1 Introduction

1.1 Origins of Pharmaceutical Industry

The origins of the pharmaceutical industry go back to the apothecaries and pharmacies that gave traditional therapies going back to the Middle Ages. 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 [8]. In year 1947, one of the first international standards was established due to the ethical and medical misconduct of the Nazis in World War II. The Nuremberg Code created 10 standards that include voluntary informed consent, risk to subject must be weighed against benefits, actions that injure human subjects must be avoided, and the subject has the right to end the experiment at any time [9]. Unfortunate events like deaths due to diethylene glycol poisoning in the US in 1930s, the thalidomide disaster in late 1950s, catalyzed the development of medicines regulation. By the year 1964 World Medical Assembly developed the Declaration of Helsinki. It became a key international document describing the ethical principles that underlie GCP. These principles provide guidance in medical research involving human subjects. The document has been amended on a regular basis, with its most recent update done in October 2013, in Brazil [10]. The World Health Organization (WHO) and the United Nations Educational, Scientific and Cultural Organization (UNESCO) in 1949 jointly established The Council for International Organizations of Medical Sciences (CIOMS) [11]. The use of statistics as a scientific field to support R&D of new medicines grew multifold since the Kefauver-Harris Amendments (1962) and continues to grow [12]. 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 proof of safety will not be sufficient for new drug approvals as had been required in the initial regulatory requirements. It issued a document titled "Proposed International Guidelines for Biomedical Research Involving Human Subjects" to help apply the principles of the Declaration of Helsinki and the Nuremberg Code, in 1982. These guidelines were revised in 1992, which resulted in the "International Ethical Guidelines for Biomedical Research Involving Human Subjects." Due to the increasing global footprint of clinical research, the International Conference on Harmonisation (ICH) was established in 1990 [13]. In 2016, the name was changed to International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH created guidelines based on 4 pillars of quality, safety, efficacy, and regulatory obligations to protect public health. These guidelines get updated on a regular basis [13]. These efforts have shown that, in the USA and all over world, since 1970s, the value of medicine has been clearly demonstrated. Human population have exhibited a longer life expectancy, a lower infant mortality rate, and the higher quality of life.

1.2 Elaboration of Clinical Trials: Origin of RCT Blinded Trials

Randomized Controlled Trial (RCT) is a classical research design of randomly allocating participants to one or the other treatments under investigation. Randomization is the fundamental characteristic of an RCT, and it describes the random distribution of participants to the study arms. RCTs aim towards supporting a conclusion that the difference in the outcomes among participants

in study arms was exclusively caused by the intervention, as randomization equalizes the study group in all other factors. Thus, RCTs set the standard of excellence in health sciences research.

Although, randomization reduces the bias of participant assignment to the intervention and control group, it does not rule out the chance of bias, for example, due to investigator/caregiver or the patients themselves or bias generated during subjective adjudication of an outcome variable. However, RCTs do help in reliable description of causality between the intervention and outcome.

Blinding as a procedure, assists in monitoring and controlling several types of biases that might unintentionally creep into the study. Two major biases, namely performance bias and ascertainment bias can be controlled using blinding. Blinding deals with four groups of people associated with a trial: the study participants, the investigator(s), the outcome assessor(s), and the data analyst(s). Depending upon the groups of people blinded, trials are categorized as open label, single blinded, double blinded, triple blinded, and quadruple blinded trials [14].

Physicians and clinical researchers have remained confident that RCTs deliver the most rigorous test of preventive, diagnostic, and therapeutic interventions. They are universally denoted as the "gold standard" of experimental medical analysis, as an undisputable starting point in preventive, diagnostic or therapeutic evaluation. An article written by Alvan Feinstein and Ralph Horwitz in The New England Journal of Medicine (NEJM) in December 1982, is the first instance of RCTs being referred to as "gold standard" [15]. 1982 appears to be late date for the first usage of the "gold standard" terminology. The authors say that they were eager to be proven wrong on this 1982 date. But despite massive searches across many textbooks, conference materials, journals, and archival collections have not thrown up any earlier date [15].

While probing details regarding RCTs on the ClinTrial.gov website the following data was seen on 24th April 2020 [16], Link. Total trials in the ClinTrial.gov database was 3,36,905. The estimated % of RCTs ranged between 3% and 20%. Remaining 80% to 97% of the registered trials are not RCTs (Figure 1-1). 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 work is carried outside of the RCT framework.

Figure 1-1: RCTs on ClinTrials.gov

Terms Synonyms	Search Results*	Entire Database**
Randomized control trial	1,862 studies	1,862 studies
trial	56,745 studies	117,642 studies
Clinical Trials	21,302 studies	42,784 studies
CLINTRIAL	17 studies	67 studies
control	50,668 studies	188,230 studies
Comparator	37,863 studies	123,329 studies
Controlled	31,962 studies	66,966 studies
Control Group	7,483 studies	24,005 studies
controlling	390 studies	1,772 studies
Randomized	56,945 studies	178,454 studies
Randomization	6,723 studies	12,452 studies
RANDOM	1,267 studies	3,526 studies
randomized clinical trial	6,234 studies	6,234 studies
clinical randomized trials	35 studies	35 studies
clinical trial	56,945 studies	118,708 studies
trial	56,743 studies	117,627 studies
Intervention Study	551 studies	1,599 studies
CLINTRIAL	17 studies	67 studies
clinical	32,693 studies	138,751 studies

