

Supplementary Materials

A Semi-automated Term Harmonization Pipeline Applied to Pulmonary Arterial Hypertension Clinical Trials

Mapper Instructions

Below we provide a copy of the instructions given to our two “first pass” term mappers as well as to the “second pass” attending physician mapper for harmonization of medical history (MH) terms.

LLT Fuzzy Matching Instructions

NOTE: To maximize reproducibility and facilitate downstream automation, please read and follow these instructions. Please do not modify the working fuzzy mapping files further (other than using sort on columns). In particular, do not transform all text into CAPS, or change the name of column headers.

Background: We are harmonizing the lowest level term (LLT) of MH terms using the MedDRA versus 21 standard (as we did previously with adverse event [AE]) terms. At this point we have cleaned and formatted the original MH data and completed exact matching on the MH terms. Out of a total of 28,720 unique MH rows, all but 7,115 were successfully matched exactly. Fuzzy matching (with fuzzywuzzy) has been run on these remaining 7,115. Fuzzywuzzy outputs the top five matches using any of its built-in fuzzy matching strategies. For MH terms we have available three sources of terms available for attempting to harmonize MH LLTs to the MedDRA standard. These included (MHTERM), available for all rows, LLT_NAME, and MHMODIFY. Fuzzy matching was thus applied to each (whenever available). Therefore, in the fuzzy matching files, there are five columns giving the candidate fuzzy matches for each of these three term sources.

Manual Fuzzy mapping instructions:

- Go through each row in your assigned mapping file and manually enter values (as described below) to indicate the “best” LLT match for each.
- During this process, only ever enter/modify values in the following columns:
 - *FZMatch_Choice_ID_MHTERM (yellow)—If the best match is available among the five (MHTERM) fuzzy matches, enter the integer (1–5) for the corresponding column where that best match term is located, i.e., *FZMatch_1_MHTERM = 1, FZMatch_2_MHTERM = 2, FZMatch_3_MHTERM = 3, FZMatch_4_MHTERM = 4, FZMatch_5_MHTERM = 5* (leave cell blank otherwise).
 - *FZMatch_Copied_Term (orange)—If an integer was entered in FZMatch_Choice_ID_MHTERM leave this cell blank. Otherwise COPY AND PASTE the exact, best matching text MedDRA standard term/phrase into this cell. This may either be copied from one of the other two sets of fuzzy matching results (i.e., MHMODIFY, or LLT_NAME), or taken directly from the MedDRA LLT standard term file, if there is a better match not identified with fuzzy matching. Please make sure only terms where currency = Y are taken from the LLT standard file.
 - *LLT_map_code (gray)—This is where the “mapping quality is recorded.” We will be using our global term match quality codes below:
 - 0 = Exact Match - DL1_FT1 and DL1_FT2 both match (i.e., MHTERM = LLT_NAME column)
 - 1 = Exact Match - DL1_FT1 match (i.e., MHTERM match)
 - 2 = Exact Match - DL1_FT2 match (i.e., LLT_NAME match)
 - 3 = Exact Match - DL1_FT3 match (i.e., MHMODIFY match)
 - 4 = Fuzzy Matching (High Confidence)
 - 5 = Fuzzy Matching (Medium Confidence)
 - Blank = Unmapped/NA

During this fuzzy matching process, this cell should be filled in for every row! Only the values 4, 5, or 6 should be entered. Enter 4 if there is a relatively clear best term match that should not require any further follow-up. Enter 5 if the “best” chosen match is not clearly ideal, and there is reasonable uncertainty that this match is best. Lastly enter 6 if no “best” match could reasonably be identified from the available information. In this case, no values will be entered for cells of the two columns described above. Cells tagged with a 6 demand follow-up by an expert with domain knowledge. Please remember to check all three sets of fuzzy match information (when available). The alternative terms can be useful in more clearly identifying a “best” match.

Dictionary of columns in fuzzy mapping files:

