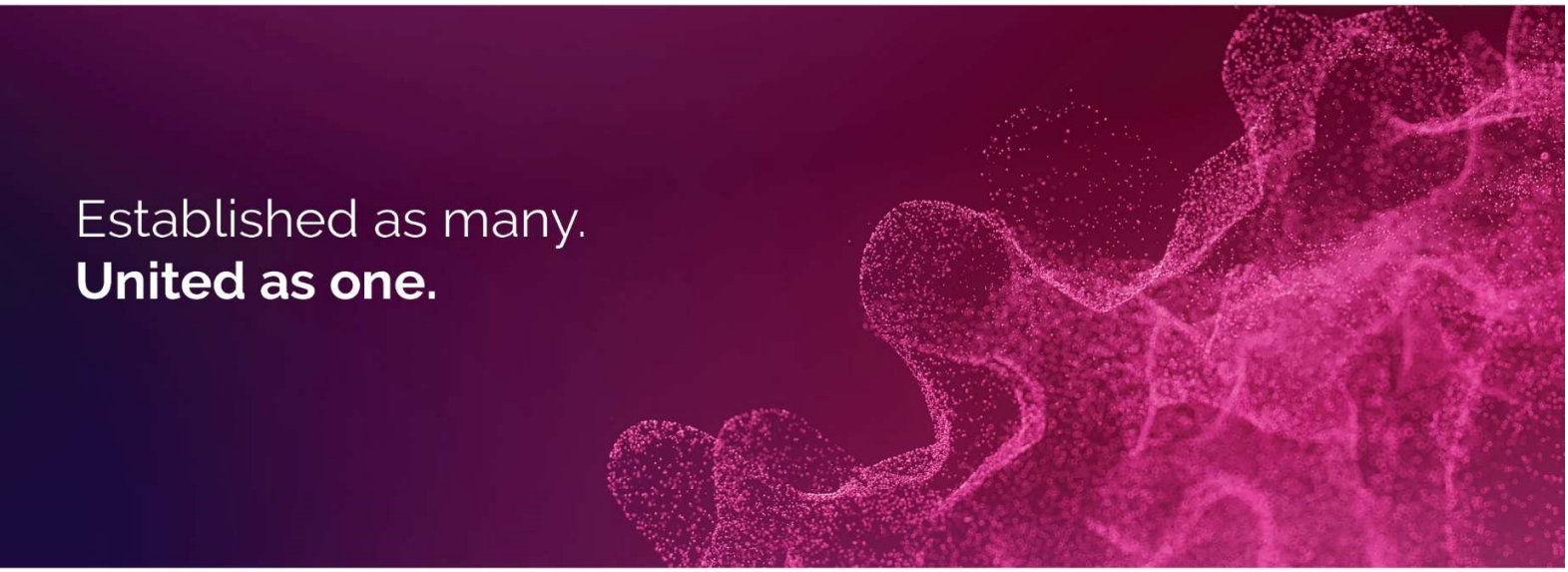
OPEN HEALTH



Development of a predictive model algorithm to identify patients with hypophosphatasia, using primary electronic healthcare records in the United Kingdom - Protocol

DATE	19/10/2021
STATUS	DRAFT V2.0
PREPARED FOR	Alexion Pharmaceuticals 121 Seaport Blvd Boston, MA 02210, United States of America
CHIEF INVESTIGATOR	David Price david@optimumpatientcare.org
PREPARED BY	OPEN Health The Weighbridge, Brewery Courtyard Marlow, SL7 2F, United Kingdom www.openhealthgroup.com

TABLE OF CONTENTS

1 Responsible Parties 3

2 Abstract 5

3 Milestones 7

List of Tables 8

4 List of Figures: 9

List of Abbreviations 10

5 Background & Objectives 11
5.1 Background and Rationale 11
5.2 Research Questions and objectives 13

6 Methods 14
6.1 Study design 14
6.2 Setting 14
6.2.1 Source population 14
6.2.2 Study Period 14
6.2.3 Eligibility Criteria 14
6.2.4 Study design definitions 15
6.3 Variables 16
6.4 Data Sources 16
6.5 Study Size 16
6.6 Data Collection and Management 17
6.7 Data Analysis 18
6.7.1 Machine learning methodology 16
6.7.1 Re-sampling 18
6.7.2 Machine Learning Algorithms 19
6.7.3 Model validation 20
6.7.4 Model application 21
6.7.5 Descriptive Analyses 21
6.7.6 Sensitivity Analyses 21
6.7.7 Missing data 22
6.8 Quality Control 22
6.9 Limitations of the Research Methods 22

7 Protection of Human Subjects and Good Research Practice 24

8 Plans for Disseminating and Communicating Study Results 25

9 Reporting 26

10 References 27

11 Appendix 30
11.1 Proposed Data Extraction Variables **Error! Bookmark not defined.**

1 Responsible Parties

The Study Sponsor has commissioned OPEN Health to develop materials for and coordinate the conduct of the study, including protocol development, ethical and local approval, data collection, analysis and presentation of the results.