This is a screenshot taken from the ClinTrial.gov website showing number of Randomized Clinical Trials. Estimates of RCT range between 3% considering "Randomized control trial (1,862)" and "randomized clinical trial (6,234)" and 20%: considering "Randomized control trial (1,862)", "Randomized (56,945)" and "randomized clinical trial (6,234)"

1.3 Modern Hospitals, Everyday Clinical Practice and Healthcare Environment

Health information provided by hospitals is vital for public health administrators. Such data, accumulated from all the relevant hospitals, are essential in formulating health policy planning for the district / state / country, by matching this data with population statistics and other demographic data, and the health resources of the district / state / country. This medical information, containing not only data about the patient's illness but also the success or failure of therapy, could be a precious source of clinical training for medical students, nurses, and related health care work force. This could also be basis of new medicines and therapies. The development of new therapeutic drugs is built upon careful observations of experienced physicians, and the comparison and analyses of administered drugs, gathered from the patient's data. This in turn leads to continuous medical progress in uncovering new entity of a disease or a new method of diagnosis [17].

One trend all modern hospitals have in common is that the amount of processed information generated per patient is constantly increasing e.g., Electronic Medical Record (EMR), patient charts, CT scans, X-rays, ECGs, pathology reports, wearable devices, sensor data, etc. When such

information is properly compiled and aggregated, it could provide data for efficient hospital administration. How each patient is treated can yield important medical data, while the number of patients handled and treated, grouped by sex, age, and diagnosis, provides basic information that could be useful for administrators [18].

This information is processed manually or partially automatically in many instances. 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 might be needed.

1.4 Real world evidence and observational studies

Real-world data are data captured in an observational manner, in a natural setting. Ordinary clinical practice, producing a never-ending flow of results from everyday practice, can be viewed as infinite sequence of unsystematic observational studies. There must be rational use of this outcome data from patients collected from everyday clinical practice as it is far more representative of the general population and medical practice than those data coming from formal clinical trials [19]. Practice-based medicine is an important way to advance science [20]. For this to happen, the treating doctor should be a "clinician researcher investigator". Doctors even in their busy practice, can see patterns, then formulate a research hypothesis, and then proceed to test it. Hence, a doctor can be good at good clinical practice (GCP) and good clinical research practice (GCRP). One can support the other and vice versa, and this way medicine advances [21].

Investigation of literature across time points from year 2000 to year 2020 shows that well designed observational studies produced treatment effects comparable to RCTs:

Year 2000: 99 reports across 5 clinical topics covering observational studies provided comparable results like those of the randomized, controlled trials. There was no systematic over estimation of treatment effects seen [22].

Year 2007: RCTs provides one kind of knowledge but could prevent us from understanding other properties of a drug. When observational epidemiological studies are carried out correctly, they can be both more conceptually complex and more powerful than an average RCT, especially in assessing drug safety. Both kinds of research if done with rigor and with scientific humility, can be supplementary / complementary to each other [23].

Year 2014: There was little evidence found for significant effect estimate differences between observational studies and RCTs regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmacological interventions. The results highlight that level of heterogeneity in meta-analysis of RCTs or observational studies is an important factor [24].

Year 2020: Cumulative 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 show the settings in which RWD analyses can robustly and consistently match the results of RCTs and the situations in which they cannot match. After careful consideration of the potential for bias, regulators can then determine when they would accept RWD in place of an RCT [25].

The data from RWE studies and Observational studies have a wealth of information which if processed accurately and summarized in a structured manner can lead us to numerous assertions and confirmations. There have been many attempts to generate clinical evidence from primary health care by systematic utilization of patient records. But this has not been easily possible due to the deficient clinical data, in-accurate input "garbage in" leading to "garbage out", insufficient follow-up, and very few fully completed case records with risk factors, co-morbidities, etc.

Both clinical practice and research based on the principles of primary care could contribute substantially as follows:

Clinical practice (1) uses medical knowledge and supports confirmation of research findings, (2) focuses on individual patients but still generates data, (3) has a short action span and immediate reward, (4) it regards authority, follows custom, earns income, and encourages research. Clinical research complimentarily (1) creates knowledge by focusing on future clinical practice, (2) focuses on groups of patients to generate data, (3) has a long action span and delayed reward of discovery, (4) questions authority, challenges custom, earns reputation, and enriches practice.

1.5 EHR across the globe

In early 2000s countries like Canada, UK, New Zealand, Estonia, etc. started collecting clinical practice data at a national level. The key challenges experienced while implementing Electronic Medical Records (EMR) faced by them were: infrastructure creation, policy & regulations, standards & interoperability, and research, development & education [26].

Resource shortages, amplified by various socio-economic problems pose substantial difficulties for development of workable healthcare solutions which are global in nature. Healthcare development is one of most important aspects for the progress of both social and economic development of the world. More than $1/4^{th}$ of the world's population has unmet healthcare needs. On World Health Day 2018, WHO introduced the concept of Universal Health Coverage (UHC): healthcare for everyone and everywhere. However, this has not been completely effective due to the financial limitations within systems and factors like aged-population-related chronic diseases influence the availability, accessibility, and quality of care [27].

