics Icons Collection Stock ... |

The following visualization provides us with the Flow diagram from data source to final usage by various usage types.

Flow diagram from data source to final usage by various usage types

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Sources | Staging area | Ware house | Data marts | Usage |
| Data access:  Monitor Screen Check Mark Symbol Padlock Data Access Icon — Stock ... | Data Transfer Icon - StructuredVector black filter data icon set. ... | Stock vector | ColourboxThe Dirty on Data Cleansing & Appending | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1F9097E3.tmp | Chapter 4 Multiple Imputation | Book_MI.utf8.md |  |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Operational system  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Coding dictionaries  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Clinical system  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\50D212FE.tmp  Flat files information | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Calculations and transformations | Data Warehouse Icons - Download Free Vector Icons | Noun Project  Curated and consistent data storage | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Operational data  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Pharmacy data  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Patient level data | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\62C2BC8D.tmp  Hospital management  Analytics - Free people icons  Researchers  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\553C02C8.tmp  Health authorities  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\69771C16.tmp  Data mining  Modern Outline Style Data Analytics Icons Collection Stock ...  Various documents |
|  | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\7DA8F17.tmpC:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3628E2E0.tmp  Questions from Alteryx Training | InterWorks | Local outlier factor - Wikipedia | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\413B5782.tmp |  |

## Real actions adopted on acquiring data

The framework defined above has been followed to reach from the source data stage to analyzable data.

## Data access

Since data is stored differently based on the type of data, different sets of tools are needed to connect to the respective data sources. E.g., structured data is stored in relational databases and uses SQL to query the data, unstructured data stored in Hadoop would use Hive, Spark or Pig, data extracts for file formats like CSV, TXT, JSON, XML, etc., and for other formats tools like Python and R are used.

The details for accessing the hospital management system are as follows: I was provided "read-only" access to the hospital database, so that there are no accidental updates to any records as well as no risk of changing any source information.

1. Install PostgresSQL locally on the system and then connect to the database as per details below.
2. Install Cygwin terminal locally on the system.
3. Login using the Cygwin terminal (the following command will prompt for password): psql -h 54.244.12.255 -p 5432 -d iaim -U iaim\_ro
4. Postgress Data Base details are as follows:
   * Hostname: 54.244.12.255
   * port: 5432
   * user: iaim\_ro
   * password: a1b2c3

This way, I was able to remotely access specific version of database without interfering in day-to-day hospital transactions.The data available in the SQL tables were used for the analysis. There were a lot of lab pathology reports uploaded into the database as pdf files have not been used to computational complexities.

## Part 1

### Details of the database

The database had approximately 200 tables. They covered various components of the hospital’s day-to-day functions right from the operational data to the patient level clinical information. The high level of classification of data types:

1. Operational tables:
   1. Hospital charges – IP, OP
   2. Operation theater charges
   3. Inventory of equipment
   4. Doctor charges, etc.
2. Reference dictionaries
   1. Disease codes
   2. Ayurvedic services
   3. Medication names
   4. Mast list of lab tests
   5. Names of city, state, Countries
3. Doctor details
   1. Doctor ID
   2. Relevant ward information
   3. Internal / Visiting / Part time / Full time
4. Patient information
   1. Patient details,
   2. Visit details
   3. Vital signs
   4. Registration details
   5. Discharge details
   6. Lab data details
   7. Diet details, etc.
5. Tables related to managing access levels and other IT related contexts

For this study, the following data was not used to avoid any controversies as well as to keep patient confidentiality:

1. Hospital monetary details
2. Doctor’s details
3. Patient details of sensitive nature – name, phone number, socio economic class, etc.

The next table presentsinventory of tables (some tables have been omitted as per the above section). The tables names marked in yellow colors are used for the creation of the analysis ready datasets from the unending puzzle of all the tables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| action\_rights | diet\_prescribed | hospital\_technical | package\_componentdetail | patient\_registration | section\_field\_options | store\_item\_batch\_details | test\_details |
| admission | discharge\_format\_detail | icu\_bed\_charges | package\_item\_charges | patient\_section\_details | service\_consumable\_usage | store\_item\_details | test\_org\_details |
| anesthesia\_type\_charges | doctor\_charges\_backup | ip\_bed\_details | package\_prescribed | patient\_section\_details\_orig | service\_documents | store\_item\_lot\_details | test\_results\_master |
| area\_master | doctor\_charges\_op\_backup | ip\_prescription | patient\_activities | patient\_section\_forms | service\_master\_charges | store\_patient\_indent\_details | test\_visit\_report\_signatures |
| bed\_details | doctor\_consultation | item\_supplier\_prefer\_supplier | patient\_consultation\_field\_values | patient\_section\_image\_details | service\_master\_charges\_backup | store\_patient\_indent\_main | test\_visit\_reports |
| bill | doctor\_consultation\_charge | manf\_master | patient\_demographics\_mod | patient\_section\_values | service\_org\_details | store\_po | tests\_conducted |
| bill\_activity\_charge | doctor\_medicine\_favourites | medicine\_dosage\_master | patient\_deposits | patient\_service\_prescriptions | services | store\_po\_main | tests\_prescribed |
| bill\_adjustment | doctor\_op\_consultation\_charge | medicine\_id\_health\_authority\_unique | patient\_deposits\_setoff\_adjustments | patient\_test\_prescriptions | services\_prescribed | store\_reagent\_usage\_details | theatre\_charges |
| bill\_charge | doctor\_org\_details | message\_recipient | patient\_details | ppfv\_form\_detail\_id | stk\_chkpt | store\_reagent\_usage\_main | diet\_charges |
| bill\_charge\_adjustment | dyna\_package\_category\_limits | mrd\_casefile\_attributes | patient\_details\_patient\_phone\_country\_code | preauth\_prescription | stock\_issue\_details | store\_reorder\_levels | url\_action\_rights |
| bill\_receipts | dyna\_package\_charges | mrd\_codes\_doctor\_master | patient\_discharge | preauth\_prescription\_activities | stock\_issue\_main | store\_retail\_customers | user\_services\_depts. |
| complaintslog | dyna\_package\_org\_details | mrd\_codes\_master | patient\_documents | prescribed\_medicines\_master | store\_adj\_details | store\_sales\_details | visit\_vitals |
| consultation\_charges | equipement\_charges | mrd\_diagnosis | patient\_general\_docs | progress\_notes | store\_adj\_main | store\_sales\_main | vital\_reading |
| consultation\_org\_details | estimate\_bill | mrd\_observations | patient\_hvf\_doc\_values | registration\_charges | store\_checkpoint\_details | store\_stock\_details | section\_field\_desc |
| deposit\_setoff\_total | estimate\_charge | operation\_charges | patient\_medicine\_prescriptions | sample\_collection | store\_estimate\_details | store\_transaction\_lot\_details | section\_master |
| diagnostic\_charges | favourite\_reports | operation\_org\_details | patient\_other\_medicine\_prescriptions | sch\_resource\_availability | store\_grn\_details | store\_transfer\_details | ha\_item\_code\_type |
| diagnostic\_charges\_backup | fixed\_asset\_master | other\_services\_prescribed | patient\_other\_prescriptions | sch\_resource\_availability\_details | store\_grn\_main | store\_transfer\_main | package\_charges |
| diagnostic\_reagent\_usage | follow\_up\_details | outsource\_sample\_details | patient\_packages | scheduler\_appointment\_items | store\_indent\_details | supp\_inv\_id | patient\_prescription |
| diagnostics | growth\_chart\_reference\_data | pack\_org\_details | patient\_pdf\_form\_doc\_values | scheduler\_appointments | store\_indent\_main | supplier\_master |  |

### Understanding Source Databases

Comprehension of the database is more than knowing how it is built with tables, views, and relationships. In order to write meaningful queries one needs to understand how real world data was decoded and stored in the database.

**Tables:**When constructing queries it is important to understand a table’s purpose. Is the table used to organize patient visit data or a list of Ayurvedic prescribed medicines or other medical personnel related information etc.?

Before writing a query look over database’s table names. In many cases the names reveal the main topic of the tables. If looking for diagnosis data, then chances are the table will be named something akin to “Patient\_diagnosis.”

The tables listed as (1) PATIENT\_DETAILS should contain details about patients. (2) IP\_PRESCRIPTIONS table should have medications prescribed to patients who have been hospitalized for some reason. (3) STATE table looks like a reference table with names of states of India.

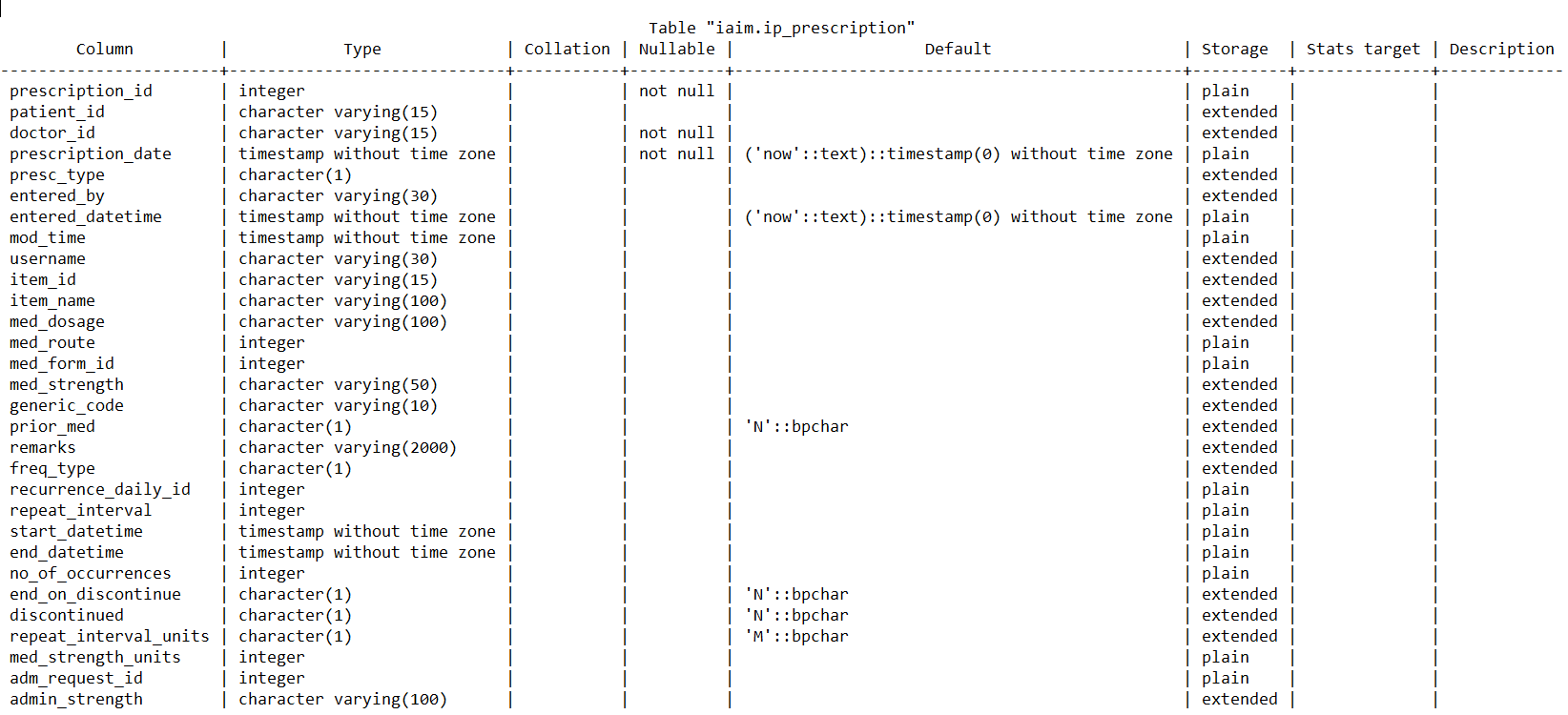
**Columns:**A table’s columns give a lot of information. It is important to have appropriate column names which explain the purpose of each column. Listed below are the column names in PATIENT\_DETAILS:

1. mr\_no
2. country
3. patient\_state
4. patient\_city
5. dateofbirth
6. patient\_gender
7. death\_date

The column names in this table are quite self-explanatory, with one exception; assumption that “mr\_no” is the unique patient identification variable.

