2.3 - Data Description

Challenge data comes from 7 related viral challenge trials, representing 4 different respiratory viruses. The challenges are *DEE1 RSV*, *DEE2 H3N2*, *DEE3 H1N1*, *DEE4X H1N1*, *DEE5 H3N2*, *Rhinovirus Duke*, and *Rhinovirus UVA*. In each of these trials, healthy volunteers were followed for seven to nine days following controlled nasal exposure to one respiratory virus. Subjects enrolled into these viral challenge experiments had to meet several inclusion and exclusion criteria. Among them was an evaluation of pre-existing neutralizing antibodies to the challenge strain. In the case of influenza H3N2 and influenza H1N1, all subjects were screened for such antibodies. Any subject with pre-existing antibodies to the challenge strain was excluded. For the rhinovirus challenge, subjects with a serum neutralizing antibody titer to RV39 > 1:4 at pre-screening were excluded. For the RSV challenges, subjects were pre-screened for neutralizing antibodies although the presence of such antibodies was not an exclusion criterion.
Symptom data and nasal lavage samples were collected from each patient on a repeated basis over the course of 7-9 days. Viral infection was quantified by measuring release of viral particles from nasal passages ("viral shedding") as assessed from nasal lavage samples via qualitative viral culture and/or quantitative influenza RT-PCR.

Symptomatic data was collected through self-report on a repeated basis. Symptoms were assessed via modified Jackson score [1] which assessed the severity of 8 upper respiratory symptoms (runny nose, cough, headache, malaise, myalgia, sneeze, sore throat and stuffy nose) and integrates daily scores over 5-day windows.

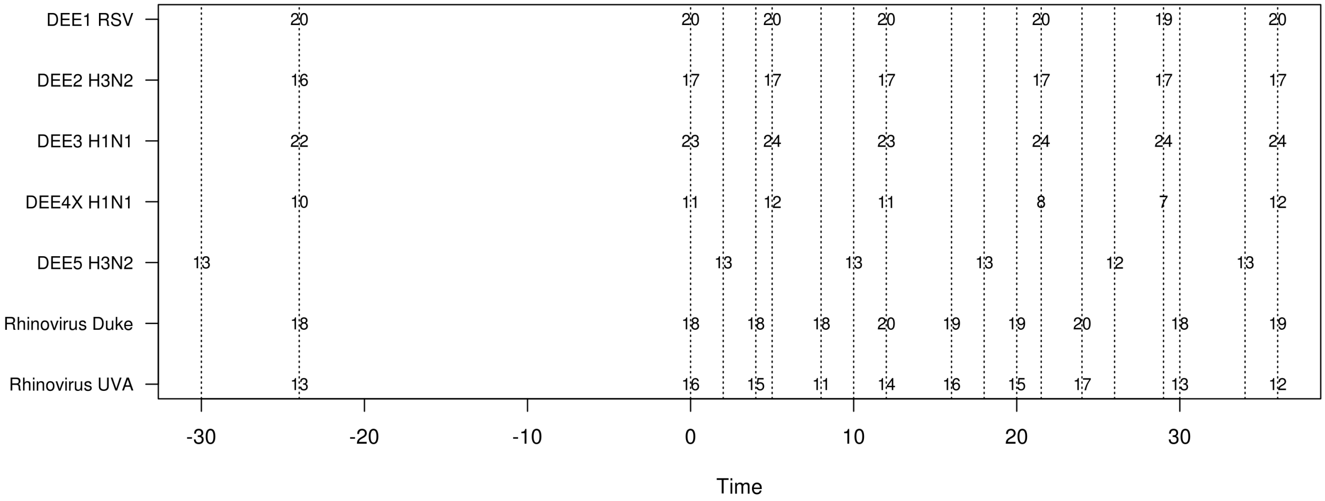
Blood was collected and gene expression of peripheral whole blood was performed 1 day (24 to 30 hours) prior to exposure, immediately prior to exposure, and at regular intervals following exposure. All patients challenged with influenza (H1N1 or H3N2) received oseltamivir 5 days post-exposure. However, 14 (of 21) patients in the DEE5 H3N2 cohort received early treatment (24 hours post-exposure) regardless of symptoms or shedding.

Four of the 7 data sets are publicly available, and test data are held out from the remaining 3 trials: *DEE4X H1N1*, *DEE5 H3N2*, and *Rhinovirus Duke*. *Rhinovirus Duke* additionally includes 7 volunteers who were exposed to sham rather than active virus. Data for these individuals will be provided to participants, but none of the sham-exposed individuals will be included in the test data. Subject counts by trial are as follows:

STUDY	Training	Test	Sham
DEE1 RSV	20	0	NA
DEE2 H3N2	17	0	NA
DEE3 H1N1	24	0	NA

DEE4X H1N1	12	7	NA
DEE5 H3N2	13	8	NA
Rhinovirus Duke	12	8	7
Rhinovirus UVA	20	0	NA
TOTAL	118	23	7

After gene expression QC, the number of gene expression profiles available by study and timepoint in the training data, excluding the samples set aside for the test set, is as follows:



These counts do not include 7 additional sham-exposed subjects in *DEE5 H3N2*. Additional timepoints are made available for the training data representing the full course of the experiment (typically up to 7 days for most studies), but test data will be limited to the early stage gene expression data (i.e. up to time 0, up to 12 hours, up to 24 hours and up to 36 hours, for Phases 1-4, respectively).