However, 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 interventions along with awareness of unintended consequences and thus supporting the clinical practice decisions and aiding to improve the healthcare process or clinical outcomes for the patients.

1.6 Role of Statistics, Analyst, Programmer

Statistics emerged as an extended stream from mathematics, operational research, and economics. Application of statistics in various fields of research like genomics, epidemiology, nutrition, biological science, biomedical research connoted the word "biostatistics". Apart from biological sciences, statistics is applied in variety of other fields like market research, insurance, trades and stocks, banking etc. The 1990s presented explosive increase in applications of computationally

demanding methods. These methods were naturally based on statistical principles, due to the emphasis on exploratory nature of the problem and the importance of data. Statistical modeling has become more complex due to the volume of data, computational requirements, and varied data sources. These developments led to creation of new roles like statistical programmer - statistical analyst, clinical programmer - clinical analyst [28]. Clinical data analysts (or clinical informatics analyst) are healthcare professionals responsible for confirming the validity of scientific experiments and data gathered. They apply their knowledge of data acquisition, data management, data analysis, and data interpretation to healthcare data, providing actionable insights that doctors, clinical scientists, and others can use. They may be responsible for automating internal and external reports, creating executive-level dashboards, and presenting information to help various stakeholders understand the operational impact of the data. Data analysts ensure that processes and protocols associated with clinical research are followed, thus improving overall care [29], [30], [31]. They provide data insights that drive clinical process improvement, such as reducing readmissions and hospital-acquired conditions [32]. Clinical data analysts generate and provide the results of clinical business intelligence to management and all stakeholders. They coordinate with other researchers (e.g., clinical strategy, clinical operations) to determine the questions to be answered as well as the appropriate measures that should be taken to ensure data analysis proves to be useful. These roles combine strengths from multiple areas to build a powerful storyline on six dimensions of quality care: safe, effective, patient centered, timely, efficient, and equitable [33].

1.7 Traditional Chinese Medicine history and philosophy

Traditional Chinese Medicine (TCM) has been in practice for more 3000 thousand years [34] through Shang Dynasty (1600-1046 BC), Han Dynasty (206 BC–220 AD), and Zhou dynasty (1100-221 BC) of China [35]. Classic of Changes (Yi Jing) and Classic of poetry (Shi Jing) are considered the oldest medical writings available for TCM [34]. Subsequent foundation of TCM are based on the four classics across different era are Inner Canon of the Yellow Emperor (Huang Di Nei Jing, ~26 BCE), Yellow Emperor's Canon of Eighty-One Difficult Issues (Nan Jing, ~106 CE), Treatise on Cold Damage Disorders (Shang Han Lun, ~206 CE), and Shennong's Materia Medica (Shen Nong Ben Cao Jing, ~220 CE) [34]. Acupuncture, herbal medicine, diet, movement, and manual therapy are considered the five main branches of TCM [36]. TCM is based on the Five Elements theory, Wood, Fire, Earth, Metal, and Water which describes the human body. A blend of Chinese philosophy, culture, ritual, and medical practices, TCM and herbal medicine demand a broad understanding which is not limited to science [37]. Zheng (syndrome), a basic unit in TCM, decides the therapeutic methods. The process of how to get the outcome is called differentiation of Zheng, which is based on the physiology and pathology of TCM helps in determining the course of treatment [36]. This narrative practice, either in the form of classics or passed down orally, retains the medical knowledge of ancient Chinese people. Until the 20th century, most Chinese could only use TCM as a medical service. After the modern medicine breakthroughs TCM has been going re-assessment efforts based on modern scientific methods [34], [38], [39], [40], [41], [42]. One such example is of Artemisinin and its derivative dihydroartemisinin which was referred to in a medical classic by Ge Hong (284–346 CE). This drug was tested by Tu and her colleagues

using modern methods. It has saved millions of lives threatened by malaria throughout the world [43], [44]. For this phenomenal work, Tu was awarded the 2011 Lasker Award for clinical research and the 2015 Nobel Prize in Physiology or Medicine [45], [46].

Researchers have found use of "case record" in advancing traditional knowledge through the history of TCM in various phases of maturity. First phase defined by the creation and development of the discipline of TCM medical records; second phase covers the standardization of writing structure of TCM medical records; third phase outlined by books written by veteran TCM doctors on recording and studying TCM medical records. Most recent stages of development account for the explosion of TCM case reports published in journals; the establishment of TCM medical records databases and application platforms integrating computer programs and artificial intelligence; and the last but not the least, many reporting guidelines have been developed in order to improve the reporting quality of case report in TCM [35].

