Ophthatome™: An integrated knowledgebase of ophthalmic diseases for translating vision research into the clinic

Praveen Raj  
MedGenome Labs Ltd

Sushma Tejwani  
Narayana Nethralaya

Dandayudhapani Sudha  
MedGenome Labs Ltd

Muthu Narayanan B  
MedGenome Labs Ltd

Chandrasekar Thangapandi  
MedGenome Labs Ltd

Sankar Das  
Narayana Nethralaya

Somasekar J  
MedGenome Labs Ltd

Susmithasane Mangalapudi  
MedGenome Labs Ltd

Durgesh Kumar  
MedGenome Labs Ltd

Narendra Pindippanahalli  
Narayana Nethralaya

Rohit Shetty  
Narayana Nethralaya

Arkashubra Ghosh  
Narayana Nethralaya

Govindasamy Kumaramanickavel  
Narayana Nethralaya

Amitabha Chaudhuri  
MedGenome Inc

Soumittra Nagaswamy (✉ soumittra.n@medgenome.com)  
MedGenome Labs  https://orcid.org/0000-0002-0401-9601
Abstract

Abstract Background: Medical big data analytics has revolutionized the human healthcare system by introducing processes that facilitate rationale clinical decision making, predictive or prognostic modelling of the disease progression and management, disease surveillance, overall impact on public health and research. Although, the electronic medical records (EMR) system is the digital storehouse of rich medical data of a large patient cohort collected over many years, the data lack sufficient structure to be of clinical value for applying deep learning methods and advanced analytics to improve disease management at an individual patient level or for the discipline in general. Ophthatome TM captures data contained in retrospective electronic medical records between September 2012 and January 2018 to facilitate translational vision research through a knowledgebase of ophthalmic diseases. Methods: The electronic medical records data from Narayana Nethralaya ophthalmic hospital recorded in the MS-SQL database was mapped and programmatically transferred to MySQL. The captured data was manually curated to preserve data integrity and accuracy. The data was stored in MySQL database management system for ease of visualization, advanced search functions and other knowledgebase applications. Results: Ophthatome TM is a comprehensive and accurate knowledgebase of ophthalmic diseases containing curated clinical, treatment and imaging data of 581,466 ophthalmic subjects from the Indian population, recorded between September 2012 and January 2018. Ophthatome TM provides filters and Boolean searches with operators and modifiers that allow selection of specific cohorts covering 524 distinct ophthalmic disease types and 1800 disease sub-types across 35 different anatomical regions of the eye. The availability of longitudinal data for about 300,000 subjects provides additional opportunity to perform clinical research on disease progression and management including drug responses and management outcomes. The knowledgebase captures ophthalmic diseases in a genetically diverse population providing opportunity to study genetic and environmental factors contributing to or influencing ophthalmic diseases. Conclusion: Ophthatome TM will accelerate clinical, genomic, pharmacogenomic and advanced translational research in ophthalmology and vision sciences. Key words: Ophthatome, knowledge base, electronic medical records, curated clinical data, defined cohort, vision sciences

Background

The field of medical records has evolved tremendously in the last century. However, the earliest history of clinical or medical record is about 4000 years old, an Egyptian case report, a surgical note dating back to 1600BC. Hippocratic Corpus and other Greek scientific texts during the 5th Century BC and Medieval Islamic physicians’ case histories developed for didactic purposes are some of the early references available indicating practices of recording patient medical history. Detailed case books, recording symptoms and hospital records consisting admission and discharge books were followed by few physicians in the Europe and US in the mid-18th and early 19th century. The precedence for the modern medical records appeared in major centres in Paris and Berlin as loose paper files in the early 19th Century. The system of assigning a clinical number and integrating all records of the patient into a single
system to organize the scattered data was established by Henry S Plummer in 1907 at St. Mary's hospital and Mayo clinic in the US. The inherent problems of paperwork and the advancement in the information and communication technologies (ICT) in the latter half of this century have propelled in an era of electronic medical records or electronic hospital records (EMR/EHR) (1). The adoption of EMR has been faster and wider in the developed countries with countries like Canada, England, Germany, New Zealand achieving >90% of the physicians or hospitals employing EMR. The adoption of EMR in India is low compared to other countries and has been implemented by very few Government hospitals and large corporate hospitals chains including major tertiary care hospitals in the field of ophthalmology (2).

There are numerous ophthalmic registries in various countries, most of them being disease specific, like for cataract, glaucoma, age-related macular degeneration, corneal diseases, visual impairment and blindness, etc, which are either single centre or multicentric or regional or national or multinational, and the contribution being either mandatory or voluntary and in some of these the entry is through a web-based portal (3). The IRIS registry by the American academy of ophthalmology and French national registry are two large databases built on data from EMRs (4,5). The primary objectives of these registries are to understand the epidemiology of visual impairment and blindness, different diseases outcomes or adverse events associated with specific treatment or surgical procedures, thereby enabling improved patient care (3).

