Defining a comprehensive verotype using electronic health records for personalized medicine

Mary Regina Boland,1 George Hripcsak,1 Yufeng Shen,1,2 Wendy K Chung,3,4,5 Chunhua Weng1,5

ABSTRACT

The burgeoning adoption of electronic health records (EHR) introduces a golden opportunity for studying individual manifestations of myriad diseases, which is called ‘EHR phenotyping’. In this paper, we break down this concept by relating it to phenotype definitions from Johannsen; comparing it to cohort identification and disease subtyping; introducing a new concept called ‘verotype’ (Latin: vere = true, actually) to represent the ‘true’ population of similar patients for treatment purposes through the integration of genotype, phenotype, and disease subtype (eg, specific glucose value pattern in patients with diabetes) information; analyzing the value of the ‘verotype’ concept for personalized medicine; and outlining the potential for using network-based approaches to reverse engineer clinical disease subtypes.

INTRODUCTION

During the seminal days of genetic research, researchers sought ways of describing and defining the complex topic of heritability while pondering over the transmission of traits1–3 before discovering the ‘gene’.4 Genetics research introduced the concept of genotypes, phenotypes, enterotypes (‘types’ based on composition of gut microbiota),5 endophenotypes (proximal disease-related phenotype with a clear genetic component regardless of disease presence)6–8 and deep phenotypes (detailed phenotype),9,10 enabling us to define human characteristics that reflect myriad disease states. The rapid adoption of electronic health records (EHR)11 introduces a new opportunity for disease characterization.

In this paper, we take a historical perspective to break down the concept of ‘EHR phenotyping’ by comparing the concept to those outlined by Johannsen.1 We also discuss the value of disease subtyping using EHR to identify related groups of patients useful for developing personalized medical treatment regimens. Then, we outline the value of network-based approaches for reverse engineering disease subtypes from EHR.

BREAKING DOWN EHR PHENOTYPING USING JOHANNSEN’S DEFINITIONS

Historical background

Mendel described the pattern of transmission of ‘characters’ (or alleles) from parent to offspring (ie, genotype) as either dominant or recessive.2,3 A dominant allele controls the expression of a trait even if an individual is heterozygous (ie, possessing only one of two copies at a single locus). A recessive allele will not affect an individual’s trait unless they are homozygous. Consequently, recessively inherited traits disappear in a generation and then reappear in subsequent generations.2,3 Later, Johannsen coined the terms phenotype, genotype, and biotype.1 These concepts were described before discovering that DNA transmits heritable characteristics to individuals.4

We illustrate the interrelationship among these concepts using eye color, a complex trait.12 Eye color can change as a result of health status13 and access to medical treatment,14 with 16 genes contributing to its heritability (genotype) in humans. Interestingly, individuals with lower social status developed darker eyes than those with high social status in Nile tilapia15 suggesting that other factors may also affect eye color. We use Johannsen’s term biotype to describe individuals with the same genotype and phenotype1 as opposed to other slightly modified definitions.16–18 One example biotype consists of individuals with a genotype for blue eyes, but possessing green eyes (darkening of eye color was found in women with many pregnancies);19 while another example biotype consists of individuals with a genotype for green eyes and possessing green eyes (normal phenotype). Interestingly, certain individuals have a different phenotype in each of their eyes (heterochromia), but their underlying genotype is the same,20 which is a third example biotype.

Hippocrates described the identification of disease subtypes.21 The characterization of disease subtypes is called ‘deep phenotyping’ by some researchers22 while others reserve it for genetic information.23 Because we focus on clinical data stored in EHR, we use ‘clinical disease subtype’21 24 25 throughout this paper. A ‘clinical disease subtype’ is any ‘type’ that stratifies a diseased population into subpopulations. Table 1 provides a summary of definitions with medical examples.