A short review of the available databases:

Post year 2000, many researchers have developed many web-based freely available databases covering topics like molecular properties, substructures, TCM ingredients, and TCM classification, based on intended drug actions. TCM Database@Taiwan database contains more than 20,000 pure compounds isolated from 453 TCM ingredients [47]. SymMap is a symptom mapping database integrating TCM with modern medicine in common aspects at both the phenotypic and molecular levels [48]. TCMSP: traditional Chinese medicine systems pharmacology database and analysis platform. It consists of all the 499 Chinese herbs registered in the Chinese pharmacopoeia with 29,384 ingredients, 3,311 targets and 837 associated diseases. It contains tools for visualization and analysis of TCM results on the network level for chemists, biologists, and pharmacologists [49]. TCMIO: Traditional Chinese Medicine on Immuno-Oncology database covering the mechanisms of TCM in cancer immunity and TCM-inspired identification of novel drug leads for cancer immunotherapy [50]. TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis [51]. YaTCM: Yet another Traditional Chinese Medicine Database for Drug Discovery [52].

A short review of the clinical trials and registries for TCM:

There are not many registries for TCM yet, but some are getting developed. The China Stroke Registry for Patients with Traditional Chinese Medicine (CASES-TCM) study is a prospective, multicenter, observational disease registry aiming to register 20,000 hospitalized patients [53]. The China Amyotrophic Lateral Sclerosis Registry of Patients with Traditional Chinese Medicine (CARE-TCM) provides an opportunity to better understand which TCM interventions patients with ALS are receiving, what the characteristics of patients with ALS are, and how these interventions impact clinical measures. This registry was initiated in March 2021. This study includes a voluntary nationwide registry. Detailed data collection will be performed every 3 months for 5 years. Baseline characteristics and 5-year survival will be collected [54].

A review of TCM trials on Clintrials.gov till Sept 2015 shows 1270 TCM trials listed on the website. Most of these trials were carried out either in China or in the USA. More than 50% of the

trials were for acupuncture and 35% for herbal medicines. 55.7% that were studies with less than 100 enrolled subjects. 8.7% of completed studies had reported results of trials. Even though there are issues, it is good to see these studies getting registered on Clintrials.gov [55]. Text data extraction using deep learning models has been tried out in TCM on literature, but online search did not filter up many such examples on patient data [56]. My search could not filter Real world evidence studies within TCM area. Since early 2000s there have been some efforts put in at the national level on outcomes research and big data. In year 2010 officially RWE concept was rolled out for TCM, and 2016 onwards this has been spread across the country. Chinese Drug administration has started using RWE in pre-approval process for Supporting evidence for investigational new drug. Chinese national authority NMPA has issued a guidance for "Clinical development of traditional Chinese medicine hospital preparations" under the overall guidance for "Key Considerations in Using Real-World Evidence to Support Drug Development" Center for Drug Evaluation, NMPA May, 2019 [57], [58].

This data shows that within TCM area there is an opportunity to carry out work based on naturally reported patient level data.

1.8 The Indian context

India has a mixed system of healthcare consisting of government hospitals, private hospitals, family doctors and private medical practices. We see this trend reflected in the actual health seeking behavior 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 [59]. India, having one of the largest rural populations in the world, it becomes imperative to be both inclusive and optimal in the use of available resources from any stream of healthcare. Along with global healthcare challenges, Indian healthcare faces specific issues like inadequate healthcare resources, insufficient funding, poor healthcare infrastructure and rural–urban disparity. The national healthcare budget is a lowly single digit percentage of the Gross Domestic Product (GDP). The number of doctors, nurses and health workers are small in comparison to high population in certain regions. The availability and utilization of technology is also at the lower end of the usage [60].

Integrating modern medicine and AYUSH systems, acronym for Ayurveda, Yoga and Naturopathy, Unani, Siddha and Homeopathy, folk medicines as well as technological advances into the India healthcare system can generate new opportunities for healthcare. These efforts will be noteworthy where healthcare systems have limitations in terms of infrastructure, expertise, and human resources. This integration will not only allow utilization of the positives of each of the medical streams but will generate new data-based evidence due to integration possible using technology. Ongoing monitoring and updating patient's data together with consistent updates to physicians can help in identifying disease and treatment patterns. Through this analysis, authorities or the government can employ specific precautions, or even start necessary facilities to avoid healthcare issues before the situation gets out of hand [61].