The numerous disease specific registries across the globe record isolated data elements like demography, diagnosis, visual acuity, investigations such as (ocular coherence tomography (OCT), fundus fluorescein angiography (FFA)), management or complications, but not all these data elements are captured together in one registry (3). However, the two large databases, IRIS Registry (Intelligent Research in Sight) (https://www.aao.org/iris-registry/about) with 50 million patients data contributed by 14,793 physicians from 2937 practices (institutional and private) and the French National Registry, Système national d'information inter-régime de l’assurance maladie (SNIIR-AM) with 65 million patients data contributed by 1546 public and private healthcare facilities have comprehensive data extracted from EMRs and have organized the data in a searchable format. The big data captured is analysed for clinical improvement (both at individual patients’ level and assess practitioner’s standard of care), assess treatment outcome, prevalence, and scientific research. However, no additional curation is performed on the data to ascertain for their data completeness, reliability, or validity (4,5).

The EMR/EHR has been in place in India’s premier private ophthalmic institutes like Narayana Nethralaya (NN), Sankara Nethralaya (SN), LV Prasad Eye Institute (LVPEI) since the last decade. The eyeSMART EMR system of LVPEI connects 176 primary vision care centres, 18 secondary, 4 tertiary centres to a central centre for excellence and has captured 2,270,584 patient’s data over an eight-year period between 2010 and 2018 (6). However, there is no published report yet on the system or volumes of data captured in their respective EMR from other ophthalmic institutes.

The electronic medical records (EMR) system is the digital storehouse of rich medical data that includes demographics, clinical (diagnosis, clinical diagnostic tests, laboratory diagnostic tests, prescription
treatment, surgical management,) and administrative (payment particulars including insurance) details of
patients’ reviews, cross-references and visits to the hospital(s). Although EMR is a repository of vast
clinical and health data on a large patient cohort collected over many years, the data lack sufficient
structure to be of any clinical value for applying deep learning methods and advanced analytics to
improve disease management at an individual patient level or for the discipline in general. Therefore,
aggregated data from hospital EMRs need to be captured in a structured knowledge base to support
clinical and translational research (CTR) (7). Medical big data analytics has revolutionized the human
healthcare system by introducing processes that facilitate rationale clinical decision making, predictive or
prognostic modelling of disease progression and management, disease surveillance, impact on public
health and research.

MedGenome Labs Ltd, India, collaborated with Narayana Nethralaya (a tertiary care state of the art
eyecare services) Bengaluru, India to create the Ophthatome™ - a knowledgebase of ophthalmic
diseases covering over 500,000 individual electronic clinical records. The study describes methods and
processes established for collecting high quality data, enriching the data with additional
endophenotyping and functionalities and presentation of the data in a web-based portal that allows basic
and advanced search functions. Finally, few case studies of the knowledgebase are given to highlight the
clinical application, relevance and efficiency of Ophthatome™.

Methods

Data

Ophthatome™ contains comprehensive clinical data captured from the electronic medical records (EMR)
between September 2012 and January 2018 from two centres of Narayana Nethralaya, a multi-speciality
tertiary eye hospital in Bangalore, India. This dataset from the Indian population has complete clinical
data of patient demography, disease specific attributes and treatment information recorded during each
patient visit. The EMR data from Narayana Nethralaya recorded in the MS-SQL database was mapped
and transferred to local MS-SQL. This was further programmatically transferred to MySQL. The captured
data was manually curated, where missing data and ambiguities in the data were resolved by working
with the clinician to preserve data integrity and accuracy. The data is stored in MySQL database
management system for ease of visualization, advanced search functions and other knowledgebase
applications. The data is completely anonymised with unique identifier number and the privacy and
confidentiality of the patient data is ensured. Written informed consent was obtained from the subjects or
legal guardians at the time of registration at the hospital. The study is approved by institutional ethics
committee of Narayana Nethralaya and adhered to the tenets of the Declaration of Helsinki.

Data curation

All demographic and clinical parameters for example, age, diagnosis, quantitative trait values were
reviewed for validity and erroneous data was cleaned or flagged and removed. For each of the
quantitative traits, the range of expected values were defined, and values falling outside the range were corrected after consultation with the clinician, or in case of ambiguity were deleted. Unstructured data, for example, disease diagnoses were reclassified using ICD10 code, personal history and family history of ophthalmic and systemic diseases was annotated with structured terminologies. *Ophthatome™* is populated with extensively annotated data that were checked for accuracy and uniformity (Figure 1). A brief overview of data types captured in *Ophthatome™* are summarized below.

(i) Age

The age was computed based on the date of birth and therefore reflects age of subjects at different visits when longitudinal data is captured. This enables catching errors for example, more than 110 years were marked as outliers, verified for accuracy of the data and corrected. Paediatric age group included subjects \( \leq \) 18yrs (8) and cases \( \gt; \) 18 years were considered adult in the cohort.