The column’s data type also gives us theclues about what kind of data is expected in the column.E.g. IP\_PRESCRIPTION table has 30 columns or variables, see the screenshot below. The “type” column provides the details on the data type presented in the table.

* The table based on the name looks like related to the IP medication prescription
* Patient\_id and prescription\_id are key variables in the table
* Patient\_id: is the unique visits ID for each patient at each visit (mr\_no is the real patient ID, not present in the current table; this is a unique code which does not change)
* item\_id and item\_name are the columns which contain the medication name and ID
* All other variablesexplaining the course of the medication are present: start\_datetime and end\_datetime (start and end of medication), med\_route (route of administration), med\_dosage (doage), freq\_type (frequency)
* There a few variables containing system related information, these could be dropped from “staging area” as they are not useful for our intended analysis



Keys and Relationships: By inspecting the table names and columns some logical reasoning gets built, but how are they related to each other needs to be explored further?This is where it makes sense to review the relevant table’s primary keys to understand what values are used to identify the tables and to see if foreign key(s) from other tables can be used to make a relation.Typically, column names in across tables are named the same. If they are not the same and the same value is stored in multiple variables then it possesses additional challenges in establishing relationships between tables.

## Part 2

### Data extraction for downstream activities

Extraction is the process of retrieving data from a source system for further use in a data repository environment. After the extraction, data can be transformed and loaded into the data warehouse. This is the most challenging aspect of Extraction-Transformation-Load, as extracting data correctly will set the stage for how subsequent processes will go. Planning and creating the extraction process is one of the most time-consuming tasks in the whole process. The source systems might be very complex and thereforedeciding which data needs to be extracted can be challenging.The data has to be extracted several times in a recurrent periodic way to ensure up-to-date data as comparedto the warehouse.

There are several important considerations to be taken into account while designing extraction methods.

1. Full extraction: the data is completely extracted from source to repository every time, if the data size is very big then this step becomes very time consuming
2. Incremental extraction: at a particular point in time, only the data that has changed as well as updated since a well-defined event will be extracted, this is less time consuming compared to the first method. But technically this approach is more difficult as “time point”, “update”, “change” are difficult to define with complete accuracy across the whole organization.

An analyst or the team involved in this process must diagnose the data. Are the data responsive to the current analysis questions? What format are they in, and how much effort is required to put them into a format expected by downstream analysis tools? Are there data quality issues, such as missing data, inconsistent values, or unresolved duplicates? Next, the analyst must decide whether to continue working with the data, and, if so, the data must be transformed and cleaned into a usable state.

The extraction of the data allows subsequent analysis steps to be kept independent of the source data that is prone to changes, inconsistencies, duplications, etc.To optimize the overall performance of data preparation process, limit the fields needed for your analysis.

### Data extracted from Hospital database

In our study, we had different versions of data, details in the table below. The analysis carried out on different versions of data provide numerically different answers, but the overall trends experienced tend to be in the same direction.

|  |  |  |  |
| --- | --- | --- | --- |
| Data version | Version 1 | Version 2 | Version 3 |
| Date time frame | From start of the hospital to Oct 2016 | From start of the hospital to Oct 2017 | PDF file version of data for 15 In Patients |
| Data domains | Lab  Vital signs  Diagnosis | All the available data in the hospital database | Specific patient visits case report forms |

## Part 3

### Data preparation

Reliable and reproducible data preparation facilitateswell-organized analysis, minimizes mistakes and inaccuracies that occur to data through processing, and makes all processed data accessible to users. Below are the steps involved in data preparation

### Merging, joining:

Combine/enrich relevant data from different datasets into a new dataset. Joining data is one of the most important functions of data transformation. A “join” is an operation that connects two or more database tables by their matching columns. This establishes a relationship between multiple tables, which merges table data together so a query can be made on the resultant data.

### Appending

Combine two similar datasets into a larger dataset

### Filtering

Rule-based reduction of a larger dataset into a smaller dataset. Data filtering includes techniques used to refine datasets. The goal of data filtering is to refine a data source to only what the user needs by eliminating repeated, irrelevant data. Data filters can be used like this to amend query results and data reports. Data filtering involves the selection of specific rows, columns, or fields to display from the dataset.

### Deduping

Remove duplicates based on specific criteria as defined. Data deduplication is a data compression process to identify and remove repeated copies of information. Deduplication allows storageof one unique copy of data in data warehouse or database. This process examinesincoming data and compares it to data that is already stored in the system. If the data is already there, deduplication algorithms delete the duplicate information while creating a reference to it.

### Cleansing

Data cleansing involves deleting out-of-date, inaccurate, or incomplete information to increase the accuracy of data. The process might include parsing data to remove syntax errors, deleting record portions, and amending typos. It could also involve fixing duplication problems that result from merging multiple datasets.Data cleansing involves identifying incorrect data values and then either correcting them or rejecting them. They deal with INVALID values in single data elements or correlation across multiple data elements. This can be helpful in improving the accuracy of data.

Automated data cleansing programs can identify wrong values but generally cannot correct them. They can correct values only through synonym lists or correlation against tables showing valid combinations of values. Most of the time they identify a wrong value but cannot correct it. The only alternative is to change the value to unknown (NULL in most systems) or to reject the observationin which the bad value is contained. If it is rejected, they can either drop the target (creating a bigger problem) or manually investigate the value and reenter it into the target after correction.

Dropping observations has many problems. First, lossof some data. The correct data in other data elements of these observations may be more important to the target than by dropping the entire observation. Added problem is that it may create a structural problem relative to other observations in the same or other tables. Rejecting an observation may have the effect of causing many other observations to be rejected later, when referential constraints are enforced upon load.

### Transforming

Convert missing values or derive a new column from existing column(s). Data transformation is the process of extracting good, reliable data from these sources. This involves converting data from one structure (or no structure) to another to integrate it with a data warehouse or with different applications. It allows to expose the information to advanced business intelligence tools to create performance reports and forecast future trends.

### Aggregating

Roll up data to have data for analysis. Data Summarization is similar to data aggregation. It refers to the creation of different business metrics through the calculation of value totals.Data aggregation is a process that searches, gathers, summarizes and presents data in different reports.

### Format revision

Format revisions fix problems that create trouble from variables having different data types. Some variables might be numeric, and others might be text. One data system could treat text vs. numeric information differently, so standardize the formats to integrate source data with the target data schema. This could involve the conversion of male/female, date/time, measurements, and other information into a consistent format. Field lengths can also be an issue—especially if the target schema has smaller character limits. In these cases, it may be necessary to standardize the length of fields by breaking up long serial numbers into their smaller parts and putting them into separate columns. Additionally, format revision could involve splitting up a comma-separated list of words or numbers into multiple columns.

### Key Restructuring

When the tables in a data warehouse have keys with built-in meanings, serious problems can develop. For example, if a patient ID serves as a primary key, changing the patient ID format in the original data source means that the number would have to change everywhere it appears in the data system. That would cause a cascade of updates that over-burden or slow down the system. By drawing key connections from one table to another, key restructuring optimizes the data warehouse for speed and efficiency.

### Bucketing/Binning

This transformation is used to change a numeric series into fixed, categorical ranges, say, from {2,5,8…} to {2-5, 6-9, 10-13…}. E.g., the seasonal fluctuations in diseases. Bucketing/binning allows isolation of noisy data. The focus away from short-term volatility provides a real representation of trends over time.

### Z-Score Normalization and Max-Min Scaling

In scaling, data ranges are re-scaled.In z-score normalization, individual data features have zero-min and unit variance. Scaling is important because datasets often contain elements in varying units and ranges. This simple transformation allows for a compelling visual check as well.

### Imputation

Missing values are one of the most common problems faced in data. The reason for the missing values might be human errors, disruptions in the data flow, privacy concerns, and so on. Whatever is the reason, missing values affect the analysis. Some algorithms drop the observations whichhave missing values. On the other hand, most of the algorithms do not accept datasets with missing values and gives an error.

### Numerical Imputation

Imputation is a better option rather than dropping because it keeps the data size. However, there is an important selection of what and how to impute to the missing values. This “imputation” is a PhD topic in itself and too vast to explain in a short space. Imputations by statistics of central tendency (mean, median, min, max), regression methods, multiple imputations (same value imputed n numbers of times), chained imputations (imputations based on a certain sequence), etc. are a few methods available.

### Handling Outliers

Before mentioning how outliers can be handled, I want to state that the best way to detect the outliers is to demonstrate the data visually. All other statistical methodologies are open to making mistakes, whereas visualizing the outliers gives a chance to take a decision with high precision. Outlier Detection with Standard Deviation and with percentiles is easily possible. As the observations are real data, whether to drop them or keep them is a dilemma faced by every researcher.

## Part 4

## What is this data used for?

It is very important to think through the data preparation stage about the data holistically.It is important to think about how people will use the data prepared. Understanding this context will help in determining which data set to use, how much data to bring into data wearhouse, and how to ultimately structure and shape the data. To get started, answer some basic questions:

### Who is doing the analysis?

Consider the end users of the final data set. E.g., is the analyst the sole user who will access and understand all parts of the data for detailed analysis? On the other hand, will someone in a different role use the data set,? If it is the second option, then trim down the data set to only those measures. In this case, join the data and fact tables to get the information. Audience is critical whilepreparing data, similar to while creating a dashboard.

### What type of questions need to be asked or answered?

In the data preparation process, it is important to understand how people will use the final data set—for complex analysis or for a quick summary. This detail influences the data preparation process significantly, determining both the amount of effort and detail. An analyst typically predicts the most common questions that people will ask of the data based on understanding of strategic business priorities, but there will likely be unanticipated questions that pop up. While preparing a data set, there is a balance between serving the immediate questions and allowing for further exploration. For example, someone may see a pharmacy sales trend during the last six months, but digging into a spike during a particular week requires deeper analysis and a daily granularity of the data.

## Part 5

## How to do it?

Data derivation, Data derivation involves the creation of special rules to “derive” the specific information wanted from the data source. Data derivation allows to create a set of transformation rules.

### Prepare the way you think

Data preparation has a lot of different components, from restructuring to reformatting to cleaning, and should not be constrained by a specific order.