Effort_Split – Temporary indices entered to split up the fuzzy matching task into separate files.
ROW_INDEX – The original row index of the entire dataset. We can always resort on this value to get all entries into their original order.
MHTERM – The Medical History term, available for every entry in the data.
FZMatch_1_MHTERM – Top scoring fuzzywuzzy match to the MHTERM.
FZMatch_2_MHTERM – 2nd highest scoring fuzzywuzzy match to the MHTERM.
FZMatch_3_MHTERM – 3rd highest scoring fuzzywuzzy match to the MHTERM.
FZMatch_4_MHTERM – 4th highest scoring fuzzywuzzy match to the MHTERM.
FZMatch_5_MHTERM – 5th highest scoring fuzzywuzzy match to the MHTERM.
FZMatch_Choice_ID_MHTERM – see above
FZMatch_Copied_Term – see above
LLT_map_code – see above
MHMODIFY – The alternative, “modified” medical history term, available for some entries in the data.
FZMatch_1_MHMODIFY – Top scoring fuzzywuzzy match to the MHMODIFY.
FZMatch_2_MHMODIFY – 2nd highest scoring fuzzywuzzy match to the MHMODIFY.
FZMatch_3_MHMODIFY – 3rd highest scoring fuzzywuzzy match to the MHMODIFY.
FZMatch_4_MHMODIFY – 4th highest scoring fuzzywuzzy match to the MHMODIFY.
FZMatch_5_MHMODIFY – 5th highest scoring fuzzywuzzy match to the MHMODIFY.
LLT_NAME – Another alternative, medical history term, available for some entries of the data. Presumably this term conforms to the MedDRA standard, but we cannot verify which version.
FZMatch_1_LL_NAME – Top scoring fuzzywuzzy match to the LLT_NAME.
FZMatch_2_LL_NAME – 2nd highest scoring fuzzywuzzy match to the LLT_NAME.
FZMatch_3_LL_NAME – 3rd highest scoring fuzzywuzzy match to the LLT_NAME.
FZMatch_4_LL_NAME – 4th highest scoring fuzzywuzzy match to the LLT_NAME.
FZMatch_5_LL_NAME – 5th highest scoring fuzzywuzzy match to the LLT_NAME.
PT_NAME – can ignore here
HLT_NAME – can ignore here
HLGT_NAME – can ignore here
SOC_NAME – can ignore here
T_LL – can ignore here
T_LL_CODE – can ignore here

Supplementary Table S1 Summary of the 28 studies harmonized in this project

ID	Trial	Phase	Treatment	Ref.
1001	AMB-320:ARIES-1	3	Ambrisentan	1
1002	AMB-321:ARIES-2	3	Ambrisentan	1
1003	AMB-320/321-E:ARIES-E	OL	Ambrisentan	2
1004	112565:AMBITION	3–4	Ambrisentan & Tadalafil	3
1005	ac052351	2	Bosentan	4
1006	BSN-352:BREATHE-1	3	Bosentan	5
1007	BSN-354	OL	Bosentan	N/A
1008	97218:AIR	3	Iloprost	6
1009	98008:AIR-II	2	Iloprost	7
1010	SERAPHIN	3	Macitentan	8
1011	CHEST-1	3	Riociguat	9
1012	PATENT-1	3	Riociguat	10
1013	AC-065A302:GRIPHON	3	Selexipag	11
1014	1140:SUPER	3	Sildenafil	12
1015	1142:SUPER-2	OL	Sildenafil	13
1016	FPH01:STRIDE-1	3	Sitaxsentan	14
1017	FPH01x	OL	Sitaxsentan	N/A
1018	FPH02:STRIDE-2	3	Sitaxsentan	15
1019	FPH02x	OL	Sitaxsentan	16
1020	FPH03:STRIDE-3	OL/3	Sitaxsentan	N/A
1021	FPH04:STRIDE-4	3	Sitaxsentan	17
1022	LVGY:PHIRST	3	Tadalafil	18
1023	mc-lvgx	OL	Tadalafil	19
1024	Treprostinil	3	Tresprostinil	20
1025	22387: TRIUMPH	3	Tresprostinil	21
1026	TDE-301:FREEDOM-C	3	Tresprostinil	22
1027	TDE-302:FREEDOM-M	3	Tresprostinil	23
1028	TDE-308:FREEDOM-C2	3	Tresprostinil	24