(ii) Disease diagnosis

The disease diagnosis in the EMR is based on ICD10 codes.(9) The ICD diagnosis was categorised into disease types and subtypes for comprehensive and informative querying of the data. For example, all types of cataract; cortical, nuclear, anterior subcapsular etc. were first grouped as cataract and then subclassified into their respective subtypes as shown in Table 1. Further, disease diagnoses entries not as per the ICD10 codes, or present in an unstructured format, or ‘free text’ in the EMR were assigned unique codes and grouped under appropriate disease types and subtypes. The unstructured disease diagnoses entries that specified two or more diseases, e.g, a) “Primary open angle glaucoma (POAG) with cataract” or b) “early cataract with mild non-proliferative diabetic retinopathy (NPDR)” were grouped under both the diseases, a) glaucoma and cataract, and b) cataract and diabetic retinopathy and appropriate endophenotypes, respectively (Table 2). Each of the disease diagnosis was also annotated with additional data for example, (i) affected ocular organ (cornea, conjunctiva, retina etc), (ii) age of onset of the disease (as congenital, age-related, unspecified etc) and (iii) possible causes of the disease such as a) infectious (viral, bacterial, fungal or parasitic) b) genetic (complex, monogenic, somatic, mitochondrial, chromosomal) c) age-related d) developmental e) systemic causes f) secondary to other ocular diseases g) trauma i) related to eye surgery j) inflammation k) contact lens l) medications m) autoimmune disease n) environmental/allergy o) unknown. These categorizations and labelling allow advanced search functions for identifying specific, well-defined cohorts and to raise specific queries.

(iii) Quantitative traits

a) Refractive error (RE)

Myopia and hyperopia were classified based on the degree as low, medium and high. The types of myopia are; low myopia \( \lt; \) -3.00D, medium myopia (-3.00D to -6.00D) and high myopia \( \gt; \) -6.00D and hyperopia; low hyperopia (+2.00D), medium hyperopia (+2.25D to +5.00D) and high hyperopia
(+5.00D) based on the spherical power in diopters (D). (10,11) Entries other than the accepted integer in positive or negative values representing hyperopia and myopia respectively, were marked and removed.

b) Intraocular pressure

Intraocular pressure (IOP) was classified as low, normal and high for the measurement values, &lt;12mmHg, 12-21mmHg and &gt;21mmHg, respectively. (12) IOP values greater than 80mm Hg were considered erroneous and were excluded as maximum measurement that can be recorded on applanation and NCT (noncontact tonometer) are 80 and 60mmHg, respectively.

(c) Central corneal thickness

Central corneal thickness (CCT) was defined as very thin, thin, average, thick and very thick for the values &lt;510µm, 510-540µm, 541-560µm, 561-600µm and &gt;600µm respectively (13). Values &lt;300µm and &gt;700µm were marked as outliers and corrected appropriately with reference to the longitudinal data if available or flagged and removed.

(iv) Systemic diseases

The details of the systemic disease(s) were converted into standard terminology such as, diabetes mellitus, hypertension, coronary artery disease, renal stones, asthma etc. These labels improve search and selection functions in the web portal. The unique unstructured free text entries were exported to an excel file and for each entry (for the common diseases mentioned above), the corresponding standard terminology was assigned manually, which was later mapped in the database.

(v) Prescription drugs

The EMR database contains a list of different prescribed ophthalmic and non-ophthalmic drugs as per their brand names. Generic names were appended for all the prescribed drug along with their brand names. Further, the drugs were also grouped as per their therapeutic or pharmacological classification based on WHO/ATC classification. For combination drugs too, the generic, and the therapeutic or pharmacological classification was appropriately appended. This was done to identify cohorts who may not have been treated with the same generic or brand name drugs, but drugs with similar mechanism of action or pharmacological properties.

(vi) Family history

Details of family history of disease(s) was available for 0.02% (12,465/561,466) of the cohort. The diseases were defined as either ophthalmic or systemic to choose cases with family history of either or both type of disease(s). Here again the unique unstructured free text entries were exported to excel, the entries defined as ‘ophthalmic’ or ‘systemic’ disease manually and mapped in the database.

(vii) Diagnostic procedural images
The diagnostic procedural images that include, electroretinogram (ERG), fundus photograph, fundus fluorescein angiography (FFA), optical coherence tomography (OCT), frequency doubling technology perimetry (FDT), Heidelberg retina tomograph (HRT3), nerve fibre analysis (GDX), confocal microperimetry (MAIA), visual field analysis (HFA), Pentacam (anterior eye segment tomography), amplitude scan (AScan), specular microscopy, color vision, electrocardiograph (ECG), are all integrated into the Ophthatome™. The images along with the visit dates are presented as deidentified data retaining the same resolution as that of the original image.

a) Quantitative values within the procedural images

The quantitative values like the visual field index (VFI) as percentage, median deviation (MD), pattern standard deviation (PSD) available in the visual field analysis image, average retinal nerve fibre layer (RNFL) thickness in micrometer, cup/disc ratio, available in the OCT image were programmatically read, and the values mapped and integrated in the database.

(viii) Visual impairment

The cases were defined either as normal vision or visually impaired based on latest unaided or best corrected visual acuity. First, the values in the unaided and best corrected visual acuity were classified as normal, visual impairment, severe visual impairment or blind based on standard WHO definition (14). Cases are defined normal if the unaided refraction is ‘normal’. When the unaided refraction was classified either as visual impairment, severe visual impairment or blind, then the classification under the best corrected visual acuity of the better eye was considered to classify the visual impairment status of the individual.