Main Author	
Name	Andrew Messali
Title	Associate Director
Degrees	PharmD, PhD
Address	Alexion Pharmaceuticals 121 Seaport Blvd Boston, MA 02210, United States of America
Affiliations	Alexion Pharmaceuticals

Chief Investigator	
Name	David Price
Title	Professor
Degrees	PhD, FRCGP, MRCGP, M.B B.Chir, B.A. (Hons)
Address	5 Coles Lane, Oakington, Cambridge, CB24 3BA david@opri.sg 01223 967 855
Affiliations	University of Aberdeen Observational and Pragmatic Research Institute (OPRI) Optimum Patient Care (OPC)

Investigators	
Name	Dave Heaton
Title	Data Analyst Team Leader
Degrees	BSc
Address	Harvey Walsh Limited, The Heath Business & Technical Park, Runcorn, Cheshire, WA7 4QX
Affiliations	Open Health
Name	Chris Rolfe
Title	RWE Consultant
Degrees	
Address	The Weighbridge, Brewery Courtyard, High Street, Marlow, SL7 2FF
Affiliations	Open Health

Name	Myriam Alexander
Title	Senior Scientific Consultant
Degrees	PhD
Address	The Weighbridge, Brewery Courtyard, Marlow, SL7 2FF
Affiliations	Open Health
Name	Fatemeh Saberi Hosnijeh
Title	Associate Scientific Consultant
Degrees	MD, PhD
Address	Marten Meesweg 107, 3068 AV Rotterdam, The Netherlands
Affiliations	Open Health

Data provider: Optimum Patient Care Record Database
Statistical analysis: Machine Learning
Advisory Committee: ADEPT

2 Abstract

Title: Development of a predictive model algorithm to identify patients with hypophosphatasia, using Optimum Patient Care Record Database in United Kingdom

Rationale and background: Within the setting of the National Health Service (NHS) in the United Kingdom (UK), primary care physicians act as gatekeeper of access to referral for secondary care. Symptom onset of hypophosphatasia (HPP) can be non-specific in the early phases and, as HPP is a rare condition, primary care physicians may not have experience recognising these symptoms and adequately referring patients to secondary care for diagnosis, which then filters back into patients' primary care electronic records via referral letters. As such, definition of a scoring algorithm based on primary care records could help identify patients at an earlier stage of the disease, allow earlier initiation of treatment and minimise long-term sequalae of HPP.

Objective: To develop and validate a scoring algorithm to aid in the diagnosis of HPP by primary care physicians, incorporating symptoms and risk factors recorded in patients' primary care electronic health records prior to HPP diagnosis.

Study design: The study is a retrospective observational case-control study accessing de-identified primary healthcare records from patients enrolled in the Optimum Patient Care Record Database in the UK. The study observation period will start at 1st January 2000 and end on 31st March 2021. HPP cases will be identified based on Read or Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) codes, with index date defined as date of first HPP diagnosis during the study period. Controls will be a random selection of non-HPP patients matched by year of birth/age, gender, date of earliest record of index case and being alive at index date. Controls will be collected with a target ratio of 1 case to 20,000 controls.

Data analysis: The pooled cases and controls will be randomly allocated to a (i) training (75%), or (ii) validating dataset (25%). For patients' electronic health records respectively in the training and validating datasets, predictor variables will include all available data items as Read or SNOMED-CT codes recorded any time prior to index date. A machine learning prediction model (the "scoring algorithm") will be developed in the training dataset and will then be tested in the validating set using statistical measures of accuracy, discrimination, and calibration. The validation step will involve estimating the predicted probability of HPP diagnosis for each control patient and rank-ordering of patients according to their predicted probabilities ("score"). As a next step, at least two clinical experts in HPP will perform a chart review of the top 10% ranked patients and score patients as 'highly likely HPP', 'likely HPP', 'unlikely HPP', 'highly unlikely HPP', 'not HPP', or 'unable to assess'. Based on

this clinical assessment, a threshold for possible or likely HPP will be defined, and the scoring algorithm will be determined.

3 Milestones

Milestones:
Protocol development: 9th November 2021
Data extraction and analyses: 17th December 2021
Interim study report: 21st December 2021
Final study reporting: 22nd April 2022

List of Tables

Table 1: Read and SNOMED CT codes for HPP in OPCRD

4 List of Figures:

Figure 1. Process of developing and validating a risk predictive tool

List of Abbreviations

Abbreviation	Definition
ALP	Alkaline phosphatase
HES	Hospital episode statistics
HPP	Hypophosphatasia
ICD-10	International Classification of Diseases 10th Revision
OPCRD	Optimum Patient Care Record Database
PPi	Pyrophosphate
TNSALP	Tissue-non-specific alkaline phosphatase
OPCS-4	Office of Population Censuses and Surveys Classification of Interventions and Procedures, version 4
HRG	Healthcare Resource Group
odonto-HPP	Odontohypophosphatasia
SNOMED-CT	Systematized Nomenclature of Medicine Clinical Terms