### Compartmentalize each step

Creating new steps for a specific set of actions keeps your flow nice and tidy. Think of steps as folders in filing cabinet - organize files by their subject, making it easier to find. Similarly, the steps in the flow should group a set of changes that capture a particular task. Keep these actions in the same step, and add a descriptive name to help you understand the flow later on. This process builds in-built documentation sharing the flow with other analysts, it lets them find and reference the same actions, giving them a way to easily make any edits.

### Modular approach to make maintenance easy

Document for reproducibility and collaboration, Staying organized throughout your preparation process is essential when you need to revisit and make a change to some step in the process. While there is a need to follow a specific set of instructions to clean the data,the data prep process will be a lot easier to edit and update.

## Exploration of the database

As the aim of this study is to understand the every dayayurvedic clinical practice, we further explored the data and thought of building one of the many datasets – using clinical information of patients. This dataset has patient’s visit wise, longitudinal data from the very first day of hospital visit to the last day of hospital visit in period of 2011 until Oct 2017. The case report form at each visit captures disease and medication data, along with demographic, background data and a few more characteristics (outlined later in the document). This data creates documented complete picture of each patient from various parts of the database including (1) In patient visits, (2) Outpatient visits, (3) diseases reported as per Ayurvedic Classification dictionary, (4) Medication prescribed, (5) Ayurvedic services prescribed. These components of data are logically arranged in one dataset by using various data transformation steps. In addition, there are new variables derived to create necessary information for the potential analyses. Let us go through the challenges experienced to assemble the “reference dataset” from the source data and practical explanation of the “data preparation” steps.

1. The database was manually explored using various SQL programming commands, the variables and observations were checked from numerous tables
2. Patient information and key variables needed to be understood: unique patient ID is MR\_NO, and unique ID for individual visit is PATIENT\_NO (many tables containing patients’ clinical information have this variable as the key variable)
3. Reference files needed to be used to reformat the coded variables
4. First section of the creation:
   1. Extract relevant data tables from the source database
   2. Transform the variables, join the tables based on logical link
   3. Create “staged data” or “snapshot of source data”
   4. Reference files (disease categories, Indian seasons) which are needed for calculations are developed using expert’s help
5. Second section of the program:
   1. Cleanse the tables
   2. Transform the tables for combining
   3. Join the tables on logical link
   4. Derive additional variables as necessary
   5. Filter the data using reference files created in the earlier section

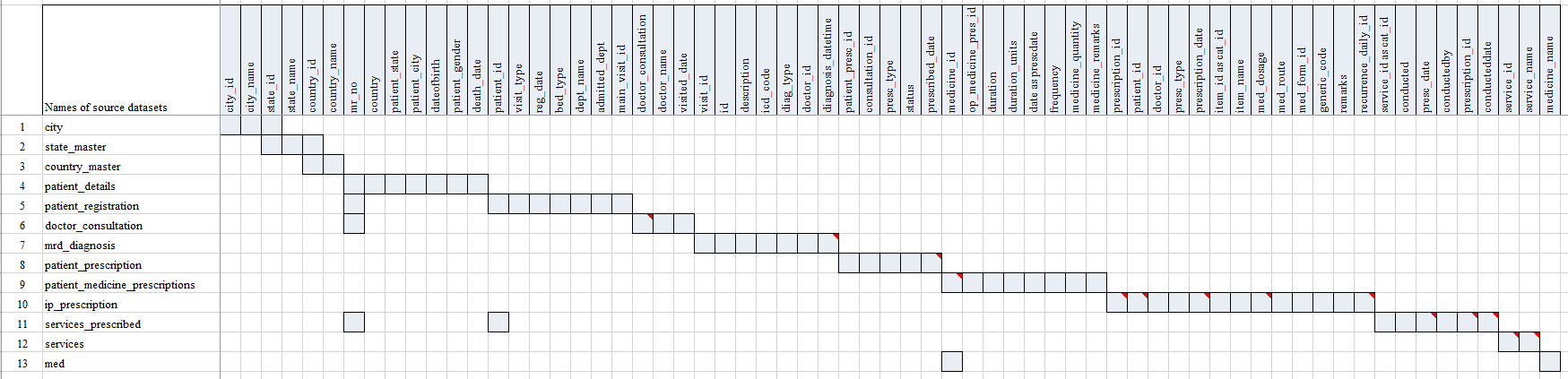
In this process, we have used 13 source datasets (5 reference datasets and 8 patient level datasets) and ~65 variables to generate the necessary snapshot of the source data. These have been re-arranged into 6 datasets and ~40 variables. 3 additional reference files are used for further processing. 1 final dataset having ~30 variables from source and ~30 newly derived variables is built.

## Flowcharts of the algorithms

Methodology section 3a, 3b, 3c, 3d:

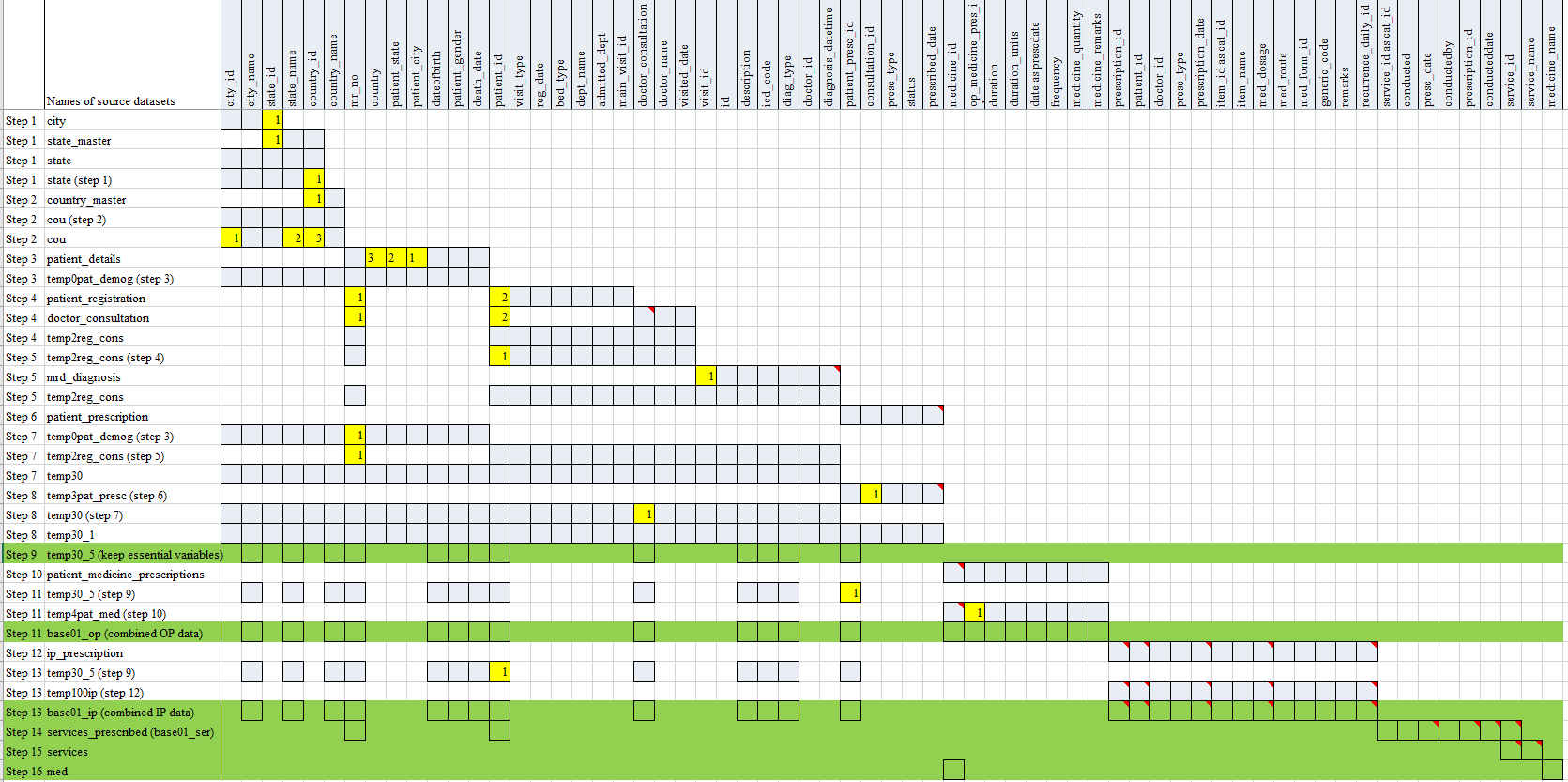
Picture option 3a:

Section 1: extraction from Source database (13 datasets, ~65 variables)



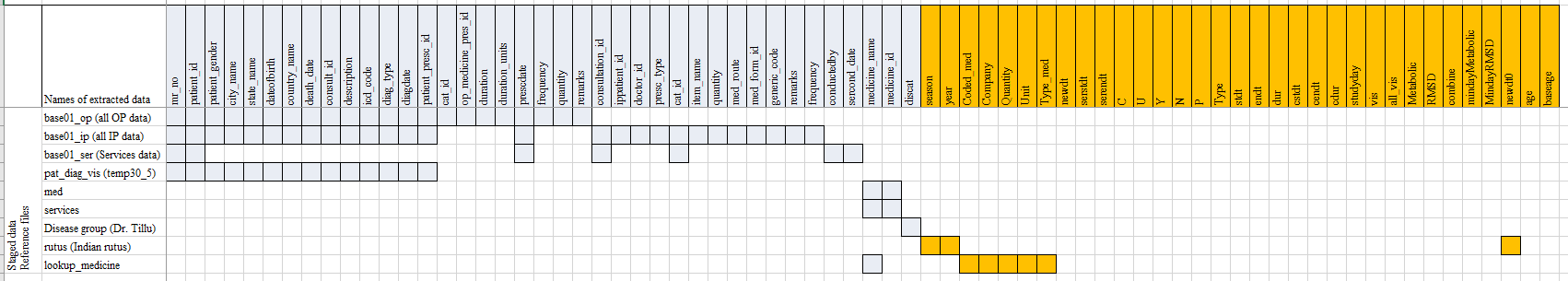
Picture option 3b:

Staged data converted into 6 datasets



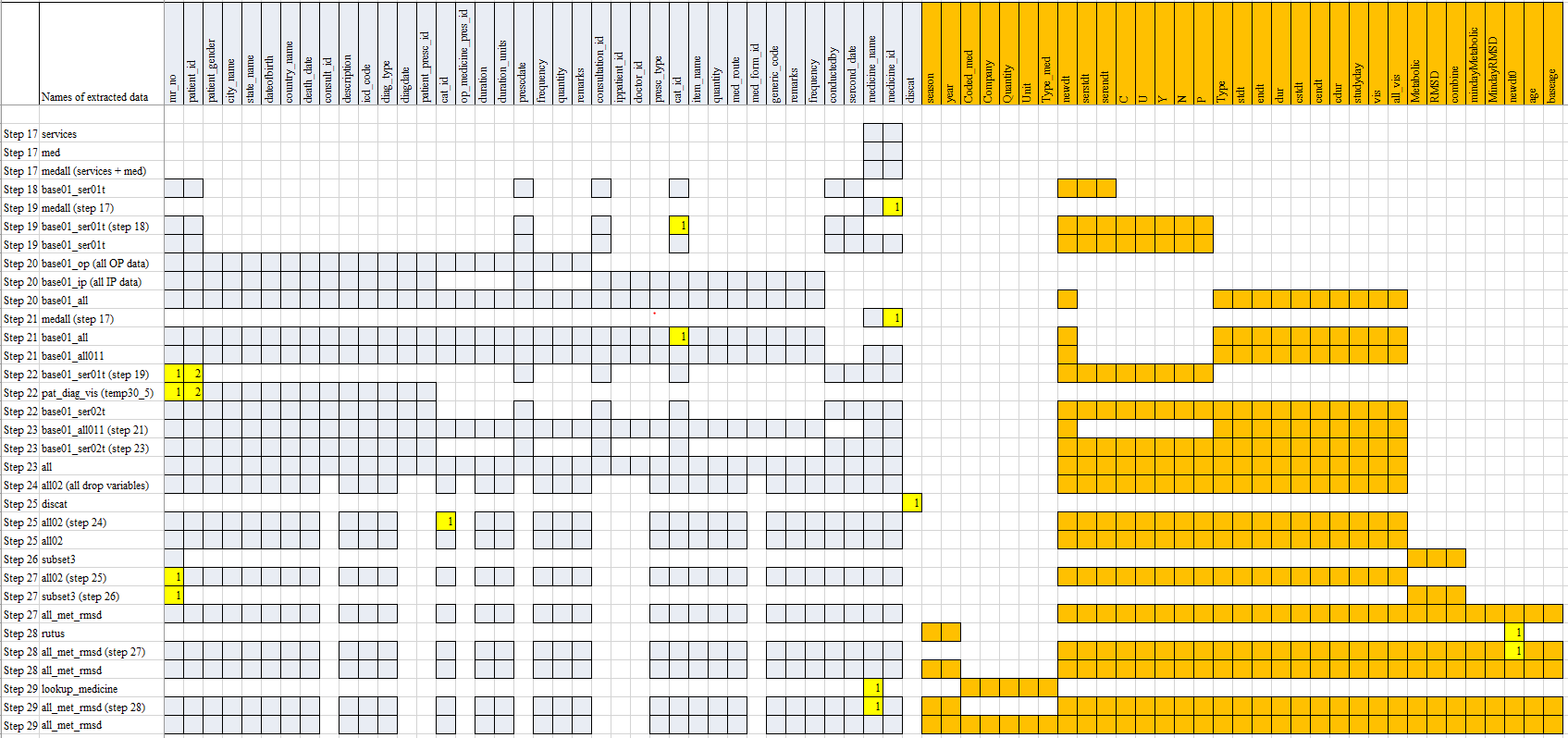
Picture option 3c:

Staged data (6 datasets, ~40 variables) + 3 reference files created based on inputs from experts



Picture option 3d:

1 Final dataset with ~30 source variables and ~30 new derived variables



Picture Option 4: (results section)

|  |  |  |  |
| --- | --- | --- | --- |
| Source data (SQL data file) | Staging data (csv files / R data files) | Data ware house (R data files) | Usage |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |  | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp  Longitudinal Patient data with disease, medication and Ayurvedic services information  ~30 variables from source  ~ 30 variables derived  ~50,000 patients  ~17,000+ patients: subsetted version for RMSD and Metabolic | Creation of additional analysis datasets  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |  |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | Actual analysis  Analytics - Free people icons |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | Learning from the existing database to be given back as learning  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp |  |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | Reference files for derivations and filtering of data | Clinical communication  C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\553C02C8.tmp |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp (txt file) |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp (txt file) |
| C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp | C:\Users\mahajvi1\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9805807.tmp (txt file) |

## Details of the reference dataset “01adsl\_met\_rmsd”

The subsequent table presents the details of the reference dataset

|  |  |  |
| --- | --- | --- |
| Variable name | Description | Derivation |
| mr\_no | Unique Patient ID | Source variable, no derivation needed  E.g. MR000001, MR040237, etc. |
| patient\_gender | Patient gender | Source variable, no derivation needed  E.g. M, F |
| patient\_id | Visit ID | Source variable, no derivation needed, the hospital database captures unique visit ID for each visit. |
| city\_name | City name | Source variable, no derivation needed |
| state\_name | State name | Source variable, no derivation needed |
| country\_name | Country name | Source variable, no derivation needed |
| dateofbirth | Date of birth | Source variable, no derivation needed, for some patients this is missing |
| newdt0 | Date of visit to hospital | Date of visit to hospital in numeric format  All the InPatient visits, OutPatient visits and Service related visits are combined from source datasets into dataset, unique visit and date combinations are created. |
| newdt | Date of visit to hospital | Character version of newdt0 |
| vis | Visit | 1. Based on all the InPatient visits, OutPatient visits and Service related visits unique visit numbers are created. 2. Visit numbers are numeric values from 1 to n, based on current version of data; a patient has maximum number 323 visits. |
| all\_vis | All visits | This variable contains maximum number of for each patient. all\_vis = max(vis) grouped by each mr\_no |
| all\_ip | All IP visits | This variable contains maximum number of for each patient for IP type of visits. all\_vis = max(vis) grouped by each mr\_no and visit type is IP. |
| all\_op | All OP visits | This variable contains maximum number of for each patient for OP type of visits. all\_vis = max(vis) grouped by each mr\_no and visit type is OP. |
| studyday | Study day | studyday = 1 when the visit minimum visit or first visit for a patient, else studyday is calculated as newdt0 – min(newdt0) + 1.  Studyday is never missing and never less than 0 for the dataset created. |
| age | Age of patient at that visit | If date of birth is non-missing for a patient, then age is calculated as round( (anydate(newdt) - anydate(dateofbirth) + 1)/365.25, digits = 0 ) |
| baseage | Age of patient at the first visit | Age at vis = 1 for each patient is stored as base age |
| death\_date | Date of death | Source variable, no derivation needed |
| cstdt | Min Start date | cstdt = min(newdt) |
| cendt | End date | cendt = max(newdt) |
| cdur | Total duration in days | cdur = max(newdt) - min(newdt) + 1 |
| stdt\_IP | Start date of IP visits | Minimum visit date for IP visits for each patient |
| endt\_IP | End date of IP visits | Maximum visit date for IP visits for each patient |
| dur\_IP | Duration of IP visits | dur\_IP = endt\_IP – stdt\_IP + 1 |
| stdt\_OP | Start date of OP visits | Minimum visit date for OP visits for each patient |
| endt\_OP | End date of OP visits | Maximum visit date for OP visits for each patient |
| dur\_OP | Duration of OP visits | dur\_OP = endt\_OP – stdt\_OP + 1 |
| serstdt | Service Start date | Minimum visit date for Service visits for each patient |
| serendt | Service End date | Maximum visit date for Service visits for each patient |
| Code | Code | Source variable, no derivation needed, ACD code |
| description | Description | Source variable, no derivation needed, description |
| Type | Type of visit | This variable identifies a visit either as IP or OP based on visit classification |
| diag\_type | Diagnosis type | Source variable, no derivation needed:  Primary or Secondary |
| year | Year | Year part of the newdt variable |
| season | Indian seasons | Derivation of Indian seasons based on the date variable for each visit:  # Add Indian rutus as new variables  # <https://www.drikpanchang.com/seasons/season-tropical-timings.html?geoname-id=1277333&year=2010>   * 01 VasantRutu * 02 GrishmaRutu * 03 Varsha Rutu * 04 Sharad Rutu * 05 Hemant Rutu * 06 ShishirRutu |
| C, N, P, U, X, Y | Values related to Services offered to patients | Source variable, no derivation needed:   * C- Cancelled * U - Condn. Unnecessary * Y -Conducted * N - Not Conducted * P - Partially Conducted |
| presc\_type |  | Source variable, no derivation needed |
| medicine\_name | Medicine name | Source variable, no derivation needed  Prescribed medicine names follow a certain predefined naming convention. Medicine name + Quantity + Producer’s name are the details recorded for each prescribed medicine. |
| item\_name | Source value of medicine name | Source variable, no derivation needed |
| quantity | Quantity of prescribed medicine | Source variable, no derivation needed |
| med\_route | Route of administration of prescribed medicine | Source variable, no derivation needed |
| generic\_code |  | Source variable, no derivation needed |
| remarks | Notes provided by doctors for medicines | Source variable, no derivation needed |
| frequency | Frequency of prescribed medicine | Source variable, no derivation needed |
| duration | Duration of prescribed medicine | Source variable, no derivation needed |
| duration\_units | Unit for duration of prescribed medicine | Source variable, no derivation needed |
| Coded\_med | Only name of medicine | Derived from medicine\_name |
| Company | Name of the company producing the drug | Derived from medicine\_name |
| Quantity | Quantity of prescribed medicine | Derived from medicine\_name |
| Unit | Unit of prescribed medicine | Derived from medicine\_name |
| Type\_med | Type of medicine | Derived based on medicine\_name. Classified into different kinds of medicines, e.g.   * Ghritam * Kashayam * Asavam * Aristham * Bhasma * Abhyanga * Cream * Rasayanam * Tablet / Gulika / Vati * … |
| cat\_id |  |  |
| distype | Disease type | Disease type as OTHER, RMSD, Metabolic   1. If a disease code is present in Metabolic list then the value is Metabolic 2. If a disease code is present in RMSD list then the value is RMSD 3. Any other disease is classified as OTHER |
| Metabolic | Metabolic | If a patient has reported any Metabolic disease at least once then that patient is given value Metabolic = 1, else Metabolic =0 |
| RMSD | RMSD | If a patient has reported any RMSD disease at least once then that patient is given value Metabolic = 1, else Metabolic =0 |
| combine | Metabolic  RMSD  Both | 1. If a patient is classified only as Metabolic diseased patient then combine = 1, 2. If a patient is classified only as RMSD diseased patient then combine = 2, 3. If a patient is classified as Metabolic as well as RMSD diseased patient then combine = 99 |
| Minday Metabolic | First day on which reported metabolic disease | First day on which any metabolic disease has been reported by a patient. |
| Minday RMSD | First day on which reported RMSD disease | First day on which any RMSD disease has been reported by a patient. |

## Other visit wise data

Along with the disease and treatment information, many variables get collected each visit. This data has been saved into a few additional tables, namely: (1) patient\_section\_details, (2) patient\_section\_values, and corresponding reference files are (3) section\_master, (4) section\_field\_options, (5) section\_field\_desc, (6) patient\_consultation\_field\_values.All these datasets put together provide us with 90+ different sections (tables) and 500+ variables of clinical datasets defined by the hospital. This requires a lot of re-arrangement of data, careful joining, transpose as well as filtering as appropriate.