(ix) Longitudinal data

Approximately fifty percent of the overall cohort in the knowledgebase have longitudinal data, i.e. have some clinical data recorded for more than one visit. In the Ophthatome™, for any visit &gt;1, an internal check of at least two datapoints for either refraction or glass prescription or slit-lamp examination was considered to short list cohort with longitudinal data.

Results

Ophthatome™ contains clinical and phenotype data of 581,466 cases (males - 318,383, females - 263,083) with 524 distinct ophthalmic disease types and 1800 disease sub-types.

The knowledgebase comprises diseases affecting 35 different anatomical parts of the orbit, lids and the eye including the anterior segment, posterior segment and eye adnexa. Figure 2A summarizes the number of patient data available for various diseases affecting different parts of the eye. The knowledgebase includes comprehensive ophthalmic clinical variables such as refraction, intraocular pressure, central corneal thickness, slit-lamp examination details, diagnosis, prescription medications, surgical
interventions, and thirteen clinical diagnostic images. The available clinical variables for each case were mapped longitudinally starting with the most recent value. Ophthatome™ is unique as it contains curated quantitative trait datapoints. Of the total of 12,234,278 quantitative datapoints (RE, IOP, CCT), 9,240,394 (77%) were within the defined range of value for the trait. Manual curation added 506,500 (5%) datapoints resulting in a total of 9,926,894 (82%) accurate datapoints (Figure 1).

Analysis of age-related ocular diseases

The top 25 ophthalmic diseases affecting <18 years (paediatric cases), >50 years and for the overall cohort are given in figure 2B, 2C, 2D respectively. A comparison between the two age groups, show that refractive error is equally prevalent in both the groups (46,569/102,590 paediatric cases and 271,147/478,876 cases in the >18 years age group), whereas presbyopia, which is a condition associated with age are completely absent in ≤18 years old and is diagnosed in 217,765 cases >18 years age. Similarly, late-onset eye disease, like age-related macular degeneration was completely absent in the paediatric group as expected and is reported in 2166 cases of >18 years age. Conversely, retinopathy of prematurity and retinoblastoma is reported in the paediatric group only (4695 and 213 cases, respectively) and completely absent in the adult. Cataract is reported in both age groups (3511/102,590 of ≤18 years and 187,314/478,876 of >18 years), but the prevalence is higher in the >18 years cohort reiterating that congenital form is rare compared to common age-related cataract. The same is observed with glaucoma as well (813 cases in ≤18 years and 25,485 cases in >18 years). Further, the prevalence of primary angle closure glaucoma (PACG) (2983 of the 25,485 (11.7%) glaucoma cases in >18 years age) is almost similar to primary open angle glaucoma (POAG) (2763/25,485 (10.8%)) in this cohort which is in accordance with the fact that prevalence of PACG is higher in Asians compared to Caucasians. The clinical data summary of the overall cohort is presented in Table 3.

Ophthatome™ contains 5,777 (0.99%) cases reporting family history of certain ophthalmic disease, 7,010 (1.20%) cases reporting a family history of systemic disease(s) and 186 cases (0.03%) reporting family history of both ophthalmic and systemic diseases. Figure 2E, 2F and 2G represents the details of systemic diseases of the family members (first or second degree relative or spouse), family history of ophthalmic diseases and disease diagnosis and family history of same disease, respectively in the 186 cases with family history of both systemic and ophthalmic diseases. Figure 2G shows that among all the common eye diseases, glaucoma and inherited retinal degenerations have a genetic component as expected.

Research and Clinical Applications of Ophthatome™

Cohort selection
Complex diseases present with various analytic challenges, including lack of homogeneity in disease phenotype, variability in disease progression rate, response to drugs and disease outcome. Therefore, aggregation of clinical data from large patient cohorts are needed to identify common as well as divergent trends in disease progression. *Ophthatome™* provides many strategies for patient selection through filters and Boolean searches with operators and modifiers. Selection of cohorts based on (i) demographics, (ii) disease type and subtype, (iii) quantitative traits like a) refractive error (myopia and hyperopia), b) intraocular pressure, c) central corneal thickness, (iv) number of visits (v) systemic diseases (vi) family history of ophthalmic and/or systemic diseases (vii) diagnostic procedural images (viii) visual field index values (ix) visual impairment status and (x) drugs prescribed, are available with the ability to build complex multiparametric searches.

A series of filters covering demography, clinical features and prescribed drugs can be applied in *Ophthatome™* (Figure 3A) to select a cohort of patient samples (10,540 cases) as represented in Figure 3B. Such a large number of disease samples can be further narrowed using additional filters to create a near homogeneous cohort sharing specific disease phenotypes for clinical or genetic analysis, for example.

**Cohort selection for deeper analysis**

Few case studies are as presented here.