5 Background & Objectives

5.1 Background and Rationale

Hypophosphatasia (HPP) is a rare genetic bone and mineral metabolism disease characterised by low serum alkaline phosphatase (ALP) activity caused by loss-of-function mutations in the gene encoding the tissue-non-specific isoenzyme of ALP (TNSALP) (1,2). Low levels of TNSALP results in the extracellular accumulation of its substrates including calcium and inorganic pyrophosphate (PPi), a potent inhibitor of mineralisation (3). Several skeletal abnormalities and systemic complications such as seizures, kidney damage (nephrocalcinosis), chronic muscle and joint pain, arise from disruption in bone and teeth mineralisation and disordered synthesis of neurotransmitters (3,4). The enzyme replacement therapy in the form of asfotase alfa (Strensiq®), developed by Alexion Pharmaceuticals, has been approved in the European Union, USA, Japan, and Canada for the treatment of perinatal, infantile and childhood HPP (4).

HPP can present in all ages and its severity is generally related to the age of onset; with patients sometimes experiencing symptoms for around 10 years before receiving a diagnosis of HPP (5). Six clinical forms are currently recognised including lethal perinatal, benign perinatal, infantile, childhood, adult, and odontohypophosphatasia (odonto-HPP) (5). The prevalence of HPP, particularly that of its milder forms, has been difficult to estimate. One study reported a prevalence of 1:100,000 live births for severe HPP in Ontario, Canada, while another estimated, using molecular diagnosis, a prevalence of 1:300,000 for severe (perinatal lethal and infantile) HPP in France, extrapolated to a prevalence of 1:6,370 for moderate (other) HPP in Europe (6). Perinatal HPP is almost always lethal near birth, whereas infantile HPP has an estimated 50% mortality during infancy, typically from respiratory complications (7). A recent study conducted in the UK using an algorithm adapted to UK electronic health records estimated point prevalence in 2017 of 0.96 (95%CI 0.71-1.29) per 100,000 population (8).

Due to the extremely low prevalence of the severe forms of hypophosphatasia, its clinical variability and overlapping phenotypic features with several more prevalent conditions, the diagnosis of hypophosphatasia in the clinical setting is challenging. However, its potential lethality and impact on the patient's quality of life, along with the recent availability of an enzyme replacement therapy, increases the relevance of the early and accurate identification of patients affected with hypophosphatasia (1,2).

Over the past decades, the potential for population scale database such as electronic health records (EHRs) database to aid in the identification of potential patients with rare disease has become recognised. Several methodologies have been assessed that include prediction modelling based on a set of predicting variables recorded prior to diagnosis. In particular, the use of machine learning techniques has been advocated, rather than a selection of predictors based on expert knowledge, as an agnostic methodology which optimises the large volume of healthcare data recorded without making any a prior hypothesis on which type of electronic health records (e.g.: symptoms, risk factors) may best support early identification of patients with rare conditions (9–12).

In 2016, the Utah algorithm was designed by US researchers, in a study funded by Alexion Pharmaceuticals, to identify patients with diagnosed and undiagnosed HPP from EHRs in the University of Utah Clinical Enterprise Data Warehouse (13). The algorithm was recently modified with respect to the National Health Service data in the UK to evaluate the epidemiology and burden of illness of HPP patients among Clinical Practice Research Datalink (CPRD) primary care and linked Hospital Episode Statistics (HES) data of UK (see section 11 appendix). The internal validation showed that the UK algorithm has a positive predictive value of 50% (probable) and 78% (probable and possible) where patients had been initially selected by Read code (8). However, there are some limitations to this study. First, the internal validation was based on a small group of patients (n=78) and showed a low predictive value for the method. Second, the validation was based on clinical expert input and therefore symptoms that are not systematically flagged by experts may have been missed, therefore an agnostic approach not presuming on which records are valuable for prediction (using machine learning algorithm) may potentially yield an improved algorithm with better predictive ability. The study was based on data up to 2018, therefore, more historical data are currently available for HPP patients.

Within the setting of the National Health Service (NHS) in the United Kingdom (UK), primary care physicians act as gatekeeper of access to referral for secondary care. Symptom onset of hypophosphatasia (HPP) can be non-specific in the early phases and, as HPP is a rare condition, primary care physicians may not have experience recognising these symptoms and adequately referring patients to secondary care for diagnosis, which then filters back into patients' primary care electronic records via referral letters. As such, definition of a scoring algorithm based on primary care records could help identify patients at an earlier stage of the disease, allow earlier initiation of treatment and minimise long-term sequalae of HPP.

In this context, the aim of this study is to develop and validate a scoring algorithm based on a machine learning prediction model using primary electronic healthcare records from the Optimum Patient Care Record Database (OPCRD) to aid in the diagnosis of HPP in patients with childhood (2-18 years) and adult (≥18 years) HPP.