Phenotypic variance

Johannsen describes two factors that introduce phenotypic variance: environmental and genetic (table 2).3 In EHR, phenotypic variance can also be introduced by variability in healthcare practice and medical decision-making among care providers26 27 or by varying documentation behaviors,28 which adds two factors that may contribute to phenotype variance: the healthcare process and documentation behavior (table 2).2

In figure 1, we illustrate influential factors that affect the traditional and EHR-based phenotypes, respectively. Many factors affect EHR phenotypes including clinicians’ documentation behavior. The experience of the person documenting can affect...
the degree of detail contained in the documentation. For example, a medical student may include more details on certain less relevant items and then miss critical items. To detect this in EHR, notes can be compared across clinicians of varying experience levels for the same set of patients. Agreement could be assessed and outliers identified (eg, highly skilled clinicians). If outliers are recorded by many EHR. However, these data are not always highly skilled clinicians (ie, outliers) may be more useful as skill and predictive ability are related. Some factors, such as lifestyle, are recorded by many EHR. However, these data are not always stored in a standard form, and may require specialized extraction methodologies. For example, smoking status can be assessed in multiple ways including using clinical notes, and billing codes. These factors can all introduce phenotypic variance. Differences between EHR data and the true patient state are described elsewhere.

**PHENOTYPING WITH EHR**

**Phenotyping as cohort identification**

Cohort identification, namely identifying patients with or without a given disease, for example, type 2 diabetes mellitus, is a popular use of EHR, called ‘EHR-based phenotyping’ or ‘EHR-driven phenotyping’. In phenome-wide association studies, EHR-based phenotyping is used to identify patient cohorts that possess a phenotype, for example, diabetes mellitus, hypothyroidism, cataracts, before associating phenotypes with genetic markers. EHR-based phenotyping is also used for electronic prescreening to determine patients’ eligibility for clinical trials. Importantly, cohort identification existed before EHR and therefore EHR are not necessarily required. However, there are situations in which cohort identification would be impractical without utilizing EHR. This is particularly true for identifying cohorts with a rare disease or outcome. Using an integrated EHR, over 33,000 HIV patients (a rare disease) were identified. A cohort of that size would be impossible, or practically unfeasible, to identify without the use of EHR. In general, EHR facilitate the process of cohort identification and often result in studies with greater power and lower cost, but they also possess their own unique set of challenges.

**Table 1** Adaption of traditional phenotyping terminology to the EHR context

<table>
<thead>
<tr>
<th>Term</th>
<th>Genetic definition</th>
<th>Adaption to the EHR context</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johannsen</td>
<td>“We do not know a ‘genotype’, but we are able to demonstrate ‘genotypical’ differences or accuracies…”</td>
<td>Clinical data redefinition</td>
<td>BRCAL alleles, TCF7L2, glucokinase, HLA alleles</td>
</tr>
<tr>
<td>Phenotype</td>
<td>“We may easily find out that the organisms in question resemble each other so much that they belong to the same ‘type’…”</td>
<td>Any phenotype, for example, diabetes, height, weight, that has related data elements extractable from EHR data</td>
<td>Height, diabetes, atherosclerosis</td>
</tr>
<tr>
<td>Biotype</td>
<td>A group of organisms characterized by having the same phenotype and genotype.</td>
<td>Using Johannsen’s definition for phenotype, we define ‘clinical disease subtype’ to be any set of characteristics that distinguishes a subset of diseased patients from the overall diseased population</td>
<td>Breast cancer and BRCA1</td>
</tr>
<tr>
<td>Hippocrates</td>
<td>Heterogeneous diseases can be classified into smaller disease ‘subtypes’ when the subtypes have different characteristics (eg, tissue-based biomarker, mutation, and symptom).</td>
<td></td>
<td>Chronic, benign, malignant</td>
</tr>
</tbody>
</table>

EHR, electronic health record; HLA, human leucocyte antigen.

**Table 2** Factors that introduce phenotypic variance in genetic and EHR data

<table>
<thead>
<tr>
<th>Type of phenotype variance</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic</td>
<td>Genetic trait could result in organisms with shorter than expected heights</td>
</tr>
<tr>
<td>Environmental (non-inherited)</td>
<td>Food shortage could result in organisms with shorter than expected heights</td>
</tr>
<tr>
<td>Healthcare process</td>
<td>Presence or absence of insurance could result in individuals with un or underreported height</td>
</tr>
<tr>
<td>Healthcare documentation</td>
<td>Presence or absence of clinician experience could result in inadequate or inconsistent measurement of individuals’ height</td>
</tr>
</tbody>
</table>

EHR, electronic health record.