Glaucoma is a heterogeneous disease with multiple subtypes, and with a myriad of treatment options. To compare patient's response to beta-blockers alone, or in combination with other anti-glaucoma medications, a combination of filters was used to select appropriate disease cohorts as shown in Figure 4. Glaucoma used as a disease type and ocular hypertension used as disease subtype, shortlists 26,298 (4.52% of the entire cohort) and 1904 cases (0.32%), respectively. Additional filters applied successively include, >1 visit – 1630 cases, high intraocular pressure – 1513 cases, drugs prescribed; cohort-A, beta blockers + other category of drugs – 363 cases, cohort-B, beta blockers + other anti-glaucoma medication (AGM) – 72 cases, and cohort-C, beta blockers alone – 33 cases. Availability of Humphrey visual field image for each of the three cohorts further narrowed the numbers as shown in Figure 4. The comprehensive clinical details of each patient can be extracted in a tabular format from the database and used for further analytics. Follow up data are available for a subset of patients under different treatment conditions and analysing cases with at least two VFI datapoints at ≥ one-year interval (Table 4) revealed number of cases that were stable and those progressed. For example, of the 9 individuals treated with beta blockers alone, 7 had stable VFI and 2 converted to primary angle closure (PAC) within 2 years. Additionally, of the 43 patients treated with beta blockers + AGM, 11 cases were not further analysed as they were diagnosed with either POAG or secondary glaucoma in the first visit along with OHT. However, 29 of the remaining 32 treated patients had stable VFI for four years, one case converted
to PAC and two individuals progressed to POAG, in 4 years. Cumulative data from drug responsive and non-responsive patient groups can be further investigated using molecular and genomic tools.

Drugs blocking growth of new blood vessels (neoangiogenesis) has improved vision dramatically in a large subset of patients with diabetic retinopathy and wet age-related macular degeneration (wet-AMD). Ophthatome™ has information of 5466 patients treated with anti-angiogenic therapy such as Avastin and Lucentis (Table 5). Follow up data on these individuals have identified drug resistance in a fraction of the treated patients opening opportunities to understand the mechanism of resistance (Unpublished data).

Seasonal allergen or seasonal changes show increased incidence of infectious or non-infectious conjunctivitis. Ophthatome™ has 12,594 cases of conjunctivitis and the subtype classification reveal 7414 cases of viral, 1604 bacterial and 912 allergic conjunctivitis. Analysing the time period of highest incidence of infection or allergy would reveal details of epidemiological trends in disease outbreak, the causative infective agent or the appearance of seasonal allergens that could be used for treatment and prophylactic disease management.

**Discussion**

Comprehensive and curated phenotype data captured over many years provide a wealth of quantitative and qualitative clinical parameters to perform deeper phenotype analysis, disease progression/prognosis, severity of disease and variations in response to therapies. Ophthatome™ is a unique database that captures eye diseases from a genetically heterogeneous population in India.

There are number of reports published from the EMR based big data registries establishing and reiterating the need for such databases. From the IRIS registry - A study to understand the incidence of post cataract surgery endophthalmitis identified the incidence as 0.04% for cataract surgery performed on 8,542,838 eyes (5,401,686 patients) between 2013 and 2017. Younger age and cataract surgery along with other ophthalmic surgery and anterior vitrectomy was found to be risk factors (15). In another study to assess the real world management of diabetic macular edema (DME), it was observed that cases presented with better visual acuity (VA) at presentation received no treatment but those presented with lower VA at index date were promptly treated with anti-VEGF which improved the VA at end of one year, and further patients who received >5 injections per year had a better VA gain compared to those who received 1-5 injections during the same period (16). In yet another analysis from the IRIS registry study, performance rate measurements with feedback to physicians and administrators had improved subsequent performance (17). The two publications from the EPISAFE programme on the congenital cataract surgery and cataract surgery report the incidence of congenital cataract between 2010 and 2012 to be 1-3/1000 birth, same as the other industrial nations, while the incidence of cataract surgery (2,717,203 eyes and 1,817,865 patients) increased from 9.86 – 11.08/1000-year between 2009 and 2012. The incidence of retinal detachment (RD) and pseudophakic cystoid macular edema being 0.99% and 0.95%, respectively with higher risk of RD for patients <60 years of age (5,18).
Two reports from the eyeSmart database, currently the only published EMR based bigdata database from India reveal incidence of dry eye disease (DED) and allergic eye disease (AED) in the population studied between the years 2010 and 2018.

Of the 1,458,830 new patients recruited 21,290 (1.46%) had recent-onset DED with age, gender, residence, occupation, socio-economic status having a significant impact of development of the disease (19), while of the 259,969 new patients (≤21years) 10.1% had AED. Boys with atopy during middle childhood and from middle to higher income families were commonly affected AED which was self-limiting by adolescence. (20).

These above reports present the real-world clinical practice and their outcomes that provide information and details on incidence, trends, risk factors, treatment outcome, prophylaxis, and quality assessment.