5.2 Research Questions and objectives

The study aims to address the following research question:
What are the identifiable predictors of HPP, including symptoms and risk factors, that may be identified using a machine learning algorithm in OPCRD?

The objective of the study is:

- To develop and validate a prediction scoring algorithm to aid in the diagnosis of HPP by primary care physicians, incorporating both symptoms and baseline risk factors recorded in patients' primary care electronic health records prior to HPP diagnosis

6 Methods

6.1 Study design

The study is a retrospective observational case-control study accessing de-identified primary healthcare records from patients enrolled in OPCRD in the UK.

A second study will be performed concomitantly from the present study and will access secondary healthcare records using the Hospital Episode Statistics (HES) database; and will be described separately.

6.2 Setting

6.2.1 Source population

The source population is all patients permanently registered in OPCRD during the observation period.

6.2.2 Study Period

The study observation period starts at 1st January 2000 and ends on 31st March 2021 (expected latest date available in OPCRD for data extraction).

6.2.3 Eligibility Criteria

All eligible patients will be enrolled into the study.

6.2.3.1.1. Inclusion criteria

All eligible HPP patients newly identified during the study observation period between 1st January 2000 and 31st March 2021 will be included in the study.

6.2.3.1.2. Exclusion criteria

Patient aged less <2 years old at first HPP diagnosis will be excluded.

As this is a retrospective case-control study designed to develop an algorithm for the detection of undiagnosed HPP patients, the following criteria will further be applied to define a set of "HPP cases" and "non-HPP controls":

- "HPP cases" will be identified based on Read codes or Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) codes (see Table 1).
- "Non-HPP controls" will be a random selection of non-HPP patients with an observation period overlapping that of the case and will be matched by year of birth/age, gender, and date of earliest record and being alive at index date. Considering the rarity of HPP, controls will be matched to cases with a target ratio of 1 case to 20,000 controls.

Table 1. Read and SNOMED CT codes for HPP in OPCRD

SNOMED CT Code	Term
190859005	Hypophosphatasia (disorder)
20756002	Adult hypophosphatasia (disorder)
30174008	Childhood hypophosphatasia (disorder)
55236002	Infantile hypophosphatasia (disorder)
708672004	Odontohypophosphatasia (disorder)
709556009	Periodontitis co-occurrent with hypophosphatasia (disorder)
190860000	Hypophosphatasia rickets (disorder)
Read Code	**Term**
C3530	Hypophosphatasia
C3531	Hypophosphatasia rickets
X40Qk	Adult hypophosphatasia
X40Qi	Infantile hypophosphatasia
X40Qj	Childhood hypophosphatasia

6.2.4 Study design definitions

Index date will be defined as the first SNOMED-CT or Read code record of HPP during the study observation period. The pre-index period will be defined as the time prior to index date, commencing at the most recent of: date of patient's registration into the database and start of the observational period on 1st January 2000.

6.3 Variables

SNOMED-CT UK edition is a UK- adapted international vocabulary for recording patient clinical information and is the terminology used alongside OPCRD for encoding of patient's records (14,15). For each study subject predictor variables will be all Read or SNOMED-CT UK codes present in the OPCRD database in the pre-index period, including general characteristics, clinical features, lab/imaging results, assessments conducted (such as blood pressure, body mass index), prescriptions, referrals to secondary care, and hospital outcomes.

6.4 Data Sources

Cases, controls, and all predictors will be identified from the OPCRD.

The Optimum Patient Care Research Database (OPCRD) is a UK based social enterprise holding UK-sourced clinical data from more than 800 general practices (GPs). OPCRD electronic health records database contains deidentified primary care records of over 12 million patients, representing approximately 18% of the UK population. The data collected includes demographic information, diagnoses, symptoms, treatments, and prescriptions issued, test results and measurements and results taken in the practice and referrals. Both clinical data and therapy data are coded using Read or SNOMED CT codes (drug codes were previously coded using British National Formulary codes).

The NHS Health Research Authority (NHS HRA) has approved OPCRD for clinical research purposes (REC reference: 20/EM/0148). Access to OPCRD data is subject to protocol approval by the Anonymous Data Ethics Protocols and Transparency (ADEPT) committee. Approval is granted to anonymised patient level data for research purposes. To comply with ADEPT, the final protocol will be uploaded on the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) or similar.

6.5 Study Size

HPP is a rare condition and so it is expected that a small number of patients will be eligible to be included as cases, and the large majority of patients will constitute the non-HPP group of controls.

We will maximise the number of cases available by maximising periods of observation and control numbers. For each HPP patient, the target control populations are 20,000 per case.

Hospital Episode Statistics (HES) is a data warehouse containing records of all patients admitted to NHS hospitals in England, with data stored on hospital diagnoses, procedures, treatment, healthcare resource use and associated costs (16). In a feasibility analysis in HES database, currently, there are 4,683 first recorded diagnosed HPP (E833) cases (between April 2010 to June 2021) in Admitted Patient Care or Outpatient setting (Table 2).