The data can be stored either a long format or a wide format. The structure of 01adsl\_met\_rmsd is wide format structure. The data encountered here is of long format, each observation is one time point per patient per variable. Each patient will have data in multiple observations, with one column containing (section number), one column-containing visit information, date information, one column-containing context (for which variable the value is captured) and the last but not the least one column-containing the actual response to the variable.This long table is transposed into 90+ different datasets with appropriate variables populated in every dataset.Next table shows the names of the datasets:

|  |  |  |
| --- | --- | --- |
| History of Present Illness | Netra Case Sheet | Visit Postures |
| Review Of Systems | Vitals | Visit Kriyas |
| Physical Examination | Visual Acuity Test | Visit Pranayamas |
| Past / Family / Personal History/History of trauma/ injury | Glass Power Prescription | Visit Mudras |
| Nurse Assessment | Other Findings | Visit Bandhas |
| Gyn History | Nephrology Case sheet | Visit Relaxation Techniques |
| OBG History | Personal History | Visit Meditation Techniques |
| SampraptiGhatakas | Vihara | Doctor Advice |
| RogiPareeksha | Wellness assesment of child | Oral Cavity |
| Session | GROWTH & DEVELOPMENT ASSESSMENT | Ear |
| RogaVinischaya | Dosha Assessment Questionnaire | Nose |
| Systemic Examination | Diet Recommendations | Complaints And Associated Complaints |
| Investigations | Operation Notes | History of Previous Illness |
| Complaint | Conduction Notes | Medical History |
| IP Case Proforma | IP - History of Present Illness | Investigation. |
| General Examination1 | IP - Gyn History | Obstetrical History |
| Local Examination | IP - SampraptiGhatakas | Food Habit |
| Nidana | IP - NidanPanchak | Activity(Vihara) |
| DashavidhaPareeksha | IP - Session | Physical.Examination. |
| Nutitional Assessment | IP - RogaVinischaya | ASTA STHANA PARIKSHA |
| General Examination | IP - Systemic Examination | DASHA VIDHA PARIKSHA |
| Breathing Practices | IP - Investigations | Date |
| Loosening Practices | IP - Complaint | Treatment Plan |
| Postures | IP - Local Examination | Final Diagnosis |
| Kriyas | IP - DashavidhaPareeksha | Physiotherapy Treatment Plan |
| Pranayamas | IP - Nutitional Assessment | Physiotherapy Short term Goal |
| Mudras | IP - General Examination | Physiotherapy Long Term Goal |
| Bandhas | Visit Breathing Practices | Yoga Treatment Plan |
| Relaxation Techniques | Visit Loosening Practices | Physiotherapy Assessment |
| Meditation Techniques | Netra Case Sheet | Physiotherapy Outcome |

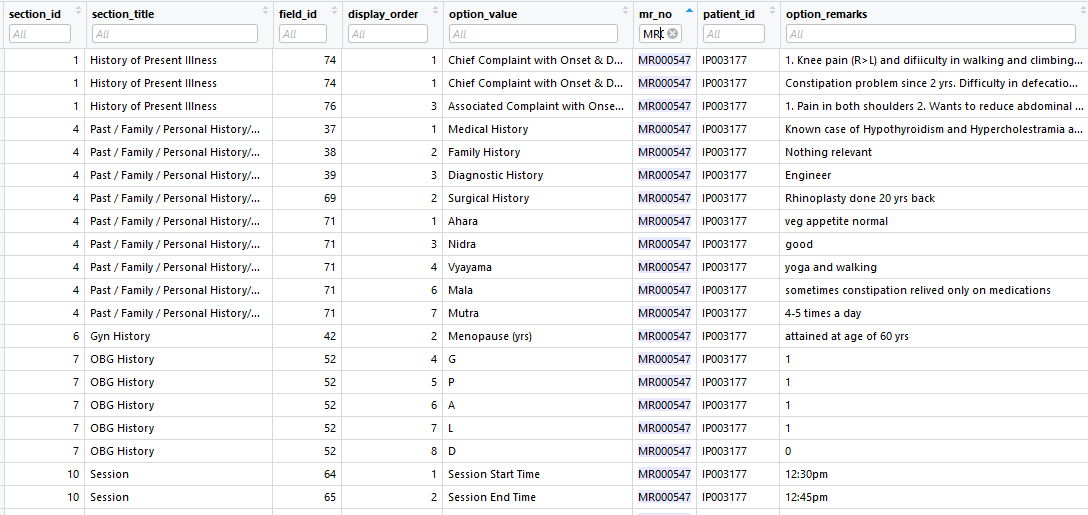
The long version of the data after transformations looks like the table below.

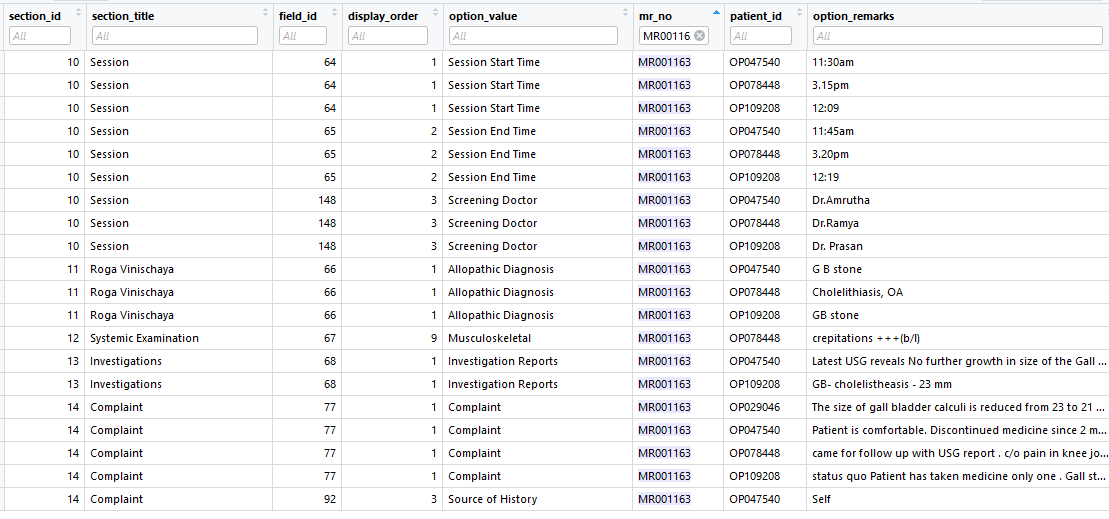
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| mr\_no | patient\_id | section\_id | field\_id | option\_remarks | section\_title | display\_order | option\_value |
| Unique Patient ID | Unique Visit ID | Case report form |  | Actual response to the variable  Data collected from patients at each visit | Case report form name | Order of the  variable on the screen | Question |
| CHI0000001 | OP167711 | 12 | 67 | Rhombergs - mild swaying, | Systemic Examination | 1 | Neurological |
| CHI0000002 | OP167712 | 12 | 67 | S1, S2 heard, | Systemic Examination | 4 | Cardiovascular |
| CHI0000003 | OP167713 | 12 | 67 | NAD | Systemic Examination | 6 | Gastrointestinal |
| CHI0000003 | OP167713 | 12 | 67 | NAD | Systemic Examination | 9 | Musculoskeletal |
| CHI0000004 | OP167715 | 12 | 67 | Direct Inginal hernia B/L, | Systemic Examination | 8 | Genital |
| CHI0000005 | OP167718 | 4 | 71 | Reduced | Past / Family / Personal History/History of trauma/ injury | 1 | Ahara |
| CHI0000005 | OP167718 | 4 | 71 | Disturbed | Past / Family / Personal History/History of trauma/ injury | 3 | Nidra |
| CHI0000005 | OP167718 | 4 | 71 | Occsional loose stools, | Past / Family / Personal History/History of trauma/ injury | 6 | Mala |
| CHI0000005 | OP167718 | 4 | 71 | Normal | Past / Family / Personal History/History of trauma/ injury | 7 | Mutra |
| CHI0000005 | OP167718 | 12 | 67 | NAD, | Systemic Examination | 5 | Respiratory |
| CHI0000006 | OP167719 | 1 | 73 | 25 yrs | History of Present Illness | 1 | Diabetes |
| CHI0000006 | OP167719 | 1 | 73 | 5-10 yurs | History of Present Illness | 2 | Hypertension |
| CHI0000007 | OP167720 | 12 | 67 | NAD | Systemic Examination | 5 | Respiratory |
| CHI0000007 | OP167720 | 12 | 67 | NAD | Systemic Examination | 6 | Gastrointestinal |
| CHI0000007 | OP167720 | 12 | 67 | B/L crepitus of knee L>R-++ive, | Systemic Examination | 9 | Musculoskeletal |
| CHI0000010 | OP167723 | 12 | 67 | Resonant ++ve | Systemic Examination | 6 | Gastrointestinal |
| CHI0000010 | OP167723 | 12 | 67 | Right knee- tenderness- Right supra patellar region, Ansirine bursa ++ive | Systemic Examination | 9 | Musculoskeletal |

Long format examples where column “section \_title” has different values.

* Patient ID: mr\_no
* Visit ID: patient\_id
* CRF name: Section\_title
* Variable name: option\_value
* Variable value: option\_remarks

The screen below represents data for a patient “MR000547” for an IP visit for 5 different CRF pages: History of Present illness, Past / Family / Personal History, Gyn History, OBG History, Session.





## Results and discussions:

It takes programming efforts as well as exploration to understand a database. In our case, we have seen approximately 50 stages to create the reference datasets from the pure Source database to get an analyzable structure. These datasets are not the only datasets created through the life of this study, similar efforts have been put in place to make the data “analysis ready”.

Learnings from the exercise:

1. Complete disease and treatment information for each patient is available in a structured database format is generated
2. Longitudinal picture of a patient’s disease can be drawn easily
3. InPatient and OutPatient information is collated at one place
4. Disease and treatment information for related diseases is present at one place
5. Time between 2 visits to the hospital for a patient can be calculated
6. Easy filtering for a disease, treatment is possible
7. Complicated subsets and creation of cohorts is possible
8. The naming of the datasets is quite logical but due to lack of documentation it was a puzzle to solve::
   1. First dataset covers: patient\_details
   2. After taking the details the patient is asked for: patient\_registration
   3. Next logical step is of: doctor\_consultation
   4. The treating doctor is able to diagnose the patient: mrd\_diagnosis
   5. Patients are prescribed medicines: patient\_prescription, patient\_medicine\_prescriptions
   6. If a patient is InPatient then the information is stored in: ip\_prescription
   7. Along with the medicines if there are services prescribed then they are in: services\_prescribed
9. Challenges:
   1. In general a database for patients should have the primary key as Patient ID: mr\_no (in our case), but the underlying database considers unique visit for each patient as a primary key between tables (Patient\_ID)
   2. In general, a variable containing same information across tables should have the same name, but in our case, each table has a different variable, making it difficult to create logical links across tables
   3. The case report form allows for multiple diseases and multiple treatments to be recorded for each patient, this causes a “clinical logic” challenge – the potential 1-1 relation between a disease and a treatment is lost, this has to be derived outside of the database using “costly” expert understanding
   4. There are multiple versions of the same table available in the database (as a programmer, it is well understood that older copies are retained in the system), but due to unavailability of the documentations, this aspect increases the complexity
10. An initiation of new role:
    1. We have seen the data generation by every day interactions and by now would have understood its potential usage. For this type of work, it is essential to have “database”, “IT”, “programming”, “clinical thinking”, “Statistical understanding”, and “connection of dots to create the big puzzle”. This expertise could be labelled as “statistical programmer”, “clinical analystand “clinical programmer” or now in our case “Ayur analyst”.
    2. This is not just an IT professional, not just a statistician, not just a data manager (cleaning the data), not a database developer to store the data, but someone who uses IT systems, uses a statistician’s mind, provides insights into building database, scientific rigor, and uses multiple technologies. Uses ideas to convert An observation to summaries to Stories and Clinical insights w/o being a doctor

|  |
| --- |
| Summary of the chapter:  * This chapter outlines the need basics of databases in brief. How are the data generated through Live system at hospital, where is the “raw or source data” stored, how is it moved to a “Staging area”, and how is it made “analysis ready” * The chapter explains a few technical details about the hospital database: how may source tables, how are they stored in ~200+ tables, out of which ~20 to 25 how tables are used to generate datasets useful for the analysis * The chapter has a few flowcharts outlining ~50+ steps to go from Live source –Staging data – transformed data - ~30+ source variables + ~30+ derived variables in 01adsl\_met\_rmsd dataset: patient level data covering Treatment and disease information * There are 90+ Case Report forms and 500+ variables in “Long format of data” covering many other ayurvedic and modern medicine parameters * This hospital has done a great job in assembling “patient level ayurvedic data” * The chapter covers the need of a “statistical programmer / analyst / clinical programmer” role to “mine the hidden gold from raw data” * **Is it ok to cover the shortcomings?** |