*Ophthatome™* though currently has >0.5 million patient records, the data is captured from a single hospital contributed by more than 100 ophthalmologists. The data was aggregated using a highly streamlined and standardized format using uniform clinical terms across departments and subspecialities in the organization, thereby avoiding ambiguity, which may arise if data is captured across many different medical facilities as is the case with other large registries. As described in the Results section, despite uniformity of data capture, manual curation of the captured data revealed lack of consistency in 23% of the data. This was resolved in further 5% through discussion with the clinicians resulting in 82% of the data being accurate. The public registries driven by clinician-reported clinical and phenotype data are most likely to contain inaccuracies and inconsistencies, which are difficult to resolve in a short time, given the size of the data and the scope of the registries and differences in annotation practices.

Yet another limitation of the large registries like the IRIS is the unavailability of the diagnostic images. In the study presenting the real-world management of DME, the data on number of DME cases is extracted based on ICD, where the ophthalmologist may have defined the case with the given code for further follow up and eligibility and does not indicate confirmed diagnosis. It is not possible to assess the severity of the disease in the absence of imaging data and thus the percent of patients managed with intervention and observation only and the outcome needs to be understood and interpreted with caution (16). *Ophthatome™* integrates diagnostic images along with the quantitative values within the image programmatically captured and mapped in the database. This adds much value to the knowledgebase as it helps to further confirm the disease severity and suitability of the chosen cohort for further analysis.

*Ophthatome™* is a licensed database with patient-consented data which enables users to build highly specific disease cohorts using combined search terms and built in logical functions for research.

*Ophthatome™* will be updated on a regular basis to capture new data for existing entries in the database as well as new patient data, though currently it is not linked real-time.
Conclusions

*Ophthatome™* represents a versatile curated database of over 1800 eye diseases with longitudinal data associated with disease outcome, drug response and secondary complications. It also provides opportunity to design genomic and pharmacogenomic studies that would help understand the biology of complex eye diseases and is a powerful tool for advanced research in eye diseases, including phase IV studies in a unique population.

Link for demo version

https://demo.ophthatome.com/

List Of Abbreviations

1. Central corneal thickness (CCT)
2. Clinical and translational research (CTR)
3. Electrocardiograph (ECG)
4. Electronic medical records (EMR)
5. Electroretinogram (ERG)
6. Fundus fluoresce angiography (FFA)
7. Intra ocular pressure (IOP)
8. Median deviation (MD)
9. Moderate to severe visual impairment (MSVI)
10. Non-proliferative diabetic retinopathy (NPDR)
11. Optical coherence tomography (OCT)
12. Pattern standard deviation (PSD)
13. Primary angle closure glaucoma (PACG)
14. Primary open angle glaucoma (POAG)
15. Retinal nerve fibre layer (RNFL)
16. Visual field index (VFI)

Declarations

Ethics approval and consent to participate

The study was approved by institutional ethics committee of Narayana Nethralaya (EC Ref No. 2020/02/01) and written consent was obtained from all subjects.

Consent for publication
Not applicable

**Availability of data and materials**

The datasets analysed during the current study are not publicly available due to the confidentiality of the patients’ clinical data.

**Competing interests**

The authors declare that they have no competing interests

**Funding**

None

**Authors' contributions**

PR – The acquisition of data and analysis and creation of new software used in the work

ST – Interpretation of data, revised the manuscript

DS – The curation of data

MNB - The acquisition of data and analysis and creation of new software used in the work

CT - The acquisition of data and analysis and creation of new software used in the work

SD - Interpretation of data, revised the manuscript

SJ - The acquisition of data and analysis and creation of new software used in the work

SM - The curation of data

DK - The acquisition of data and analysis and creation of new software used in the work

NP – The acquisition of data

RS - The acquisition of data, revised the manuscript

AG - The acquisition of data, revised the manuscript

GK – Revised the manuscript substantially

AC – Conception, interpretation of work, substantially revised the manuscript

NS – conception, design of the work, the acquisition, analyses and interpretation of data, drafted the work