References

- GALIE N. Ambrisentan in Pulmonary Arterial Hypertension. Randomized, double-blind, placebo-controlled, multicenter, efficacy studies (ARIES) group. *Circulation* 2008; 117:2966–2968
- Oudiz RJ, Galiè N, Olschewski H, et al; ARIES Study Group. Long-term ambrisentan therapy for the treatment of pulmonary arterial hypertension. *J Am Coll Cardiol* 2009;54(21):1971–1981
- Galiè N, Barberà JA, Frost AE, et al; AMBITION Investigators. Initial use of ambrisentan plus tadalafil in pulmonary arterial hypertension. *N Engl J Med* 2015;373(09):834–844
- Channick RN, Simonneau G, Sitbon O, et al. Effects of the dual endothelin-receptor antagonist bosentan in patients with pulmonary hypertension: a randomised placebo-controlled study. *Lancet* 2001;358(9288):1119–1123
- Rubin LJ, Badesch DB, Barst RJ, et al. Bosentan therapy for pulmonary arterial hypertension. *N Engl J Med* 2002;346(12):896–903
- Olschewski H, Simonneau G, Galiè N, et al; Aerosolized Iloprost Randomized Study Group. Inhaled iloprost for severe pulmonary hypertension. *N Engl J Med* 2002;347(05):322–329
- Olschewski H, Hoepfer MM, Behr J, et al. Long-term therapy with inhaled iloprost in patients with pulmonary hypertension. *Respir Med* 2010;104(05):731–740
- Pulido T, Adzerikho I, Channick RN, et al; SERAPHIN Investigators. Macitentan and morbidity and mortality in pulmonary arterial hypertension. *N Engl J Med* 2013;369(09):809–818
- Ghofrani H-A, D'Armini AM, Grimminger F, et al; CHEST-1 Study Group. Riociguat for the treatment of chronic thromboembolic pulmonary hypertension. *N Engl J Med* 2013;369(04):319–329
- Ghofrani H-A, Galiè N, Grimminger F, et al; PATENT-1 Study Group. Riociguat for the treatment of pulmonary arterial hypertension. *N Engl J Med* 2013;369(04):330–340
- Sitbon O, Channick R, Chin KM, et al; GRIPHON Investigators. Selexipag for the treatment of pulmonary arterial hypertension. *N Engl J Med* 2015;373(26):2522–2533
- Galiè N, Ghofrani HA, Torbicki A, et al; Sildenafil Use in Pulmonary Arterial Hypertension (SUPER) Study Group. Sildenafil citrate therapy for pulmonary arterial hypertension. *N Engl J Med* 2005;353(20):2148–2157
- Rubin LJ, Badesch DB, Fleming TR, et al; SUPER-2 Study Group. Long-term treatment with sildenafil citrate in pulmonary arterial

- hypertension: the SUPER-2 study. *Chest* 2011;140(05):1274–1283
- 14 Barst RJ, Langleben D, Frost A, et al; STRIDE-1 Study Group. Sitaxsentan therapy for pulmonary arterial hypertension. *Am J Respir Crit Care Med* 2004;169(04):441–447
 - 15 Barst RJ, Langleben D, Badesch D, et al. STRIDE-2 Study Group. Treatment of pulmonary arterial hypertension with the selective endothelin-a receptor antagonist sitaxsentan. *J Am Coll Cardiol* 2006;47(10):2049–2056
 - 16 Benza RL, Barst RJ, Galie N, et al. Sitaxsentan for the treatment of pulmonary arterial hypertension: a 1-year, prospective, open-label observation of outcome and survival. *Chest* 2008;134(04):775–782
 - 17 Sandoval J, Torbicki A, Souza R, et al; STRIDE-4 investigators. Safety and efficacy of sitaxsentan 50 and 100 mg in patients with pulmonary arterial hypertension. *Pulm Pharmacol Ther* 2012;25(01):33–39
 - 18 Galie N, Brundage BH, Ghofrani HA, et al; Pulmonary Arterial Hypertension and Response to Tadalafil (PHIRST) Study Group. Tadalafil therapy for pulmonary arterial hypertension. *Circulation* 2009;119(22):2894–2903
 - 19 Oudiz RJ, Brundage BH, Galie N, et al; PHIRST Study Group. Tadalafil for the treatment of pulmonary arterial hypertension: a double-blind 52-week uncontrolled extension study. *J Am Coll Cardiol* 2012;60(08):768–774
 - 20 Simonneau G, Barst RJ, Galie N, et al; Treprostinil Study Group. Continuous subcutaneous infusion of treprostinil, a prostacyclin analogue, in patients with pulmonary arterial hypertension: a double-blind, randomized, placebo-controlled trial. *Am J Respir Crit Care Med* 2002;165(06):800–804
 - 21 McLaughlin VV, Benza RL, Rubin LJ, et al. Addition of inhaled treprostinil to oral therapy for pulmonary arterial hypertension: a randomized controlled clinical trial. *J Am Coll Cardiol* 2010;55(18):1915–1922
 - 22 Tapson VF, Torres F, Kermeen F, et al. Oral treprostinil for the treatment of pulmonary arterial hypertension in patients on background endothelin receptor antagonist and/or phosphodiesterase type 5 inhibitor therapy (the FREEDOM-C study): a randomized controlled trial. *Chest* 2012;142(06):1383–1390
 - 23 Jing Z-C, Parikh K, Pulido T, et al. Efficacy and safety of oral treprostinil monotherapy for the treatment of pulmonary arterial hypertension: a randomized, controlled trial. *Circulation* 2013;127(05):624–633
 - 24 Tapson VF, Jing Z-C, Xu K-F, et al; FREEDOM-C2 Study Team. Oral treprostinil for the treatment of pulmonary arterial hypertension in patients receiving background endothelin receptor antagonist and phosphodiesterase type 5 inhibitor therapy (the FREEDOM-C2 study): a randomized controlled trial. *Chest* 2013;144(03):952–958