**Need to decide where to put the following section: “Chapter: covert the database to analyzable format”, “Chapter: understanding of the database”**

The process of data transcription:process of cleaning to confirm that the data were collected and evaluated according to the predefined assumptions and match the source data continues over the life of data.Different types of data transformed into analyzable data sets to address administrative questions, legal requirements, insurance compliance, regulatory compliance, and varied research questions. Based on the informed consent of a patient, the data is used to generate various publications and reports for different audiences

Health authorities and top level journals require the data along with research papers. The processes outlined in the above sections provide the necessary “fit for purpose” data.The analyzable data set, is the result of many decisions made by varied people, as explained above. If there are errors, flaws, or biases in the processing of raw data, such problems will not necessarily be identified in the analyzable data set. After data have been entered in computerized form, new variables are generated mathematically to serve as the basis for later analyses. The final cleaned analyzable data set consists of various components (participant characteristics and primary outcome, pre-specified secondary and tertiary outcomes, adverse event data, and exploratory data).

The full analyzable data set is generally the most useful set of data to share from a trial, with large and likely important benefits to science and society.

Analyzable data may increase the scientific return on the funder's investment in the trial and the benefits to the public and future patients.

Second, the full analyzable data set allows other researchers to reproduce the original analyses and carry out alternative, scientifically valid analyses of the primary study aim. Such additional analyses help determine how robust the original analyses are.

Third, meta-analytic syntheses of the results of similar trials increase the statistical power for detecting effects and maximize the evidentiary value of the clinical trial knowledge base.

Finally, analyzable data allows for further scientific discovery through additional secondary analyses, as well as the conduct of exploratory research to general hypotheses for additional studies.

Physicians and other scientists are good and getting better at producing data. But we must become proficient—with or without the help of technology—at mining and managing the data in ways that will allow us to use it to maximum effect.

The mission of health care institutions – restoring patient’s health – demands effective and efficient medical data for evidence-based intervention. Installing an appropriate health care data management system with valid case definition enables efficient data extraction, improves communication for clinical decision making in medical practice, and clinical research, and upgrades the quality of health care services. Healthcare professionals are responsive to improve recording, distributing, monitoring, and implementing preventive measures to decrease morbidity. This requires consistent, complete, comprehensive, and accurate information which attracts more attention in the health care industry.

The health care industry uses a paper-based record (PBR) and/or electronic health record (EHR) system to manage patient’s data. The EHR has become an integral part of medical care, which transforms health care service quality and improves clinicians’ satisfaction and facilitates patients’ decision. Accurate information from EHR enables physicians’ order entry and measures clinical validity, which in turn upgrades the quality of patient care. This functionality is crucial during diagnosis and therapy, which benefits medical and legal practices too.

Key to the challenge of being able to use the flood of information that is threatening to overwhelm us will be the development and use of “intelligent agent software,” programs that can automate commonly performed tasks and learn from their interactions with people. Such software could conceivably identify unrecognized opportunities to analyze data, solve problems, bring in interdisciplinary expertise, and integrate and prioritize diverse data sources in large, complex, and distributed information systems. To be truly useful, we would need the agents to know:

• What parts of particular sets of information are relevant to a specific individual and the current situation?

• Which medical references pertain to a specific patient’s condition?

• To which web sites a physician should refer a patient for relevant information

• How to recognize potential unexpected relationships between the diverse information sources?

## Challenges

Even in the 15th century, many expressed concerns about the problem of too much information. The 21st centrury scenario of flow of information has been increasing at almost exponential levels, and now it threatens to drown us in data. It poses a major challenge when evaluating its reliability, dependability, while obtaining useful information. Perhaps more significantly, this hurdle probably discourages a large number of people from working with data in the first place. The speed and efficiency of your data prep process directly affects the time it takes to discover insights. American healthcare is in danger of being overwhelmed by data. We undoubtedly require a plan to keep up with this information onslaught. Grasping new knowledge and collecting contemporary information requires more time, work force and money than most providers cannot afford. Right now, the data explosion threatens to destabilize American or rather world medicine.

Despite continued advances in data management technologies, it remains tedious to examine a newly acquired data set and ‘wrangle’ it into a form that allows meaningful analysis to begin.

Hardly do you have current, complete information available on transaction systems. Metadata dictionaries and repositories generally have very low accuracy. The poor attention paid to creating and maintaining accurate data in data dictionaries and metadata repositories is now hurting corporations to the tune of millions of dollars in unnecessarily complex data movement projects and/or in having to accept low-quality data (data that is even of lower quality than the source databases it is extracted from) for use in decision making.

The reasons for poor data quality in metadata repositories are many. Repository technology has lagged behind database design evolution. Repository solutions are generally insufficient in content to collect all the needed information. They are passive, which means that they can get out of step with the data without noticeable impacts on normal operations. There has been little motivation on the part of data administration staff to keep them current.

## Clinical data understanding

### Background

In the previous chapter, we have seen how the data has been collected and then converted into analyzable formats. Let us see of some of the question raised in the first chapter could be answered now through understanding the data contents.

In understanding the data, the first step is to understand what the individual tables convey on their own. Next step is to logically combine these tables to create a narrative or story which clearly defines the condition/situation of the patient in the database. This could include the demography, disease type and severity, underlying medical conditions, current medications if any and overall lifestyle of the patient, the first dose of medication, was there any interruption or medicine terminated, did the patient follow the dosing as well as visit schedule and how compliant was he/she; basis the data collected in these individual tables.

This data collected should be classified, explained, analyzed, explored and interpreted via the following ways:

1. How is the data collected:
   1. Is it collected via paper and then transferred electronically or is it direct electronic capture?
   2. Is the collected data continuous or discrete?
   3. Have proper checks been put in place during the electronic capture in order to avoid incorrect data entry. E.g. Character information entered for numeric data, alpha numeric or negative data or unrealistic data entry.
   4. Are appropriate drop-down menus, check boxes in place with minimal free text area in order to avoid both missing and incorrect data, thus avoiding quality problems
   5. Is data categorized? If yes, are the categories logical as well as optimal? E.g. 20 categories of hair color, invites variability in classification. Collecting data at a less granular structure, data reduction, than can be discerned in reality also invites variability and results in information loss.
   6. What information needs to be collected and how much needs to be collected should be clearly based on the need of the evaluation, too much data creates a lot of noise in the information thus blurring the actual picture as well as missing of vital information in an even to minimize data collection can be detrimental for accurate evaluation.
2. Is the common information across all the tables and visits accurately captured? The consistency should be not only in what is captured but how is it captured and stored within the tables for future usage. E.g. a patients demographic data like sex, race, height(unless a arthritis or osteoporosis patient), age(if an acute disease) should be same across visits.
3. How complete is the data? Is the start of medication date and the end of medication date accurately captured? Are there any missing dates for Concomitant medications or Adverse Events? Missing dates lead to loss of vital information during the evaluation of the patient’s status. Are the laboratory parameters captured with accuracy of level of detection/ quantification as well as in appropriate units?
4. What is the data that is collected? Is everything that was to be captured is adequately documented? This serves for future usage and analysis. Is the data collected consistent and complete and providing clinical insights and scientific validity when analyzed.
5. Completeness, consistency and clinical and scientific validity become the crux for any data captured, representing the population.

The pictorial representation below summarizes the cycle of understanding the hospital data so that a meaningful interpretation can be arrived at.

This provides an overview of looking at the same data in different ways and how each section provides an additional view to the patient data thus converting the collected information into a meaningful patient story.

Data understanding from an observation – patient – disease to a clinical picture

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 1. Understand data from variable and observation point of view | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 6. Translations of thoughts from mind to paper for future use | | |  |  | 2. Consistent data across case report forms across visits | | |
|  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | Data checks to understand the collected data | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 5. Completeness to provide overall clinical picture | | |  |  |  |  |  | 3. Consistency across patients (as appropriate) | | |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | 4. Consistency within the disease area / Therapeutic area | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

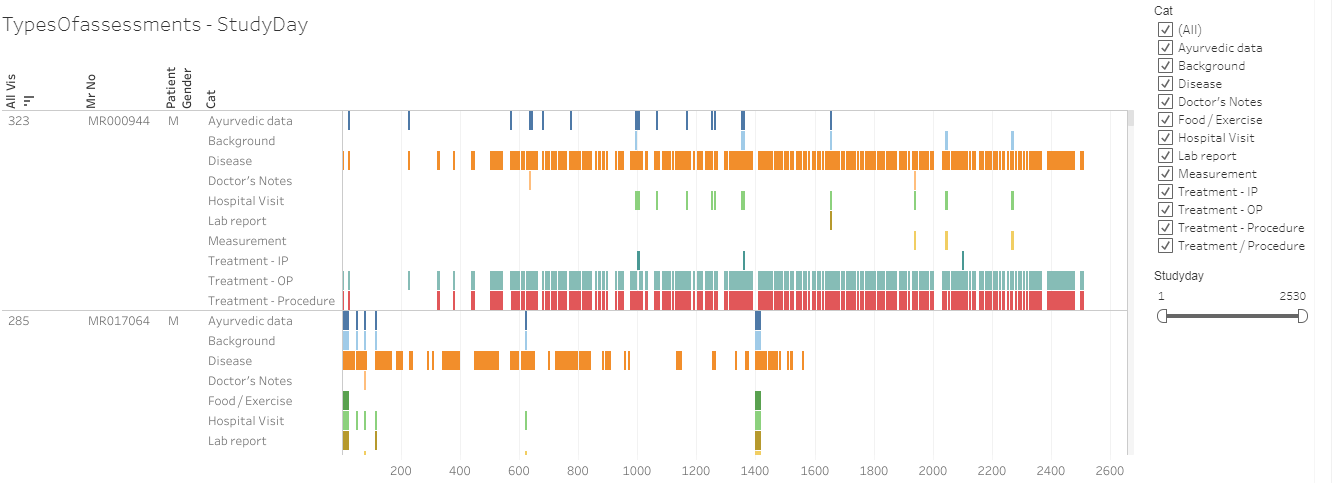
### Understanding the hospital data:

Data Classification

1. Ayurvedic data
2. Background
3. Disease
4. Doctor's Notes
5. Food / Exercise
6. Hospital Visit
7. Lab report
8. Measurement
9. Treatment / Procedure

This data helped in understanding the pattern of data collection. Below is an explanation on how data was analyzed to look for missing data and a snapshot of the same:

If there is any non-missing data present in a particular variable or a column then a pseudo value “Yes” is assigned, if the data is missing then a pseudo “Blank or No” value is assigned for the purpose of analysis. Tableau display is presented below as an illustration.



