**Data Input specifications**

Table of Contents

[1. About This Guide 2](#_Toc146704638)

[1.1. General 2](#_Toc146704639)

[1.2. Important Notice 2](#_Toc146704640)

[2. Input file format 3](#_Toc146704641)

[2.1. Types of Signals (Inputs) 3](#_Toc146704642)

[2.2. Input File 3](#_Toc146704643)

[2.2.1. Organizing Input Files 3](#_Toc146704644)

[2.2.2. Input File Format 3](#_Toc146704645)

[3. Appendix A 5](#_Toc146704646)

[3.1. Input File Examples 5](#_Toc146704647)

1. About This Guide
   1. General

This document was created to allow a retrospective validation of AlgoMarker on deidentified sample data coming from medical records of individuals treated by certain Health System. The document contains detailed instructions on the data to be extracted and how to format it so that tools that are created by Medial EarlySign could read the data and analyze the input characteristics and later on also the performance of the model.

* 1. Important Notice

Copyright © 2023 by Medial EarlySign Ltd. All rights reserved.

All intellectual property rights in this publication are owned by Medial EarlySign Ltd. and protected by applicable copyright laws and international treaty provisions. Medial EarlySign Ltd. retains all rights not expressly granted. No part of this publication may be reproduced in any form whatsoever or used to make any derivative work without prior written approval by Medial EarlySign Ltd.

Medial EarlySign Ltd. reserves the right to revise this publication and/or make improvements or changes in the prod­uct(s) and/or the program(s) described in this documentation at any time without prior notice. The information in this document is provided in good faith. However, it is provided without any warranty that it is accurate or complete and on the express understanding that Medial EarlySign Ltd. shall have no liability whatsoever to other parties in any way arising from or relating to the information or its use.

Any product described in this publication is furnished under a license agreement.

All trademarks or registered trademarks are the property of their respective owners.

1. Input file format

Each AlgoMarker has different requirements for mandatory and additional input signals.

Please refer to the specific AlgoMarker guide to get list of relevant input signals with their type.

* 1. Types of Signals (Inputs)

Signals are data input parameters sent as part of a request to the AlgoMarker, and then used by the model. These may include, for example, various individual person’s clinical or lab data, demographics, events, and more.

Each signal is defined by two components: list of timestamps and values.

A Signal type is defined by a fixed number of time channels and a fixed number of value channels. For example, Hemoglobin has 1 time channel of type numeric integer and 1 value channel of type numeric float. Gender has only 1 value channel of type categorical, and 0 time channels.

Each value channel can be either:

1. **Numeric** – float or integer. The channel may be associated with a measurement unit. Numeric channels may be received using a different measurement unit, and then go through a pre-defined conversion to the default measurement unit of the signal.
2. **Categorical Text** – The channel belongs to a predefined list of categories, defined by text labels. We have internal mappings of those categories into medical concepts the model uses. For example, we receive ICD-9 and ICD-10 codes with textual prefix. More details appear in the signals file comments column.

* 1. Input File
     1. Organizing Input Files

Input data can be arranged into one or more files. Each file may contain different signals, and each signal may be present in more than 1 file. Also, a patient can appear in more than 1 file. All Input files should be uploaded to the same Input subfolder.

Once the ‘Execute script’ is activated, all input files located in the Input folder will be treated. Consequently, data relating to a single patient may be distributed between multiple files.

For file examples, please refer to **Appendix**.

* + 1. Input File Format
* Files should be in encoded UTF-8
* Lines are delimited by an EOL (end-of-line) character.
* Each line in a file represents a single data point - a single patient data spans across several lines.
* First line in each file is a header, and should exactly include the following 5 labels, separated by a TAB character:

|  |  |
| --- | --- |
| **Field #** | **Name** |
| 1 | ID |
| 2 | Date |
| 3 | Signal |
| 4 | Value |
| 5 | Unit |