Table 2. feasibility analysis; case numbers per age groups between April 2010 and June 2021

Age Band	Patients	% (n=4,683)
<1	54	1.15%
1-4	112	2.39%
5-9	93	1.99%
10-14	91	1.94%
15-19	159	3.40%
20-29	475	10.14%
30-39	584	12.47%
40-49	619	13.22%
50-59	646	13.79%
60-69	702	14.99%
70+	1,136	24.26%
Missing	12	0.26%

6.6 Data Collection and Management

All data analysis and reports presented to Alexion will be aggregated and contain no patient-identifiable data. Following ADEPT approval, Open Health will be given a secure access to OPCRD server with the extracted data and will directly query anonymised patient data on the OPCRD server. Derivation of variables and all data analyses will be performed by an experienced data analyst at OPEN Health in accordance with OPEN Health standard operating procedures.

6.7 Data Analysis

The data analysis will be performed using Python version 3.10.0.

6.7.1 Machine learning methodology

A machine learning methodology will be applied to the database to enable the development of a risk prediction tool without making a prior clinical hypothesis on relevant symptoms or risk factors for HPP. An overview of the process is presented in Figure 1.

Machine learning makes predictions from complex data through inductive inference rather than classical statistical models (17). In machine learning, it is assumed that the "machine" is able to learn the properties of a given dataset with *m* samples (i.e. observations, rows) and *n* features (i.e. independent variables, predictors, columns), and then apply these properties to a new dataset with the same features. The learning process is termed training and is achieved with a training dataset, whereas the subsequent application of the learned properties in a new dataset is termed validation and is achieved with a testing set (Figure 1) (17,18). In addition to model development, machine learning can select a subset of predictors to achieve the best possible performance, called model selection (19). To detect undiagnosed HPP patients, an Python-based machine learning framework will be employed for model development and model selection (20). The Boruta algorithm will be used to identify a smaller set of the most prognostic features. Boruta is a feature selection algorithm with a statistical foundation not necessitating human input (21).

6.7.1 Re-sampling

In this study, cases and controls will be randomly allocated to a (i) training (75%), or (ii) validating dataset (25%).

Outcome imbalance is commonly encountered in healthcare datasets (i.e. the outcome of interest is a rare event). Training based on a small number of events is likely to generate a model with poor accuracy compared to training based on frequent events (e.g. an outcome that is close to 50% probability). Specifically, machine learning algorithms tend to produce a trivial model with all negative prediction in order to reach the highest accuracy. To tackle this problem due to outcome imbalance, two different sampling methods, over-sampling (over-sample the positive cases) and under-sampling (under-sample the negative controls) as well as class weight approach within the algorithms, may be applied to the training set to produce a balanced sample (22,23).