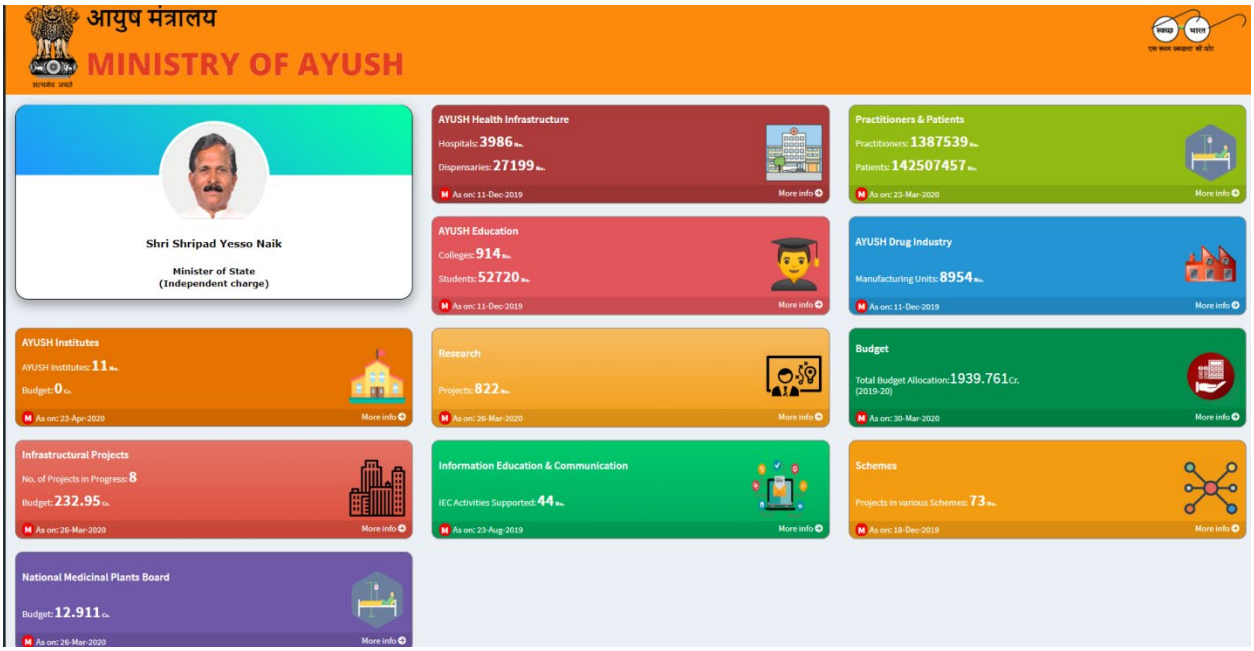
The application of such assessed and analyzed data are substantial as the information brings out more and more knowledge about an area or a person. In under-resourced countries such as India,

the resultant benefits are expanded further, suffering is reduced, hospital resources are saved and socio-economic improvements that lift a nation's wellbeing are recognized. Moreover, India could achieve improved healthcare delivery, care audit, epidemiological study and quick response to epidemics and bring economic benefits to individuals by reducing the healthcare cost.

1.9 National level efforts AYUSH

This section covers efforts put in by the Government of India at the National level. The Ministry of AYUSH is a part of Government of India to promote and expand use of AYUSH systems of health care and medicine in India. 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 [3]. One of AYUSH affiliates, CCRAS has developed National AYUSH Morbidities and Standardized Terminologies E-Portal NAMASTE 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 [62]. To publicize the merits of AYUSH systems across the globe, a web-based portal for Research publications was launched in 2011, which is being maintained by NIIMH Hyderabad [63]. Ayurveda Hospital Management Information System (A-HMIS) is a complete IT platform for all functions of health care delivery systems and patient care in AYUSH centers. THERAN (THE Research Application Nexus): HMIS is developed by Siddha Central Research Institute, Chennai under Central Council for Research in Siddha. These are 2 additional examples of ongoing work under Ministry of AYUSH [64]. Ministry of AYUSH has created AYUSH GRID to cohesively implement projects under Digital India Movement. It is an amalgamation of IT projects planned for advancement of AYUSH pan India [65]. The dashboard available on AYUSH homepage (accessed on 25th May 2020), reveals the following facts: the screenshot shows almost 10 crore patients treated at some or the other point (Figure 1-2, Figure 1-3). 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" [66].

Figure 1-2: Screenshot of a dashboard from AYUSH website



Dashboard from AYUSH website covering: AYUSH institutes, Infrastructure projects, National medicinal plants board, Health infrastructure, Education and communication, Research, Practitioners and patients, Drug Industry, Budget, and Schemes

Figure 1-3: Ayurvedic practitioners and patients from AYUSH dashboard

Practitioners & Patients

Practitioners

#	Facility	Ayurveda	Unani	Siddha	Yoga	Naturopathy	Homoeopathy	Sowa-Rigpa	Total
1	Registered Practitioner(IQ & NIQ)	443704	51110	9125	0	2485	293455	0	799879
2	Institutional Qualified Registered Practitioner	294162	38672	5685	0	2349	246792	0	587660

Patients

#	System	IPD	OPD			Total	
		CHC	CHC	Dispensaries	PHCs	IPD	OPD
1	Ayurveda	1390950	16557395	71486719	8455545	1390950	96499659
2	Unani	102884	2220231	8109909	1504284	102884	11834424
3	Siddha	210016	10617115	1634022	16858888	210016	29110025
4	Yoga	15741	1564861	50150	1282064	15741	2897075
5	Naturopathy	18636	138187	210634	72176	18636	420997
6	Sowa Rigpa	0	0	7050	0	0	7050

Close

Screenshot from AYUSH website covering: Approximately 4.5 lakhs Ayurvedic practitioners and more than 10 crore treated patients. Data regarding different medical systems within AYUSH has been presented. Ayurvedic practitioners and patients treated by ayurvedic interventions are considerably more than by any other traditional medical practice.