**Phenotyping as disease subtype discovery**

Another use case is for identifying novel disease subtypes using clinical data from EHR. This ‘disease subtyping’ depends on the identification of a higher-level ‘parent’ phenotype, that is, the disease. Before EHR, identifying disease subtypes was challenging and in many cases it required glaring phenotypic differences. For example, types of diabetes were initially distinguished by age, namely juvenile (type 1) and adult (type 2) diabetes. Over time, these categories were made more descriptive with insulin dependent (type 1) and non-insulin dependent (type 2), which eventually gave way to ‘type 1’ and ‘type 2’. Therefore, disease subtyping was possible before EHR using clinical data (eg, observations, chart review). However, it was challenging as initial observations (juvenile vs adult) regarding a disease subtype were often incomplete. In genetics, disease subtyping often occurs by identifying genetic or molecular ‘biomarkers’ (ie, disease subtype) that segregate a diseased population into subpopulations. An example of EHR-based clinical disease subtyping includes identifying a patient subpopulation...
A patient may have a disease status indicating that their breast cancer is ‘critical’ second disease. For example, if a patient has advanced diabetes then their status for a second disease unaffected (advanced breast cancer still present). The loop at the top of disease status indicates that a disease hospital due to a car accident and is in a well-controlled condition as examples of patient status. For disease status, we include ‘early’ (eg, stage i), ‘advanced (eg, stage iii)’ as examples. A patient may have a disease status indicating that their breast cancer is ‘advanced or stage iii’ while their disease status would remain unaffected (advanced breast cancer still present). The loop at the top of disease status indicates that a disease’s status can affect the status of a second disease. For example, if a patient has advanced diabetes then their status for a second disease—retinopathy—could be affected.

VEROTYPE: THE PATIENT’S ‘TRUE’ TYPE
Learning from Johannsen’s definition of biotype as a group of organisms with the same phenotype and genotype, we introduce a new concept called ‘verotype’ from the Latin word vere, meaning truly or actually. This higher level ‘type’ defines a unique combination of genotype, phenotype, and disease subtype for an individual. We named it verotype because it indicates the true subpopulation that a patient belongs to, for example, diabetic with unique glucose pattern (ie, disease subtype), for example, diabetic with unique glucose pattern (ie, disease subtype), and genetic underpinnings related to the disease.

The conventional approach
Before large-scale data mining of EHR, clinical disease subtyping was performed by collecting clinical data from patients with a given disease, for example, Parkinson’s disease (PD). Patients were then clustered based on their observed clinical findings and statistically significant clusters were considered PD subtypes. Afterwards, the relationship between each PD subtype and outcome had to be established and verified. Using EHR enables researchers to develop algorithms for disease subtype classification, and to identify clinical features associated with a disease subtype, for example, estrogen/progesterone negative breast cancer.

The high-throughput approach
EHR offer the opportunity to develop novel methods for investigating new disease characteristics using clinical data, for example, laboratory values. Furthermore, novel disease subtypes identified from EHR have the potential for predicting patient outcomes more accurately than predefined subtypes. This is particularly true for poorly characterized mental...
diseases,\textsuperscript{65} for example, PD,\textsuperscript{58} depression\textsuperscript{50} and amyloid lateral sclerosis.\textsuperscript{49} 51

A proposal to apply a network-based approach to clinical disease subtyping

This led us to look to ‘network medicine’\textsuperscript{66} for a solution. Network medicine involves integrating knowledge from various sources including genes, biological pathways, protein–protein interaction complexes, and so on, to identify tailored biomarkers for disease treatment.\textsuperscript{66} Network approaches were used in genetics to identify regulatory pathways from gene expression.\textsuperscript{67} Not limited to genetics, some researchers have applied network approaches to demonstrate that social influences contributing to the development of obesity are as strong as genetic factors.\textsuperscript{68} Leveraging this expertise,\textsuperscript{69} we can apply network medicine methodologies to EHR\textsuperscript{12} to reverse engineer\textsuperscript{72} disease subtypes by integrating various data sources within EHR (eg, laboratory results, medications, visits), and linking them to external sources (eg, PubMed). To achieve this, we can treat each medical entity\textsuperscript{70} 71 as a ‘marker’ for a clinical disease subtype. These markers can then be associated with various diseases or disease severities (eg, chronic, acute), using a high-throughput approach similar to those used in genetics.\textsuperscript{24} 72–74 We can use laboratory values (for laboratory test entities),\textsuperscript{62} 63 dosage level (for medication entities) or the frequency of specialist visits (for specialist entities) and so on. These EHR markers and their expression values are related to typical gene expression data used in genetics studies.\textsuperscript{75} 76 Similar work was performed using the National Health and Nutrition Examination Survey.\textsuperscript{77}

In genetics, network-based approaches were used to attain meaningful results because non-network-based association studies often lacked statistical power to analyze individual genes.\textsuperscript{82} Some network approaches search for hubs of interesting genes within a network,\textsuperscript{83} while others integrate various types of data (protein–protein interactions, gene expression) to find genes considered ‘important’ to the disease of interest.\textsuperscript{69} 85 For clinical disease subtyping using EHR, a network approach would be useful to identify EHR ‘markers’ or combinations of EHR markers associated with a certain disease.