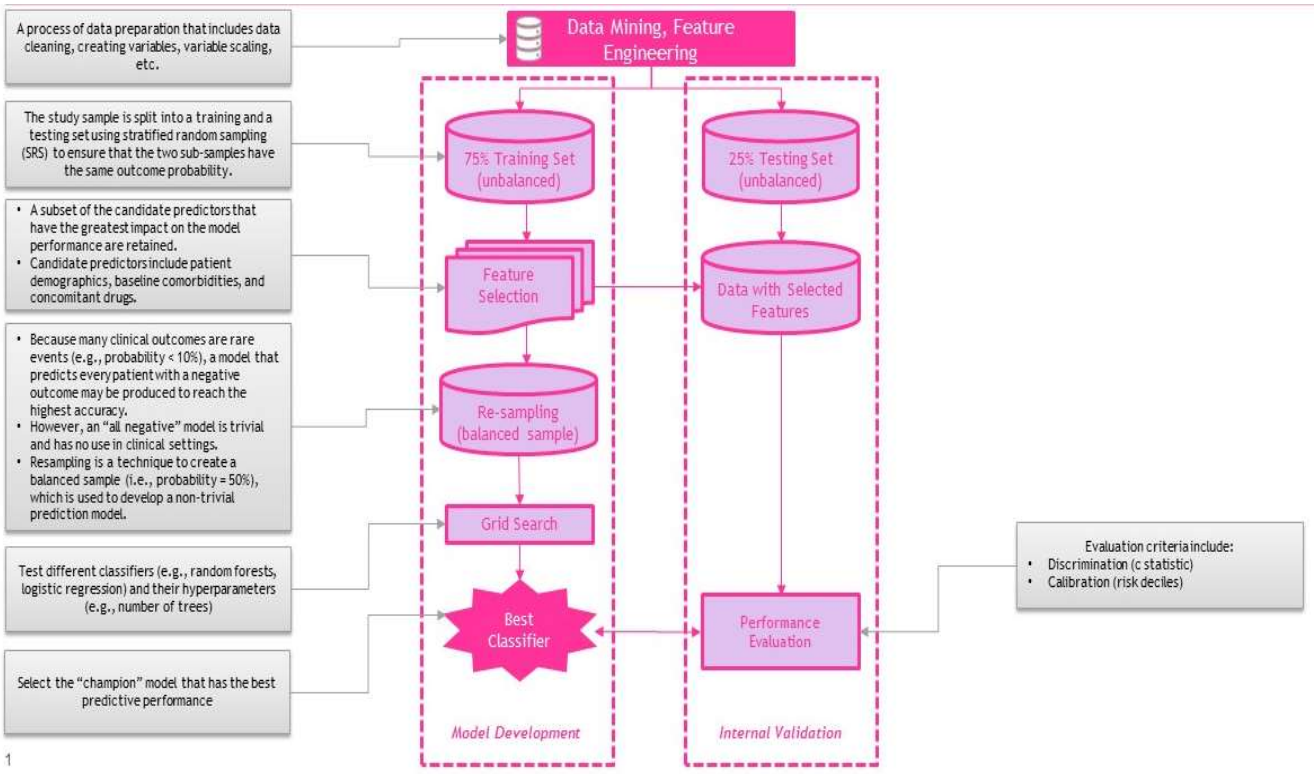
6.7.2 Machine Learning Algorithms

For patients' electronic health records respectively in the training and validating datasets, predictor variables will include all available data items coded in Read or SNOMED CT and recorded any time prior to index date.

A machine learning prediction model will be developed in the training dataset using cases and controls and will be tested in the validating set.

Three machine learning algorithms will be implemented to identify the best performing classifier: random forest (RF) (24), Light Gradient Boosting Machine (LightGBM) (25) and Extreme Gradient Boosting (XGBoost) (26). The class to predict will be a flag of HPP diagnosis. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimisation of an arbitrary differentiable loss function. LightGBM is an open-source implementation of gradient boosting designed to be efficient and more effective than other implementations. XGBoost, another implementation of gradient boosting concept, uses a more regularised model formalisation to control over-fitting, which gives it better performance.

Figure 1. Process of developing and validating a risk predictive tool



6.7.3 Model validation

The prediction model will be internally validated using statistical measures of accuracy, discrimination, and calibration. The objective of the internal validation is to evaluate the quality of the derived prediction tool using a testing set. A full validation of the prediction tool will be conducted following the steps described below.

Accuracy: accuracy measurements will include the sensitivity (true positive), specificity (true negative), positive predictive value, and negative predictive value.

Discrimination: The ability of the derived predictive tool to discriminate between cases and non-cases will be estimated by measuring the area under the receiver

operating characteristic curve (AUC). The value of the AUC (c-statistic) represents the probability that a randomly chosen case is correctly predicted with greater risk score than a randomly chosen non-case (27). The cut-off of an accepted AUC for reasonable and strong discrimination are 0.7 and 0.8, respectively (28).

Calibration: Calibration refers to the agreement between observed outcomes and predicted outcomes. Hosmer-Lemeshow goodness-of-fit test is the most popular technique to evaluate the calibration for a risk prediction model (29). Hosmer-Lemeshow test assigns study subjects to risk strata, typically deciles, based on each subject's predicted probability of an outcome event. The risk strata can then be used to implement tailored treatment interventions to improve health outcomes for patients with different risks.

The validation step will involve estimating the predicted probability of HPP diagnosis for each patient in the validating set and rank-order patients according to their predicted probabilities.

As a next step, at least two clinical experts in HPP will perform a chart review of the top ranked patients (10%) and score patients as 'highly likely HPP', 'likely HPP', 'unlikely HPP', 'highly unlikely HPP', 'not HPP', or 'unable to assess' based on their expert clinical knowledge and according to current UK guidelines.

6.7.4 Model application

Based on this clinical assessment, a threshold for likely HPP will be defined, and the scoring algorithm will be determined.

6.7.5 Descriptive Analyses

General characteristics (age at index date, age of HPP symptom onset, gender, ethnicity) of the cases and controls in the training and validation sets will be described using mean (standard deviation) for continuous variables, or numbers (percentages) for categorical variables.

6.7.6 Sensitivity Analyses

The model performance will be tested in validation sets for different age group: 2-18 years old, ≥18 years old.

6.7.7 Missing data

Missing data will not be substituted.

6.8 Quality Control

All data underwent quality control at OPCRD. The NHS Health Research Authority (NHS HRA) has approved OPCRD for clinical research purposes (REC reference: 20/EM/0148) (30). The Anonymous Data Ethics Protocols and Transparency (ADEPT) committee, an independent body of experts and regulators, has been commissioned by the Respiratory Effectiveness Group (REG) to govern the standards of research conducted on internationally recognised databases, including OPCRD. The committee comprises scientists with statistical and epidemiological experience, members with specific OPCRD-related expertise, independent clinical experts and lay members adhering to UK standards. Any research project conducted on OPCRD data needs to be reviewed and ethically approved by the ADEPT committee prior to any data being accessed. The ADEPT committee will be responsible for reviewing applicant study protocols for scientific quality.