1.10 Science of ayurveda

There are some 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 in his talk at TDU, huge observational data for Ayurvedic medicines covers large number of books and manuscripts, 57 authentic books (Drug and cosmetic act 1940, page 47) [67], more than 4500 diseases including subtypes and conditions (Ayusoft database) [68], more than 81,000 formulations (TKDL database) [69], more than 4,00,000 Practitioners in India [66], infinite documents, references, experiential data, living tradition and knowledge in public domain [7]. Dravyaguna (Pharmacology), Bhaisajya Kalpana (Pharmaceutics), Nidana (Diagnosis) and Chikitsa (Management principles). This data points to a validated knowledge base.

Dr. D. B. Vaidya has explained the concept of reverse pharmacology to understand the action mechanism of Ayurvedic intervention. He says that there is huge amount of observational data available, showing relatively low side effects that have been reported. This existing data should be used as basis to carry out large interventional Ayurvedic trials to assess safety, efficacy, and pharmacokinetic information. This approach will be economical and could be less time consuming compared to the sequential drug development or the hierarchical model used in western medicine. He further talks about integrating meticulously documented experiential and experimental observations [70].

Prof R. H. Singh has opined that lab-based research experiments within Ayurvedic area during the last 50 years have not been rewarding. On the other hand, literary experiments to make a few of the classical Ayurvedic texts accessible to masses have been extremely useful. This situation warrants newer strategies of scientific research without compromising on the fundamental principles of Ayurveda [4].

Prof. Bhushan Patwardhan writes that there are substantial similarities between the traditional systems like Ayurveda and modern medicines. Ayurveda emphasizes on health promotion, disease prevention, early diagnosis, and personalized treatment. The modern medicine system approach uses predictive, preventive, and personalized medicine (PPPM). In case of Ayurveda, the evidence can be drawn from two main sources: (1) Evidence based on historical and classical nature of clinical practice supported by credible and accepted documentation. (2) Evidence based on ongoing scientific research to support various theories, medicines and procedures used in Ayurvedic medicine [71] [72].

Dr. Ram Manohar has expressed that Ayurveda is based on 5000 years of clinical practice. Hence, practice-based clinical trials should complement natural ways to gain insights [73].

Dr. Baghel's interpretation is that one should think of Ayurveda being in the developmental phase like any other medical systems. Like many other scholars he thinks that Ayurveda is a pure science based on logical explanation, which is called *Darshana*. Ongoing research in Ayurveda should impact academics, pharmacy, and practice in a profound way to convert data into information, information into knowledge and knowledge into wisdom [74].

As per Prof. Darshan Shankar's analysis as of 2015, at the national level, Ayurveda receives a meagre 3% of the Central health Budget and at the State level, making it difficult to fund any meaningful research projects. Despite Ayurveda's strengths, it has some limitations in current scenario. To advance the science there is a need to embrace tools of information technology to organize its vast multifaceted data, in searchable formats. Meticulously documented clinical experiences interpreted through Ayurveda-biology will expand rejuvenation of healthcare in India [59].

All the thought leaders cited here point to the strengths of Ayurveda as well as the immediate needs. They have pointed out that the research must be 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 to find new ways of tackling problems [75].

1.11 Potential opportunities for Real World Data analysis within Ayurveda

The western medicines are developed using a method called as hierarchical method where it tries answering 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 including open-label randomized studies but in general the clinical trial testing usually concludes with a blinded, randomized controlled trial (RCT). As already pointed out, RCTs are the "gold standard" of evidence generation as they offer most internal validity and minimal bias [76]. These studies could be complemented by using case studies, case series. This "one step at a time" approach has worked very well in the western medicine framework.

Ayurveda has been practiced for more than a millennium and is widely accepted in India as a worthy medical system. Over the last few decades' people across the world have gained knowledge and realized the importance of the age-old medical system and are constantly driven towards it. Although, having been in practice for ages Ayurveda still does not enjoy the recognition which the Western medicine does. Hence there needs to be a structured approach towards making this possible. The untapped potential of Ayurveda needs to be scientifically communicated globally for a wider reach for it to be utilized as a public health tool for promotion of health and prevention of diseases [76].

Ayurvedic vaidya usually use paper-based case report to record a patient's Ayurvedic parameters along with other details of medical consultation. These are typically not exchanged with other vaidyas. There is a huge amount of data available on paper and if digitized could be a big revolutionary step. Increased use and interoperability with electronic medical records of digital Ayurvedic patient management systems are required. Based on a report published by AYUSH [66], there are 4.5 lakh registered Ayurvedic practitioners. Even if 5% of doctors start using EMRs, i.e., 22,500 doctors and if data for 2 new patients is entered every day (~225 working days) for the whole year, 50 lakh unique patients' data can be generated in a single year. Currently, this gold mine of data has not been built yet.

1.12 What this study aims to contribute to

This study aims to contribute to interests of multiple stakeholders involved in clinical research. Based on arguments given above about availability of observational data, availability of technology, the Indian context and the specific case about Ayurveda and need to understand the science behind it, the study would highlight multiple use cases. The study would highlight tools and information generated through these tools and this could be interchangeably used by interested stakeholders including:

1.12.1 Hospital management

To keep any hospital functioning smoothly, operational insights from routine hospital data are important to improve management and efficiency of day-to-day activities. How many patients are present in the database? What are the characteristics of these patients? What is the gender distribution and age group distribution? Which countries, states, cities do they come from? How many times do they visit the hospital? What is the number of In-Patients & Out-Patients? What kind of assessments are done at each visit? What is the duration of visits for a patient? Which diseases are getting treated? Are the patients benefiting? How do you measure benefit?