All authors have read and approved the manuscript
Acknowledgements

Not applicable

References

1. Gillum RF. From papyrus to the electronic tablet: a brief history of the clinical medical record with lessons for the digital age. Am J Med. 2013 Oct;126(10):853–7.
2. Srivastava SK. Adoption of Electronic Health Records: A Roadmap for India. Healthc Inform Res. 2016 Oct;22(4):261–9.
3. Tan JCK, Ferdi AC, Gillies MC, Watson SL. Clinical Registries in Ophthalmology. Ophthalmology. 2019 May;126(5):655–62.
4. Parke DW, Rich WL, Sommer A, Lum F. The American Academy of Ophthalmology’s IRIS® Registry (Intelligent Research in Sight Clinical Data): A Look Back and a Look to the Future. Ophthalmology. 2017;124(11):1572–4.
5. Daien V, Korobelnik J-F, Delcourt C, Cougnard-Gregoire A, Delyfer MN, Bron AM, et al. French Medical-Administrative Database for Epidemiology and Safety in Ophthalmology (EPISAFE): The EPISAFE Collaboration Program in Cataract Surgery. Ophthalmic Res. 2017;58(2):67–73.
6. Das AV, Kammari P, Vadapalli R, Basu S. Big data and the eyeSmart electronic medical record system - An 8-year experience from a three-tier eye care network in India. Indian J Ophthalmol. 2020;68(3):427–32.
7. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE–An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc AMIA Symp. 2009 Nov 14;2009:391–5.
8. Khadilkar VV, Khadilkar AV. Revised Indian Academy of Pediatrics 2015 growth charts for height, weight and body mass index for 5-18-year-old Indian children. Indian J Endocrinol Metab. 2015 Aug;19(4):470–6.
9. World Health Organization. (2004). ICD-10: international statistical classification of diseases and related health problems: tenth revision. 2nd ed. World Health Organization.;
10. McMonnies CW. Clinical prediction of the need for interventions for the control of myopia. Clin Exp Optom. 2015 Nov;98(6):518–26.
11. American Optometric Association. Optometric Clinical Practice Guideline: Care of the Patient with Hyperopia. American Optometric Association, St. Louis, MO; 1997.
12. Sihota R, Angmo D, Ramaswamy D, Dada T. Simplifying “target” intraocular pressure for different stages of primary open-angle glaucoma and primary angle-closure glaucoma. Indian J Ophthalmol. 2018;66(4):495–505.
13. Galgauskas S, Juodkaite G, Tutkuvienė J. Age-related changes in central corneal thickness in normal eyes among the adult Lithuanian population. Clin Interv Aging. 2014;9:1145–51.
14. Bhushan Punani and Nandini Rawal. Visual impairment handbook. In: Second edition. Ahmedabad, India: Blind people association, India; 2000. p. 1–10.

15. Pershing S, Lum F, Hsu S, Kelly S, Chiang MF, Rich WL, et al. Endophthalmitis after Cataract Surgery in the United States: A Report from the Intelligent Research in Sight Registry, 2013-2017. Ophthalmology. 2020;127(2):151–8.

16. Cantrell RA, Lum F, Chia Y, Morse LS, Rich WL, Salman CA, et al. Treatment Patterns for Diabetic Macular Edema: An Intelligent Research in Sight (IRIS®) Registry Analysis. Ophthalmology. 2020;127(3):427–9.

17. Rich WL, Chiang MF, Lum F, Hancock R, Parke DW. Performance Rates Measured in the American Academy of Ophthalmology IRIS© Registry (Intelligent Research in Sight). Ophthalmology. 2018;125(5):782–4.

18. Daïen V, Le Pape A, Heve D, Villain M, Bremond Gignac D, Collaborators of the Epidemiology and Safety Program (EPISAFE). Incidence and Characteristics of Congenital Cataract Surgery in France from 2010 to 2012: The EPISAFE Program. Ophthalmic Res. 2017;58(2):114–6.

19. Donthineni PR, Kammari P, Shanbhag SS, Singh V, Das AV, Basu S. Incidence, demographics, types and risk factors of dry eye disease in India: Electronic medical records driven big data analytics report I. Ocul Surf. 2019 Apr;17(2):250–6.

20. Das AV, Donthineni PR, Sai Prashanthi G, Basu S. Allergic eye disease in children and adolescents seeking eye care in India: Electronic medical records driven big data analytics report II. Ocul Surf. 2019;17(4):683–9.

Tables

Table 1 Categorization of diseases with ICD codes into disease types and subtypes for comprehensive and information querying in the *Ophthatometm*
| ICD_code | Diagnosis_name                                      | Disease | Sub_type                                         |
|----------|---------------------------------------------------|---------|--------------------------------------------------|
| H25011   | Cortical age-related cataract, right eye           | Cataract| Cortical age-related cataract                    |
| H25012   | Cortical age-related cataract, left eye            | Cataract| Cortical age-related cataract                    |
| H25013   | Cortical age-related cataract, bilateral           | Cataract| Cortical age-related cataract                    |
| H25019   | Cortical age-related cataract, unspecified eye     | Cataract| Cortical age-related cataract                    |
| H25031   | Anterior subcapsular polar age-related cataract, right eye | Cataract| Anterior subcapsular age-related cataract        |
| H25032   | Anterior subcapsular polar age-related cataract, left eye | Cataract| Anterior subcapsular age-related cataract        |
| H25033   | Anterior subcapsular polar age-related cataract, bilateral | Cataract| Anterior subcapsular age-related cataract        |
| H25039   | Anterior subcapsular polar age-related cataract, unspecified eye | Cataract| Anterior subcapsular age-related cataract        |
| H2504    | Posterior subcapsular cataract                     | Cataract| Posterior subcapsular cataract                   |
| H25041   | Posterior subcapsular polar age-related cataract, right eye | Cataract| Posterior subcapsular age-related cataract        |
| H25042   | Posterior subcapsular polar age-related cataract, left eye | Cataract| Posterior subcapsular age-related cataract        |
| H25043   | Posterior subcapsular polar age-related cataract, bilateral | Cataract| Posterior subcapsular age-related cataract        |
| H25049   | Posterior subcapsular polar age-related cataract, unspecified eye | Cataract| Posterior subcapsular age-related cataract        |