* Any subsequent line should include the following 5 fields, separated by a TAB character, in this order:

|  |  |  |  |
| --- | --- | --- | --- |
| **Field #** | **Name** | **Description** | **Restrictions** |
| 1 | ID | The unique identifier of the patient. | Length between 1 and 40.  Valid characters: a-z A-Z 0-9 \_ - . ^ \* / % # |
| 2 | Date | A comma-separated list of zero, one or more timestamps.  E.g. a GENDER signal will have no timestamp in its data object. A GLUCOSE signal will have a single timestamp in each data object. An Admission signal will have two timestamps separated by a comma in each data object. | Each timestamp represents a valid date, formatted yyyy-MM-dd, e.g. 2021-08-12  In case of no timestamp, an empty string should be used |
| 3 | Signal | A signal name, as defined in the AlgoMarker list of signals. In the lungflag.signals.xlsx | Part of the list of signals |
| 4 | Value | The signal’s value – textual representation (at the specific time given above).  A special character ‘|’ is used as a seprator for signals that required multiple values. For example, Blood Pressure that is represented as 120|80 for the systolic, diastolic values | A valid numeric/float number if the channel type is numeric. When it’s categorical, it will be a string. |
| 5 | Unit | The signal’s unit if applicable, as defined in the lungflag.signals.xlsx | In case of no unit, an empty string should be used |

a

1. Appendix A

This appendix provides examples of the ‘File API’ input and output files.

* 1. Input File Examples

Scoring Input File

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Date | Signal | Value | Unit |
| 82437 |  | BDATE | 19820101 |  |
| 82437 |  | GENDER | F |  |
| 82437 | 2017-08-01 | Basophils# | 0.1 | 10^9/l |
| 82437 | 2017-08-01 | Eosinophils# | 0.1 | 10^9/l |
| 82437 | 2017-08-01 | Hemoglobin | 12.7 | g/dl |
| 82437 | 2017-08-01 | Hematocrit | 40 | % |
| 82437 | 2018-09-01 | DIAGNOSIS | ICD9\_CODE:496 |  |
| 82437 |  | MEMBERSHIP | 20170101,20190916 |  |
| 67531 |  | BDATE | 19510101 |  |
| 67531 |  | GENDER | F |  |
| 67531 |  | MEMBERSHIP | 20171201,20181231 |  |
| 67531 | 2017-12-11 | RBC | 4.62 | 10^12/l |
| 67531 | 2017-12-11 | RDW | 16.2 | % |
| 67531 | 2017-12-11 | WBC | 9.2 | 10^9/l |
| 67531 | 2017-09-25 | Basophils# | 0.1 | 10^9/l |
| 67531 | 2017-09-25 | Eosinophils# | 0.3 | 10^9/l |
| 67531 | 2017-09-25 | Platelets | 168 | 10^9/l |
| 67531 | 2017-09-25 | RBC | 4.85 | 10^12/l |
| 67531 | 2017-09-25 | RDW | 13.4 | % |
| 67531 | 2017-09-25 | WBC | 11.2 | 10^9/l |
| 67531 | 2018-12-04 | Basophils# | 0.1 | 10^9/l |
| 67531 | 2018-12-04 | Eosinophils# | 0.2 | 10^9/l |
| 67531 | 2018-12-04 | RBC | 4.61 | 10^12/l |
| 67531 | 2018-12-04 | RDW | 13 | % |
| 67531 | 2018-12-04 | WBC | 9.2 | 10^9/l |
| 67531 | 2018-06-08 | Basophils# | 0.1 | 10^9/l |
| 67531 | 2018-06-08 | Eosinophils# | 0.3 | 10^9/l |
| 67531 | 2018-06-08 | Hemoglobin | 13.8 | g/dl |
| 67531 | 2018-06-08 | Hematocrit | 40 | % |
| 67531 | 2018-06-08 | Lymphocytes# | 2.5 | 10^9/l |
| 67531 | 2018-06-08 | MCH | 30.6 | pg |

Performance Analysis Input File – can be plugged together with scoring input file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Date | Signal | Value | Unit |
| 82437 | 2018-09-01 | DIAGNOSIS | ICD9\_CODE:162.2 |  |
| 82437 | 2018-09-01 | Cancer\_Stage | II |  |