The analysis carried out shows that for majority of the patient and for majority of the visits, the disease data and medication (Treatment /Procedure) are entered and is not missing. Most of the other categories are not entered as consistently as they should have been. This lack of information should be addressed.

Understanding the data from a variable point of view

The following analysis provides information about distinct number of patients present in individual table as well as unique values present for various variables.

1. There are many CRF pages with very little number of patients.
2. There are many variables where captured data is in a “free text” format. These provide the doctors flexibility during data entry and the complete information gets captured.
3. This creates challenges in analysis. Same data value could be represented in different cases, minor spelling mistakes, human errors in data handling, etc. This causes lower data accuracy

E.g. Let us take a look at the following row in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sec001\_var008\_Diabetes | 1064 | 4124 |  |  | Background |

1. sec001\_var008\_Diabetes: presents CRF page 1 “History of Present Illness” and variable number 8 “Diabetes”.
2. This variable has 1064 unique values entered by different doctors for different patients.
3. There is some entered data for 4124 distinct patients and
4. This variable is categorized as variable capturing data for category “Background”.

Another example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sec001\_var018\_Associated Complaint with Onset & Duration | 5102 | 4549 |  |  | Disease |

1. sec001\_var018\_Associated Complaint with Onset & Duration: presents CRF page 1 “History of Present Illness” and variable number 8 “Associated Complaint with Onset & Duration”.
2. This variable has 5102 different values entered by different doctors for different patients.
3. There is some entered data for 4549 distinct patients and
4. This variable is categorized as variable capturing data for category “Disease”.

This data provides the following areas of improvements:

1. Standardization of CRF pages for better data capture – there are some standards already available in healthcare industry, they could be implemented at the hospital.
   1. CDISC standards
   2. ISO standards
2. Implementation of the standard dictionaries to capture data uniformly and make analysis ready, some of these could be directly implemented; some could be modified for our requirements.
   1. LOINC standards
   2. NCI USA Code lists
   3. MedDRA dictionary
   4. ICD Codes
3. There are some variables present in more than 1 CRF page, this duplication should be removed.
4. After consulting experts in specific fields make some variables mandatory with appropriate values, pre-defined drop down lists.
5. Data quality as well as amount of data for Ayurvedic parameters should be improved to recreate patient information.

These steps should help in improving the data quality and data consistency.

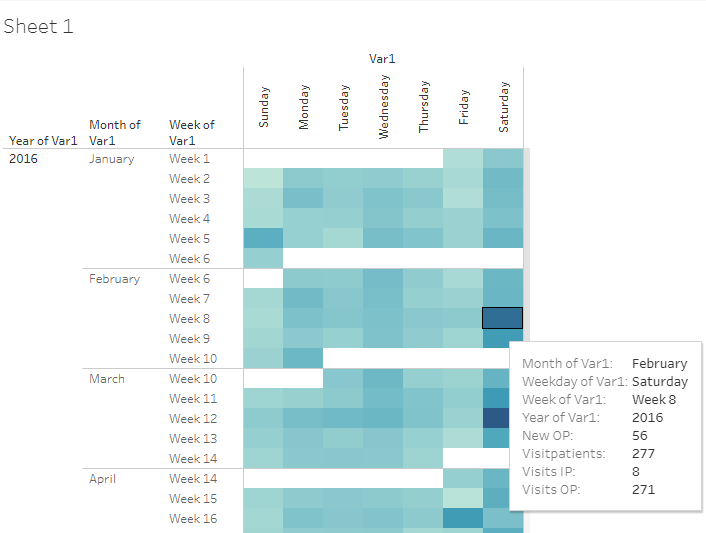
Visit Pattern summary; consistent data across case report forms across visits

The following analysis provides information about visit patterns from year 2011 to 2016.

* Unique combinations of individual visits for each patient are created for each day for each year.
* Columns in the display are: (1) Year, (2) Month, (3) Week of month, and (4) Days of a week from Sunday to Saturday.
* Each of the resulting cell is colored in shades of blue from light blue to dark blue based on the frequency count of number of patients
* Each cell has additional information included in the tooltip feature for the following summaries:
  + New Out Patients added on that day
  + Total number of patients visiting on that day
  + Total number of Inpatient visits on that day
  + Total number of Outpatient visits on that day

Interpretation:

* The number of patients visiting hospital on weekdays is less than the number of patients visiting on weekends. This information would help in employing staff across different Line Functions to adequately cover services for patients.
* If the hospital conducts special events then the number of patients goes up.
* Inpatients are considerably less than Outpatients which is quite understanble.

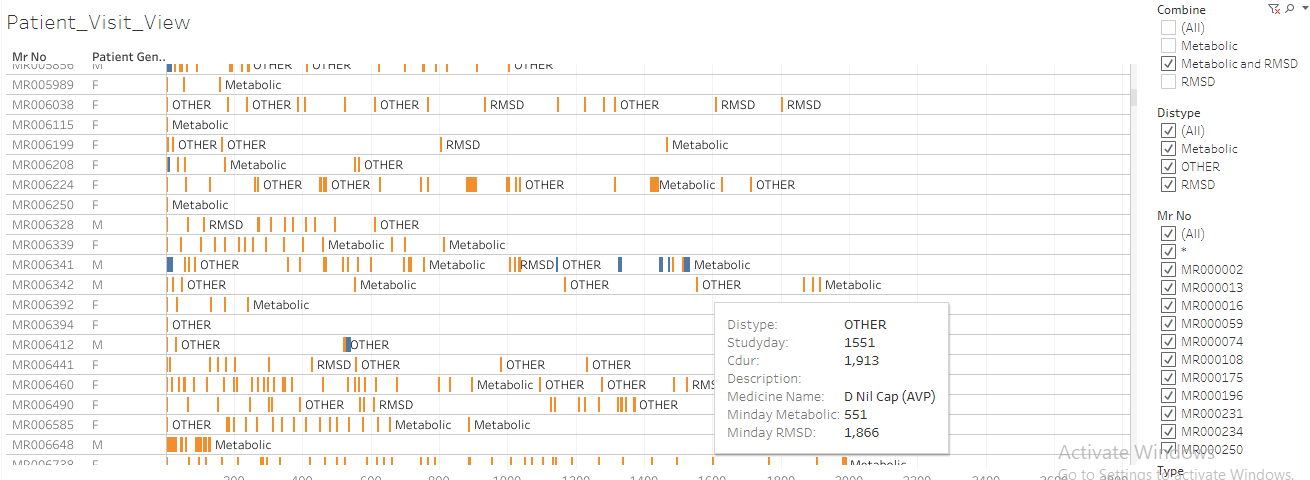


Patient visit view; completeness to provide overall patient view

There are 2 analysis views created for individual patient data. These show the longitudinal nature of data.

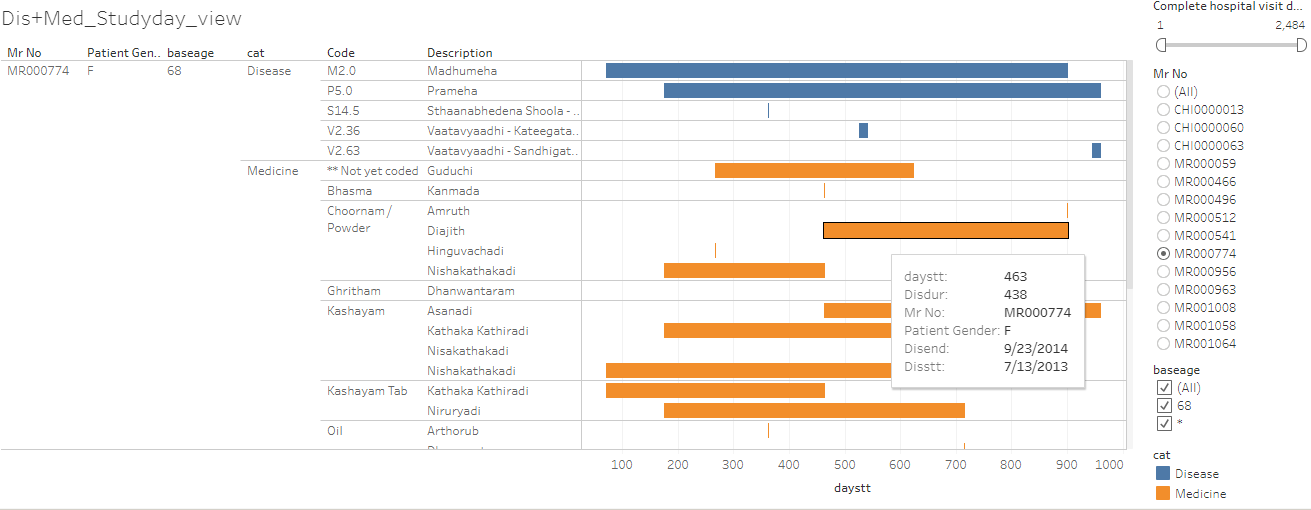
Columns displayed in the first view are as follows: Patient ID (mr\_no), gender, study day

1. The Inpatient visits are displayed in blue color and Outpatient visits are displayed in Orange color.
2. The tooltip contains information about the following data points not displayed on the page:
   1. Studyday: Study day
   2. Cdur: Total duration of hospital visits.
   3. Description: Disease description variable accompanying ACD codes.
   4. Medicine Name: Medicine provided at that visit.
   5. Minday Metabolic: First day on which any metabolic disease has been reported by patient.
   6. Minday RMSD: First day on which any RMSD disease has been reported by patient.