Regular analysis of data would give insights which will allow for operational proficiencies, cost savings and eventually profitability. Finally, patients' satisfaction would improve if a hospital functions efficiently.

1.12.2 Clinicians or treating doctors

Let us understand what kind of benefits the clinicians can have from this study. Traditionally many clinicians in their own practice use paper CRF to capture patient data. The hospital has electronic data capture system, which is essentially doing the same job, but data capture is electronically done. Are they ready to adopt to a new way of working? Does onboarding training at hospital cover this part of the job? Do they see this as an additional burden or an integral part of day-to-day work? How can experience of clinicians using the Health Information system be viewed? Do doctors like data entry part of job? Do the doctors have time for real data entry while consulting patients?

Answers to some of these questions can be generated indirectly by understanding the quality of the database and data contents. Does hospital management take any feedback from the clinicians who are the primary "end users" of the data capture system? This timely feedback loop should provide great inputs into evolution of the system both operationally as well as scientifically.

The electronic data capture provides unique opportunities to clinicians such as they have access to the data from other practicing clinicians. This gives indirect learning opportunities of understanding treatment protocols, treatment variations employed, rare diseases treated at the hospital. Retrospective analysis of disease and treatment should provide ideas about disease variations, appropriateness of documentation, disease – disease combinations, disease – treatment combinations. These documented combinations could be clinically meaningful, could be season wise, gender wise, age wise varying. Retrospective analysis of treatments should provide tendencies of treatment prescription such as use of classical treatments, herbo mineral treatments.

Data review and analysis tools developed for this thesis can enhance patient and doctor interactions.

1.12.3 Universities and students

Teaching material for students: Over the years, teaching methodology has not transformed even though there are quite a lot of advances in modern methodology. Can learning objectives of different kinds be tackled by making insights generated from this data available to the students. For example, can complex disease-treatment relationships be made easily visually available. Can interesting ways of explaining text and advice from Ayurveda which have contemporary relevance be created? For example, occurrence of certain diseases in certain geographic areas or season. Very few large-scale studies like this have taken place in Ayurveda, so this study with large amounts of observational data can provide exploratory opportunities to describe findings, textually as well as diagrammatically. Tools developed here can be used as a supplementary material for any MD / PhD student.

Scientific literature generation by researchers: Most hospitals in India or any part of the world mainly focus on treatment and not on research publications. Can a research team be put together for medical communication who publish papers as their primary job? If there is no known profile of patients visiting an Ayurvedic hospital and if this data can be generated and represented in the right form, it will provide novel information. How can we measure the strengths and weaknesses of an Ayurvedic practice? How should researchers evaluate changes in results of a practice over time? Is it possible to build new hypothesis? Is prescribed treatment truly personalized? Is it possible to trace back the treatment regimen followed and compare it to classical fundamentals? Is there a way to compare the demographics and patient characteristics from a Ayurvedic hospital against a mainstream western hospital? Which are rare diseases identified in the database? Can a clinically meaningful document be written, like a case series about this rare disease?

How can anyone use these data as “secondary use”? Can this data be used by insurance companies? Do the approved labels of medicines and prescriptions in the database match each other? Metal based formulations are questioned by non Ayurvedic community, what insights can be drawn about the rasa-aushadhis? Which are these medicines? For what diseases are they given and for what duration? Before providing the metal-based treatment and after providing the metal-based treatment, is there any difference in duration seen in treatments? What is the percentage of patients that are prescribed these medicines and what is the percentage of duration of all the duration of treatment given to these patients?

1.12.4 Policy makers – AYUSH and relevant ministries, insurance sector:

Policy makers and insurance companies can use insights from this data to decide on which treatments to cover for insurance. Government agencies can use this data to promote Ayurveda as a medical system throughout the world. They can make policies similar to policies that govern western medicine.

1.13 Introduction to real life data

As presented above, the potential benefits of gathering and analyzing such data are huge. A quick history of how this ability to collect data has come about is given below. This history also

highlights the challenges of collecting and analyzing data in general on top of challenges associated with clinical research and Ayurveda specifically. With the passage of time, revolutions in technology have continually increased the creation of information and its exchange. With the advancement for communication from spoken to written, it became simpler to create texts, books thereby documentation; thus aiding transfer of knowledge from one person to another as well as from generation to generation without losing any data in translation. With increased and improved writing, compilation of articles, tables and records, there came a time where storing them became important, thus came in the 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 [77].

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 issued in Dec 2018 that the world's cumulative data will grow from 33 zettabytes 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 growth of data has been seen in every industry, in every corner of the world. The relentless increase in the quantity and flood of information denotes an important professional opportunity, but a challenge simultaneously for those in medicine and science [78].

As indicated by Toby Cosgrove MD, from Cleveland, medical information doubles every 73 days, in the year 2020, as compared to approximately 3.5 years in 2010. An estimated 8,00,000 papers were published in 5,600 medical journals every year. It is projected that 12,000 new articles and 300 randomized controlled trials will be added to Medline each week, and that new medical articles will appear at a rate of one every 26 seconds [79]. 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 the data. Data is being produced, obtained, and stored in numerous number of structures [80].