Table 2 Categorization of unstructured disease diagnosis (non-ICD diagnosis) into disease types and subtypes for comprehensive and information querying in the Ophthatome™

| ICD_code | Diagnosis_name                                      | Disease | Sub_type                                         |
|----------|---------------------------------------------------|---------|--------------------------------------------------|
| UK1486   | Traumatic cataract LE                              | Cataract| Traumatic cataract                                |
| UK2123   | Partially absorbed traumatic cataract              | Cataract| Traumatic cataract                                |

Table 3 Clinical data summary of the overall cohort

| Clinical parameters information available (at least one datapoint) | Number of cases |
|-------------------------------------------------------------------|-----------------|
| Diagnosis                                                         | 568,324         |
| Refraction                                                        | 534,253         |
| Intraocular pressure                                              | 520,874         |
| Central corneal thickness                                         | 11,460          |
| Systemic history                                                  | 239,966         |
| Medications prescribed                                            | 231,223         |
| Ophthalmic surgery performed                                      | 134,686         |
Table 4 Number of cases treated with beta blocker alone or along with other AGM and at least 2 VFI datapoints at >1 year interval

| # of cases prescribed with BB only and >1 year interval between the VFI datapoints | 9 |
| # cases prescribed with BB and other AGM and >1 year interval between the VFI datapoints | 43 |

Table 5 Diseases and number of cases treated with different anti-angiogenic drugs

| Anti-VEGF molecules          | Avastin | Lucentis | Accentrix | Eylea | Razumab | Anti-VEGF (not specified) |
|------------------------------|---------|----------|-----------|-------|---------|---------------------------|
| Total number of cases        | 3706    | 942      | 860       | 35    | 310     | 279                       |
| Diseases wise distribution   |         |          |           |       |         |                           |
| Wet AMD                      | 60      | 17       | 23        | 3     | 7       | 5                         |
| AMD                          | 102     | 36       | 38        | 8     | 17      | 10                        |
| Diabetic retinopathy         | 1272    | 256      | 190       | _     | 65      | 75                        |
| Retinal vein occlusion       | 496     | 122      | 119       | 4     | 28      | 26                        |
| Myopic CNMV                  | 4       | 1        | _         | _     | _       | _                         |
| Pathological myopia          | 5       | 3        | 1         | _     | 1       | 1                         |
| Neovascular glaucoma         | 199     | 37       | 27        | 1     | 10      | 12                        |
| Retinopathy of prematurity   | 2       | _        | _         | _     | _       | _                         |
| Retinal vascular disorder    | 579     | 174      | 65        | 10    | 18      | 21                        |
| Retinal vasculitis           | 20      | 3        | 2         | _     | _       | 1                         |
| Chorioretinopathy           | 20      | 7        | 4         | 1     | 1       | 3                         |

Figures
Figure 1

Shows the number of cases with clinical diagnosis and the number of curated accurate datapoints in the Ophthatome™ knowledgebase.
Figure 2A summarizes the number of patient data available for various diseases affecting different parts of the eye in the database. Figure 2B shows the top 25 ophthalmic diseases of the paediatric cases (≤ 18 years) in the database. Figure 2C shows the top 25 ophthalmic diseases of the >50 years cohort in the database. Figure 2D shows the top 25 ophthalmic diseases of the overall cohort in the database. Figure 2E represents common systemic diseases of the 186 cases with family history of both ophthalmic and
systemic diseases. Figure 2F presents details of family history of ophthalmic diseases of the 186 cases with family history of both ophthalmic and systemic diseases. Figure 2G shows the details of ophthalmic disease and family history of same disease in the 186 cases with family history of both ophthalmic and systemic diseases.
Figure 2A summarizes the number of patient data available for various diseases affecting different parts of the eye in the database. Figure 2B shows the top 25 ophthalmic diseases of the paediatric cases (≤ 18 years) in the database. Figure 2C shows the top 25 ophthalmic diseases of the >50 years cohort in the database. Figure 2D shows the top 25 ophthalmic diseases of the overall cohort in the database. Figure 2E represents common systemic diseases of the 186 cases with family history of both ophthalmic and systemic diseases. Figure 2F presents details of family history of ophthalmic diseases of the 186 cases with family history of both ophthalmic and systemic diseases. Figure 2G shows the details of ophthalmic disease and family history of same disease in the 186 cases with family history of both ophthalmic and systemic diseases.
Figure 3A represents a flow chart of choosing filters successively to obtain a specific cohort. Figure 3B is the snapshot of specific cohort shortlisted after selecting filters successively as shown in Figure 3A.
Figure 3A represents a flow chart of choosing filters successively to obtain a specific cohort. Figure 3B is the snapshot of specific cohort shortlisted after selecting filters successively as shown in Figure 3A.
**Figure 4**

shows combination of filters used, and the cohorts obtained that would help to study and understand the response of beta blockers alone or in combination with other anti-glaucoma (AGM) drugs in cases with ocular hypertension.
Figure 4 shows combination of filters used, and the cohorts obtained that would help to study and understand the response of beta blockers alone or in combination with other anti-glaucoma (AGM) drugs in cases with ocular hypertension.