Columns displayed in the second view are follows: Patient ID (mr\_no), gender, base age, category, Code, description, study day

1. The Diseases are displayed in blue colored bars and treatments prescribed are marked in orange colored bars.
2. Bars are created as followed: Duration between minimum and maximum reported date for a disease as well as prescribed treatment is calculated, this duration is displayed on the visualization. This calculation needs improvements (algorithm is getting developed to derive individual episodes of diseases as well as individual episodes of treatment assignment)
3. The tooltip contains information about the following data points not displayed on the page:
   1. Daystt: Start of event in days (disease as well as treatment)
   2. Disdur: Duration of event in days (disease as well as treatment)
   3. Disstt: Start date of event
   4. Diend: End date of event



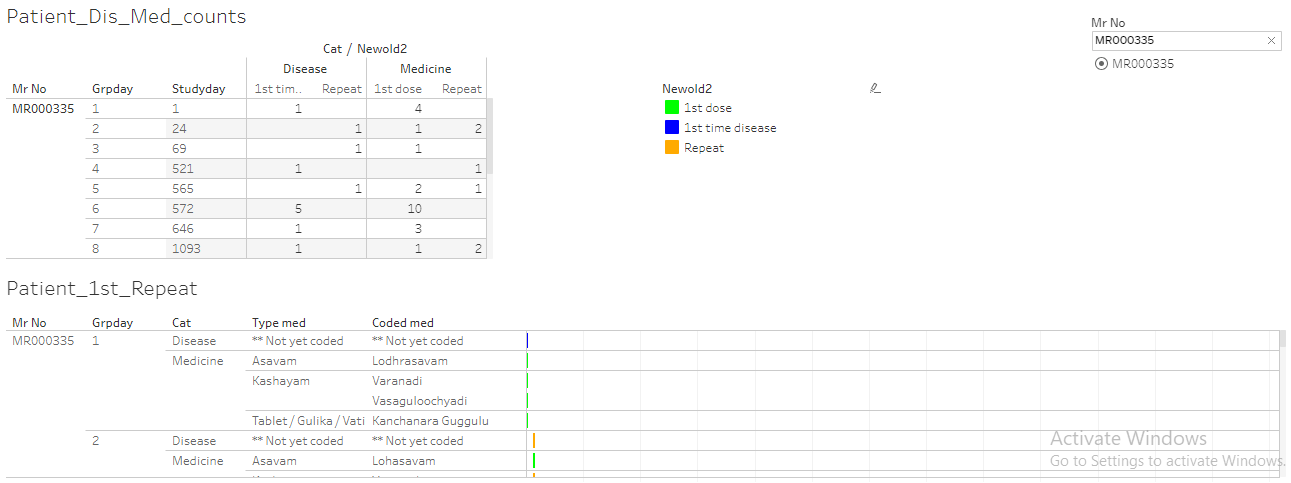
Disease and treatment view, listing and summary; Consistency within the disease area / Therapeutic area

There are 3 analysis views created for individual patient data.

1st part of the analysis:

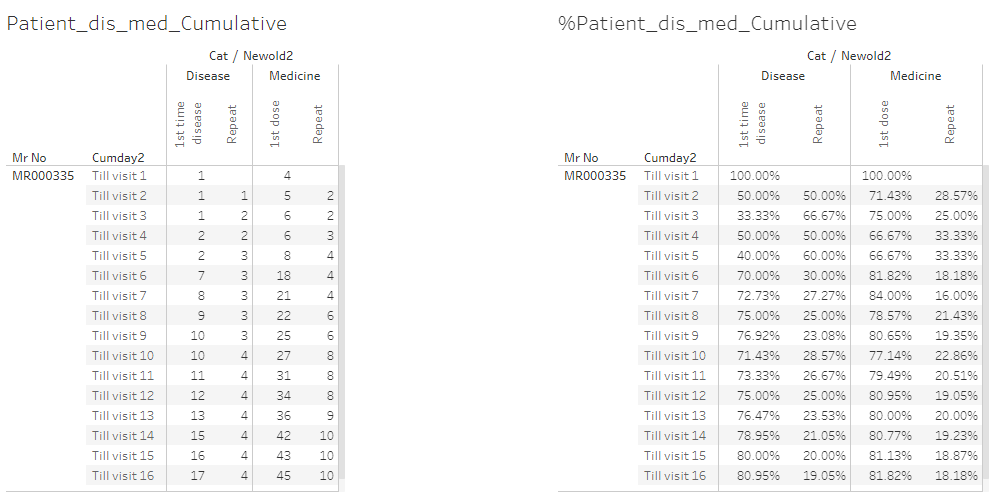
Patients are treated as they come to hospital. This visual provides a patient level view of number of diseases reported for the first time and then repeated, similarly treatment prescribed for the first time vs. a repeat of treatment.

1. When a disease is reported very first time then that is considered “1st time disease reported”, any subsequent repetition is considered as “Repeat”.
2. When a treatment is prescribed very first time then that is considered “1st time treatment prescribed”, any subsequent repetition is considered as “Repeat”.
3. These 2 calculations are repeated through the data for each patient.



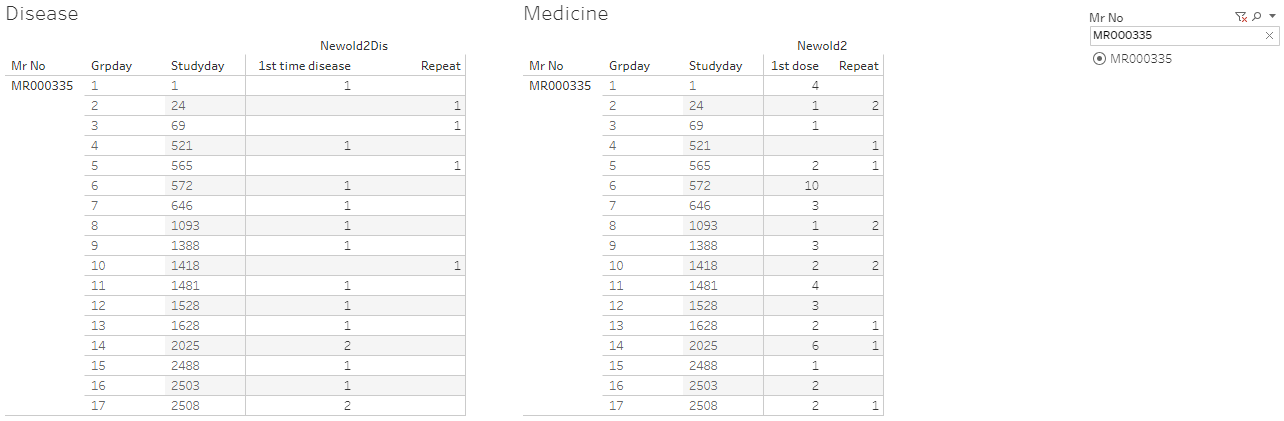
2nd part of the analysis:

This is a cumulative view for an individual patient. This provides a summary of what would have happened to a patient till a certain visit number. There are 2 tables created, first with absolute numbers and second with percentages.



3rd part of analysis: This is another version of display of diseases and treatments for individual patients “non-overlapping or non-cumulative” version.

1. Each line is a patient visit. 1st disease, Repeat disease, 1st treatment and Repeat treatment columns are displayed.
2. Studyday column shows the visit day.



Interpretation:

1. When a new disease is reported, usually a new treatment or treatments are reported, if there is only a new treatment added then it could indicate, the earlier treatment may not have worked, or it explains the treatment regimen.
2. If only new diseases are added and no new treatment is added then the same treatment could work for multiple diseases.
3. The above example shows for patient MR000335,
   1. There are 17 visits
   2. There are 17 distinct diseases reported and 45 distinct treatments, services prescribed.
   3. 4 out of 17 diseases are repeated and 10 out of 45 treatments have been repeated.
   4. Based on the 3rd view, at visit 6 on day 572, there is only 1 new disease reported and 10 new treatments are added. At visit 14 on day 2025, there are 2 new diseases reported and 6 new treatments added.
4. These visualizations allow the treating doctor insights into newer diseases getting reported as well as what newer treatments have been prescribed at what time points.

Completeness to provide overall clinical picture; converting thoughts to data

1. Unique patient ID is assigned to individual patient. Each visit to the hospital is also uniquely captured. A visit is classified into 2 types (1) in-patient visit where a patient is hospitalized for certain duration, (2) out-patient visit where a patient visits hospital, meets a designated doctor and goes back home. The Unique Patient ID and visit number are used as “key variables” to join information from various tables.
2. Disease conditions are coded using ACD dictionary. This dictionary was developed at C-DAC Pune and covers approximately 4500 disease conditions. At the time of analysis, the database contained approximately 900 disease conditions.
3. There are a few records where the Coded disease value is missing, it will be coded as “\*\* Not yet coded”
4. Information for the diseases and medicines from different tables are merged into 1 large table based on Patient ID (mr\_no) + Visit ID (Patient ID), this provides a side by side view of Disease conditions and prescribed medicines for the same.

The data captured and curated in a standardized manner allows the physician to understand the complete patient profile thereby providing the complete and consistent information of the patient and disease condition. Along with it, it also provides clear information of the drugs prescribed and the compliance information. This information can be easily passed on from one physician to another in case of an emergency or review, thus benefiting the patients.

Similarly, a standard database or a standardized data mart containing multiple patient information, along with their diseases and co-morbidities and all the diagnostic details provides a wealth of data to the researchers. This data can be utilized for future drug development or prescription of optimal treatment regime. Further this data also provides the researchers a view of demographic and climatic influences on diseases, treatments and recoveries.

Along with the doctors, patients and researchers this standardized data is also of great importance to the Government, if shared with them managing all the data privacy norms, as it would help them devise healthcare policies which are benefiting the actual demography. This would create a good competition amongst the insurance companies thus benefiting the patients by providing them appropriately design policies.

Regular updates to these databases can add a wealth of data which can support and aid the clinical research thus enabling continuous improvement in patient lifestyle management, medical practice, diagnostic tools, policy frameworks and as a larger goal help the government and medical institutions in understanding the teething problems, root causes for certain diseases, lapses in compliances, rural and urban complexities if any and many more.

The kind of data we store and how we store it is the first step towards achieving larger goals. Similarly, studying the demographic data and standardizing the data capturing methodology, clarifying too many versus too few and thus optimizing the number of variables stored would be the next step.

Data understanding from an observation – patient – disease to a clinical picture

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 1. Understand data from variable and observation point of view | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 6. Translations of thoughts from mind to paper for future use | | |  |  | 2. Consistent data across case report forms across visits | | |
|  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | Data checks to understand the collected data | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 5. Completeness to provide overall clinical picture | | |  |  |  |  |  | 3. Consistency across patients (as appropriate) | | |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | 4. Consistency within the disease area / Therapeutic area | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Usage for continuous improvement

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 1. Improve the knowledge of people using the database for data entry | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 6. Solid databased evidence for many stake holders | | |  |  | 2. Regular check on the data entry by an independent person | | |
|  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | Data checks to understand the collected data | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 5. Clinically meaningful Longitudinal view from “an observation” to “summary” | | |  |  |  |  |  | 3. Is the data getting entered in similar fashion across patients (as appropriate) | | |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | 4. Is a disease area / Therapeutic area looking consistent from data capture point of view | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

### Methods and Material

Data quality dimensions:

(1) Accuracy of the data: Values in the data match the intended values in the real data

(2) Completeness: the extent to which the value matches the real value and provides full meaning

(3) Consistency: data values representing the true state, for a variable, across variables, visits and across patients

(4) Timeliness: how much time it took from the actual visit to get the data entered into the system

(5) Current state: is the value most up to date or would the value require any modification due to clinical judgement, common sense, statistical plausibility, etc.

(6) Granularity: level of detail intended and captured in the data

(7) Specificity: how unambiguous the data value is

(8) Precision: how precise it is, numerical value, clinical meaning

(9) Attributable: is it source generated by doctor, lab report, scans, derived based on multiple visits and elements

These steps are described at a general level so that they can be applied to any project. From the data-oriented point of view, the steps include: (1) identifying data to be collected, (2) defining data elements, (3) observing and measuring values, (4) recording those observations and measurements, (5) processing data to render them in electronic form and prepare them for analysis, and (6) analyzing data.

A general assumption is that the more structured the data, the higher the degree of accuracy and ease of processing.

### Results

fafaf

### Discussion

afafaf

<http://people.dbmi.columbia.edu/~chw7007/papers/chapter%2010.pdf>