The generated algorithm will be validated with clinical experts with relevant experience in the diagnosis and identification of HPP in all age groups included in the study.

6.9 Limitations of the Research Methods

A limitation of studies in rare diseases is the low numbers available for research, even when screening patients from large datasets such as OPCRD. For this reason, a relatively wide study observation period has been included, however confidence intervals in the predictor models and descriptive analyses may be wide.

Diagnosis of HPP sometimes remains difficult to establish and therefore some patients may not or not yet have received a confirmed HPP diagnosis during the study period.

The model will be developed using OPCRD, a UK database, and the scoring algorithm may not necessarily be generalisable to other countries, with different healthcare systems. However, to optimise generalisability, the algorithm will be developed in two stages, first being developed a training set and then being tested on a validation

set. However, to further enhance generalisability, future studies would need to validate the algorithm using alternative electronic healthcare records from other databases or other countries (external validation).

The datasets represent information collected for clinical and routine use rather than specifically for research purposes, and therefore information recorded may not necessarily be complete (e.g.: laboratory variables, health indicators such as weight and smoking status).

The dataset will be using primary electronic healthcare records from OPCRD, however information collected in secondary healthcare settings may be a more valuable source of important predictors for the diagnosis of HPP. For this reason, concomitantly to this study, a separate study will be run using the HES database in England.

Validity and completeness of individual patient records cannot be assessed due to the nature of electronic health records data.

Finally, we will use the clinical data collected between 2000 and 2021. As standard medical practice evolves over time, predictive accuracy of the model may be affected.

7 Protection of Human Subjects and Good Research Practice

This study will comply with all applicable laws, regulations, and guidance regarding patient protection including patient privacy, and consistent with the ethical principles of the Declaration of Helsinki (31) and the requirements of the European Union General Data Protection Regulation (GDPR) (32).

This study has been designed and will be conducted according to the requirements of EncePP (33) and International Society for Pharmacoepidemiology (ISPE) (34) guidance for Good Pharmacoepidemiology Practices, as appropriate.

Permission for the current study will be requested from The Anonymous Data Ethics Protocols and Transparency committee (ADEPT).

8 Plans for Disseminating and Communicating Study Results

All research using OPCRD must be registered on established study databases such as the ENCePP or similar as a requirement of ADEPT.

There is an intention to present the results of the study at scientific conferences and to publish it in peer reviewed journals. Publications will be developed according to Alexion policies and authorship will be determined in accordance with the International Committee of Medical Journal Editors (ICMJE) guidelines.

9 Reporting

An interim study report will be prepared containing the first set of analysis of data available by 21st December 2021. A full study report will be prepared once all data analysis is complete.

10 References

1. Bianchi ML. Hypophosphatasia: an overview of the disease and its treatment. Osteoporos Int J Establ Result Coop Eur Found Osteoporos Natl Osteoporos Found USA. 2015 Dec;26(12):2743–57.
2. Seefried L, Dahir K, Petryk A, Högler W, Linglart A, Martos-Moreno GÁ, et al. Burden of Illness in Adults With Hypophosphatasia: Data From the Global Hypophosphatasia Patient Registry. J Bone Miner Res Off J Am Soc Bone Miner Res. 2020 Nov;35(11):2171–8.
3. Whyte MP. Hypophosphatasia: An overview For 2017. Bone. 2017 Sep;102:15–25.
4. Genest F, Rak D, Petryk A, Seefried L. Physical Function and Health-Related Quality of Life in Adults Treated With Asfotase Alfa for Pediatric-Onset Hypophosphatasia. JBMR Plus. 2020;4(9):e10395.
5. Bangura A, Wright L, Shuler T. Hypophosphatasia: Current Literature for Pathophysiology, Clinical Manifestations, Diagnosis, and Treatment. Cureus. 12(6):e8594.
6. Fraser D. Hypophosphatasia. Am J Med. 1957 May;22(5):730–46.
7. Mornet E, Yvard A, Taillandier A, Fauvert D, Simon-Bouy B. A molecular-based estimation of the prevalence of hypophosphatasia in the European population. Ann Hum Genet. 2011 May;75(3):439–45.
8. Jenkins-Jones S. Evaluation of the Utah algorithm for identifying hypophosphatasia to estimate prevalence in the United Kingdom. Proceedings of BRS 2017; 2017; Journal of Musculoskeletal and Neuronal Interactions (JMNI).
9. Colbaugh R, Glass K, Rudolf C, Tremblay Volv Global, Lausanne, Switzerland M. Learning to Identify Rare Disease Patients from Electronic Health Records. AMIA Annu Symp Proc. 2018 Dec 5;2018:340–7.
10. Sun AZ, Shu Y-H, Harrison TN, Hever A, Jacobsen SJ, O'Shaughnessy MM, et al. Identifying Patients with Rare Disease Using Electronic Health Record Data: The Kaiser Permanente Southern California Membranous Nephropathy Cohort. Perm J. 2020 Feb 7;24:19.126.
11. Doyle OM, van der Laan R, Obradovic M, McMahon P, Daniels F, Pitcher A, et al. Identification of potentially undiagnosed patients with nontuberculous mycobacterial lung disease using machine learning applied to primary care data in the UK. Eur Respir J. 2020 Oct;56(4):2000045.
12. Gruber S, Krakower D, Menchaca JT, Hsu K, Hawrusik R, Maro JC, et al. Using electronic health records to identify candidates for human immunodeficiency virus pre-exposure prophylaxis: An application of super learning to risk prediction when the outcome is rare. Stat Med. 2020 Oct 15;39(23):3059–73.
13. Biskupiak J, Sainski A, Yoo M, Brixner D, Iloeje U. Utilization of an algorithm to identify individuals at risk for hypophosphatasia (HPP) within an electronic health record (EHR) database [abstract FR0329]. 31st Annual Meeting of the American Society for Bone and Mineral Research; 2016 Sep 16; Atlanta, GA, USA.
14. Wardle M, Spencer A. Implementation of SNOMED CT in an online clinical database. Future Healthc J. 2017 Jun;4(2):126–30.