Clinical Data

Available Predictors

The available clinical and demographic variables available are Age and Gender, as well as whether the patient received early oseltamivir treatment (*DEE5 H3N2* only) and whether the patient received sham exposure rather than virus (*Rhinovirus Duke* only). Study demographics for virus exposed subjects are summarized below.

Study Virus	N	Age (Years)	% Male	%Early Treatment
DEE1 RSV	20	26.9 ± 6.3	55.0	0

DEE2 H3N2	17	27.4 ± 5.0	52.9	0
DEE3 H1N1	24	25.0 ± 4.5	70.8	0
DEE4X H1N1	19	24.2 ± 4.5	47.4	0
DEE5 H3N2	21	26.0 ± 6.6	57.1	66.7
Duke Rhinovirus	20	27.9 ± 5.7	70.0	0
UVA Rhinovirus	20	20.1 ± 2.3	60.0	0

Outcome Variables

Three outcome variables are available, each of which is to be predicted for one of the three subchallenges. These variables are

- Subchallenge 1: SHEDDING_SC1, a binary variable indicating presence of virus in nasal swab following exposure
- Subchallenge 2: SYMPTOMATIC_SC2, a binary variable indicating post-exposure maximum symptom score >= 6
- Subchallenge 3: LOGSYMPTSCORE_SC3, a continuous variable indicating the log of the maximum symptom score+1

Clinical Data File

A clinical data file is provided which includes clinical/demographic data, outcome variables, and information matching sample IDs to expression data at each timepoint. As such, each patient occurs multiple times in the file because multiple timepoints are represented per patient. The file is of the form

STUDYID	SUBJECTID	AGE	GENDER	EARLYTX	SHAM	SHEDDING_SC1	SYMPTOMATIC_SC2	LOGSYMPTSCORE_SC3	TIMEHOURS
DEE1 RSV	RSV012	24	Female	NA	NA	1	1	1.07918124604762	-24
DEE1 RSV	RSV019	21	Male	NA	NA	0	0	0	5
DEE1 RSV	RSV018	22	Male	NA	NA	1	1	0.903089986991944	5

where the variables are defined as

Variable Name	Variable Description
STUDYID	Study name
SUBJECTID	Unique patient ID
AGE	Patient age
GENDER	Patient gender (Male or Female)
EARLYTX	=1 if patient received early oseltamivir, =0 if patient received oseltamivir at day 5, or NA for cohorts other than *DEE5 H3N2*
SHAM	= sham if patient received sham exposure, NA otherwise

SHEDDING_SC1	=1 if patient exhibited viral shedding, =0 if viral shedding not observed
SYMPTOMATIC_SC2	=1 if max symptom score >=6, =0 if max symptom score < 6
LOGSYMPTSCORE_SC3	$\log_{10}$(max symptom score +1)
TIMEHOURS	Time of gene expression profile (hours)
SAMPLEID	Gene expression sample ID
CEL	CEL file name

Granular Symptom Data

For the training data, we have provided the daily symptom scores for each of the 8 individual symptoms which are used in computing the modified Jackson score (which is transformed to compute the LOGSYMPTSCORE_SC3 outcome). These are provided in order to give more granularity as to the particular symptoms exhibited by the subjects. Because these are aspects of the outcomes to be predicted, they will not be provided for the test data, and should not be used directly in predictive models.

The granular symptom data file, ViralChallenge_training_SymptomScoresByDay.tsv, is of the form

STUDYID	SUBJECTID	STUDYDAY	SX_RUNNYNOSE	SX_COUGH	SX_HEADACHE	SX_MALAISE	SX_MYALGIA	SX_SNEEZE	SX_SORTHROAT
Rhinovirus UVA	1	-1	0	0	0	0	0	0	0
Rhinovirus UVA	1	0	0	0	0	0	0	0	1
Rhinovirus UVA	1	1	1	0	0	0	0	1	1
Rhinovirus UVA	1	2	1	0	0	0	0	0	1

where the variables are defined as

Variable Name	Variable Description
STUDYID	Study name
SUBJECTID	Unique patient ID
STUDYDAY	Day for which the symptoms are reported
SX_RUNNYNOSE	Runny nose severity 0-4
SX_COUGH	Cough severity 0-4
SX_HEADACHE	Headache severity 0-4
SX_MALAISE	Malaise severity 0-4
SX_MYALGIA	Myalgia severity 0-4
SX_SNEEZE	Sneezing severity 0-4
SX_SORETHROAT	Sore throat severity 0-4
SX_STUFFYNOSE	Stuffy nose severity 0-4

Expression Data

Gene expression profiling was performed on the Affy Human Genome U133A 2.0 array. Both a raw and normalized version of the gene expression data are available for use in this challenge. Both versions contain only profiles that pass QC metrics including those for RNA Degradation, scale factors, percent genes present, β-actin 3’ to 5’ ratio and GAPDH 3’ to 5’ ratio.