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, ECGs, questionnaire data, or self-reported data (e.g., symptoms, quality of life).

Raw data, unprocessed source data, like unrefined gold buried deep in a mine is a precious resource. It is often: (1) Inconsistent, containing both relevant and irrelevant data, (2) Imprecise, containing incorrectly entered information or missing values, (3) Repetitive, containing duplicate data. To utilize the raw data to its fullest potential, it needs to be extracted, filtered through, understood, and transformed into analyzable format. 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. 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" [81] [82] [83] [84].

1.14 Introduction to clinical data understanding

The health care industry uses either a paper-based record keeping method and/or electronic health record (EHR) system to manage patient data. More and more organizations are using electronic data capture, but the practicing doctors in individual clinics may still be documenting observations on paper. The EHR has become an integral part of medical care, which transforms health care service quality and improves physicians' satisfaction and facilitates patients' decision. Accurate information from EHR enables physicians' decision making 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 [80].

Health authorities and top-level journals require the data to be submitted along with research papers. The analyzable data set, is the result of many decisions made by varied people, as explained above. The errors, flaws, or biases in the processing of source data, will not necessarily be identified in the analyzable dataset. After the electronic data entry, new variables are generated to support further analysis. The final cleaned analyzable datasets consist of various components such as participant characteristics and primary outcome, pre-specified secondary and tertiary outcomes, adverse event data and exploratory data [85].

Physicians and other scientists are 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 [86]. The full analyzable dataset is generally the most useful set of data to share, with large and likely important benefits to science and society. Secondly, the full analyzable dataset provides scientific validity to the outcome and ensures replication and repeatability. Further, meta-analysis increases the statistical power of detecting effects and maximizes the value of the outcome in the clinical knowledge base. Finally, analyzable data allows for further scientific discovery through additional secondary analyses, as well as the conduct of exploratory research to generate hypotheses for additional studies [87].

1.15 Introduction to study of demographics and patient characteristics

We will proceed with understanding the data contents and see if any of the questions raised earlier can be solved. Demography is the study of the population. It explains the composition, the distribution and the data trends seen in the population. Roles and functions for demography studies can be broadly defined as, (1) population projections, (2) inputs into government budget, (3) evidence-based policy, and (4) communication of vital statistics [88]. There is very little data on

the profile of patients accessing traditional systems of medicine. A comparative study of profile of patients using an Ayurveda clinic and modern medicinal clinic will help in understanding of utilization of services and preference for health seeking behaviour.

1.16 Introduction to study of diagnostics and interventions

Diagnosis is one of the most important aspects in the process of treatment of the disease or condition. It is a patient-centered, cyclic process of gathering information, analyzing information, determining the health condition, and defining the type of intervention and continuously monitoring the progress till the desired state of functions/doshas is arrived at. Ayurvedic treatment involves removal of the causative factors. It assists in getting the functions/doshas into balance. The success of a treatment is possible only by timely and accurate diagnosis, a tailor-made intervention accompanied by an effective collaboration of the physician and the patient [89]. The term comorbidity refers to the coexistence of multiple diseases in relation to a primary disease in a patient. Patients report multiple diseases during their visits to the hospital. Some of these reported disorders are expected and some are unexpected. There are known as well as unknown disease combinations present due to biological linkages. Clinical and epidemiological studies indicate that disease comorbidities have a great impact on health status, selection of appropriate treatments and health system costs. Understanding comorbidities and their etiology is key to identify new preventive and therapeutic strategies.

Finally, all these steps (1) accessing and understanding real life data, (2) converting that data into analyzable format, (3) understanding demographics and patient characteristics, and (4) understanding diagnostics and interventions should help in building transdisciplinary evidence to increase the scientific understanding outside of the community, then increase the confidence and thereby widening the user base.

1.17 Structure of the thesis document

The thesis is structured as follows: Chapter 2 covers study design and numerous methods applied for data analysis. In section 2.7, table (Table *2-2*) presents a consolidated view of various analysis, context about different analysis and their possible relationships with each other. Detailed observations of these experiments are documented in chapter 3 Results. Challenges with respect to overall system architecture and data collection are documented in chapter 3. Chapter 4 provides some solutions to these challenges. Supplementary interpretations are further illustrated in chapter 4. Chapter 5 covers several conclusions drawn based on the analysis and key take away message. Chapter 6 appendix provides detailed material generated for this PhD work. Appendix section 6.2 provides details about the master dataset developed for whole of this study. Appendix section 6.3 provides information regarding the source database. Appendix section 6.4 provides a table to track all analysis presented in the thesis. Appendix section 6.5 provides R, SQL, and D3js programs developed to generate different datasets and analysis. Chapter 7 covers the references.

All the tables and figures are cross referenced and clicking on the table number or figure number in the document takes the reader to that table or figure. Column “Figure number and analysis name (linked to the actual figure)” in Table *2-2*: Proposed methods and analysis use cases allows a reader

to go the actual link of specific analysis. “Link to analysis” part under any figure or table allows a reader to go to actual link of specific analysis.