15. SNOMED CT [Internet]. NHS Digital. [cited 2021 Oct 8]. Available from: https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct
16. Thorn JC, Turner E, Hounsome L, Walsh E, Donovan JL, Verne J, et al. Validation of the Hospital Episode Statistics Outpatient Dataset in England. PharmacoEconomics. 2016 Feb;34(2):161–8.
17. Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.
18. Song X, Mitnitski A, Cox J, Rockwood K. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. In IOS Press; 2004. p. 736–40.
19. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;97(1–2):273–324.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
21. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36(11):1–13.
22. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57.
23. Weiss GM, McCarthy K, Zabar B. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? :7.
24. randomforest2001.pdf [Internet]. [cited 2021 Oct 12]. Available from: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf
25. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–57. (NIPS'17).
26. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2016 [cited 2021 Oct 12]. p. 785–94. (KDD '16). Available from: https://doi.org/10.1145/2939672.2939785
27. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
28. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928–35.
29. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997;16(9):965–80.
30. Optimum Patient Care. Optimum Patient Care Research Database.https://opcrd.co.uk/. Accessed October 7, 2021.
31. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Jama. 2013;310(20):2191–4.
32. General Data Protection Regulation (GDPR) – Official Legal Text [Internet]. General Data Protection Regulation (GDPR). [cited 2021 Oct 6]. Available from: https://gdpr-info.eu/
33. ENCePP Home Page [Internet]. [cited 2021 Oct 6]. Available from: http://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml

34. Guidelines for Good Pharmacoepidemiology Practices (GPP) - International Society for Pharmacoepidemiology [Internet]. [cited 2021 Oct 6]. Available from: https://www.pharmacoepi.org/resources/policies/guidelines-08027/

11 Appendix

Table 1 | Utah algorithm criteria for selecting patients with hypophosphatasia and their applicability to UK electronic health records in the Clinical Practice Research Datalink (8)

Original Utah algorithm criterion (13)	**Utah criterion as modified for UK EHRs** (8)
One or more diagnostic record with the ICD-9 code 275.3 (‘Disorders of phosphorus metabolism’)	One or more diagnostic record with the ICD-10 code E83.3 ('Disorders of phosphorus metabolism and phosphatases') in secondary care data, or with the Read Code C353000 ('Hypophosphatasia') or C353100 ('Hypophosphatasia rickets') in primary care data
AND/OR	
Two or more 2 low age- and sex-adjusted alkaline phosphatase (ALP) test results, with no normal test results and no exposure to bisphosphonates	≥ 2 low alkaline phosphatase (ALP) test results, with no normal test results and no exposure to bisphosphonate; all as recorded in primary care data
AND	
Clinical, biochemical, histological or radiographic evidence of at least one manifestation of HPP: seizures; respiratory failure in children < 5 years old; elevated serum pyridoxal-5'-phosphate or inorganic pyrophosphate (PPi); elevated urine phosphoethanolamine; family history of HPP; radiographic evidence of hypomineralization or histological evidence of osteomalacia; history of or treatment for non-traumatic fractures; premature tooth loss; craniosynostosis; multiple fractures; rickets	Clinical recording of at least one manifestation of HPP: seizures; respiratory failure in children < 5 years old; family history of HPP; hypomineralization or osteomalacia; history of or treatment for non-traumatic fractures; premature tooth loss; craniosynostosis; multiple fractures; rickets
AND NO	
Diagnosis of hypothyroidism	Diagnosis of hypothyroidism
AND	
ALP activity that is all low	ALP activity that is all low

CPRD, Clinical Practice Research Datalink; EHRs, electronic health records