The raw data are available as Affy CEL files, one per individual per timepoint.

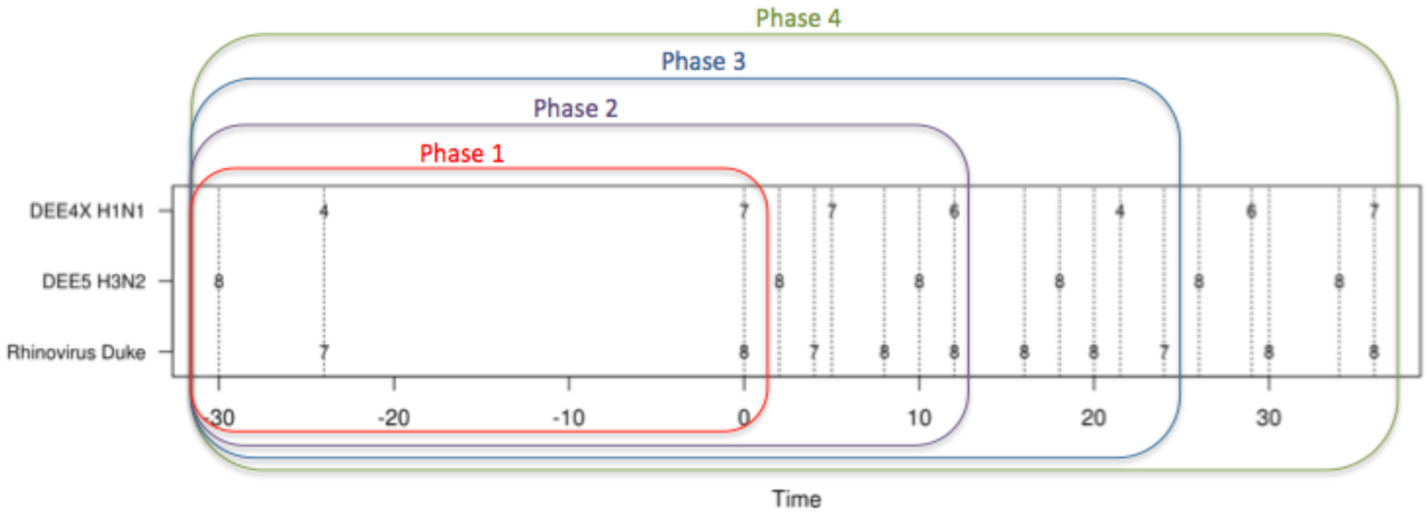
The normalized data have been RMA adjusted and are available as a single data tab-separated text file whose top left corner looks like this:

FEATUREID	1590_266948_HG-U133A_2_41828_DU08-02S09371.CEL	1590_266969_HG-U133A_2_41972_DU08-02S14579.CEL	1590_91213_H133A2_22274_DU08-02S14613_T2.CEL
1007_s_at	7.72812962392557	7.65465729424535	8.11790252209913
1053_at	6.96088502681371	7.19618433790847	7.20658724064259
117_at	9.61813647199283	10.2417665144579	11.1358903269326
121_at	8.48633919412534	8.5095025409127	8.47653190385772

Here the first row is the sample CEL file name (for all but the first entry), which can be matched to patient through the clinical data file. The first column is the probeset ID (FEATUREID), which is annotated in HG-U133A_2.na35.annot.csv, alternate annotations may be available through Bioconductor or other sources. The numerical entries represent the normalized log$_2$-transformed expression by sample and probeset.

Test Data

Test data will be released according to phase. The number of samples by timepoint are shown below.



Data Downloads

Training data files are available in the following locations:

Data Type	Version	Filename	Location
Clinical		ViralChallenge_training_CLINICAL.tsv	syn6043449
Granular Symptoms		ViralChallenge_training_SymptomScoresByDay.tsv	syn6043450
Expression	Raw	ViralChallenge_training_EXPRESSION_CEL.tar.gz	syn6043347
Expression	Normalized	ViralChallenge_training_EXPRESSION_RMA.tsv	syn6043448
Annotations	Affymetrix	HG-U133A_2.na35.annot.csv.zip	syn5684262

YOU MUST BE A REGISTERED CHALLENGE PARTICIPANT TO ACCESS THESE DATA.

[1]
Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, et al. (2008) Time lines of infection and disease in human influenza: a review of volunteer challenge studies. Am J Epidemiol 167: 775–785.