Figure 3 shows how each marker’s expression pattern indicates a certain patient state: presence/absence of diabetes, presence/absence of difficult to manage diabetes and so on. If we integrate the analysis of EHR markers then a distinct diabetes pattern (clinical disease subtype) emerges. However, if markers are analyzed in isolation of each another then the resulting conclusion may differ drastically from the ‘true patient state’. For example, if only hemoglobin A1C is analyzed then it is possible that the patient does not have diabetes (figure 3). However, hemoglobin A1C values can be misleading when the patient is being treated for diabetes, but when EHR markers are integrated then a distinctive diabetes pattern emerges and the likelihood of the patient having diabetes depends on the expression of each marker. Integrating markers\textsuperscript{86} 87 is applicable for both EHR-based phenotyping (eg, to identify that a patient has type 2 diabetes) and for clinical disease subtyping. Importantly, in order for a set of markers (or an integrated pattern of markers) to be considered as a clinical disease subtype then they must be able to stratify the diseased population into a subpopulation with some distinguishing characteristics.\textsuperscript{78}

However, because of the diversity among EHR entities, the type of ‘expression’ must be based on the entity of the marker.\textsuperscript{88} For instance, EHR entities contain Boolean values (eg, presence/absence of International Classification of Disease, revision 9 codes), numerical values (eg, hemoglobin A1C) and nominal values (eg, medication name). A network-based approach would be useful because each marker’s expression pattern would then either increase or decrease the probability that a patient belongs to a certain disease subpopulation (figure 3). Once a novel subtype is characterized, the expression of EHR markers could be used to identify the subtype.

RECOMMENDATIONS

We envision that more effective and personalized disease treatment regimens will be possible when each patient’s genotype, phenotype, and clinical disease subtype information is integrated\textsuperscript{89} to form the patient’s complete ‘verotype’. We posit that a network approach would be useful for integrating genetic and EHR markers because it would reduce the computational complexity\textsuperscript{90} introduced by having multiple EHR markers and multiple genes (polygenic) associated with one disease. When a patient’s true disease subtype is known then clinicians can plan a more effective and personalized treatment plan for that
We break down the concept of ‘EHR phenotyping’ by relating it to definitions defined by Johannsen.1 We relate it to cohort identification and disease subtyping. We also coin a new term, verotype, to group patients who have the same genotype, phenotype, and clinical disease subtype. We recommend using a verotype to develop personalized medical treatment regimens. Finally, we outline the potential for a network-based approach based on the type of marker. Each pattern indicates the most likely patient state based on the marker’s expression level.

HbA1C, hemoglobin A1C.

CONCLUSIONS
We break down the concept of ‘EHR phenotyping’ by relating it to definitions defined by Johannsen.1 We relate it to cohort identification and disease subtyping. We also coin a new term, verotype, to group patients who have the same genotype, phenotype, and clinical disease subtype. We recommend using a patient’s verotype to develop personalized medical treatment regimens. Finally, we outline the potential for a network-based approach to reverse engineer clinical disease subtypes using EHR markers.

Acknowledgements The authors would like to thank: Gregory Hruby, Dražhič Naklji, Silis Jiang, Junfeng Gao, Nicole Weiskopf and Riccardo Miotto for useful discussions on the topic of phenotyping during various laboratory meetings. CW is principal investigator, developed ideas, and wrote the manuscript. GH, YS and WKC provided useful insights, discussion and feedback. CW is principal investigator, developed ideas and wrote the manuscript. GH, YS and WKC provided useful insights, discussion and feedback. CW is principal investigator, developed ideas and wrote the manuscript. GH, YS and WKC provided useful insights, discussion and feedback. CW is principal investigator, developed ideas and wrote the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES
2 Mendel G. Versuche über Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn 1866;4:3–47.
72 Teo YY. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 2008;19:133–43.