KEY
White-- All clear
Green-- Different names, same thing
Yellow-- Different ways to capture same variable
Red-- exists in one, not the other; red sus

Description	ST-GPR	Dismod	Red notes	Red status
	age_demographer	age_demographer		
both: must exist in the data set and cannot be null	age_end	age_end		
	age_gp		-seems like this combines age_start and age_end (example: age_start = 0.5, age_end = 1, age_gp = 0.5-0.99) -documentation doesn't say it's necessary so I think we can ignore	IGNORE
	age_group		-thought there was an association with age_gp, but seems not (values are either 0 or NA, so maybe we can ignore?) -I can't figure out what this is	IGNORE
ST-GPR: must exist in the data set and cannot be null	age_group_id		-I think this is our main age difference concern -basically matches	Must add to convert to ST-GPR; no dismod equivalent

			age_start and age_end with the ihme age group id	
	age_issue	age_issue		
	age_sex_specific		-either 0,1,NA -not sure what this is but I think we can ignore	IGNORE
both: must exist in the data set and cannot be null	age_start	age_start		
	bundle_id	bundle_id		
	bundle_version_id		-either NA or 1 (?) -I think we can ignore this	IGNORE
	case_definition	case_definition		
	case_name	case_name		
	cases	cases		
		cases_stillbirth	-pretty sure this is cause specific, I think we can ignore	IGNORE
		cases_top	-pretty sure this is cause specific, I think we can ignore	IGNORE
		cause	-this is the cause, should be the same for every row so I think we can ignore	IGNORE
		clinical_data_type	-basically just saying if its claims data -I think we can	IGNORE

			ignore	
		concatNID_loc_id_YYYY	-a bunch of unique identifying IDs strung together, I think we can ignore (have not seen this in other dismod bundles)	IGNORE
	crosswalk_origin_id		-all NA or 1 -I think we can ignore this	IGNORE
	crosswalk_origin_seq		-all NA or 1 -I think we can ignore this	IGNORE
	crosswalk_parent_seq	crosswalk_parent_seq		
DisMod: optional, but must be 0, 1, or null(?); column names must be valid		Bunch of cvs	-I have no idea if ST-GPR takes covariates, but if it does we should just leave them in there I guess?	??
	data.type		-specifies if its new/reextracted microdata, or who did a new extraction -I think we can ignore this	IGNORE
DisMod: must exist in the data set and are allowed to contain nulls	design_effect	design_effect		
DisMod: must exist in the data set and are allowed to contain nulls	effective_sample_size	effective_sample_size		

		extractor	-this just says who extracted -kinda similar info to data.type in ST-GPR, but I think we can ignore	IGNORE
	field_citation_value	field_citation_value		
	file_path	file_path		
	flag_to_add		-either 1 or NA -not sure	IGNORE
		ghdx.data.type	-'Scientific Literature' or blank -I havent seen this in other bundles, I think we can ignore	IGNORE
		gold_standard	-this is identifying the data that is gold standard case definition -I feel like we can ignore this if there isnt an equivalent ST-GPR column, but not positive	IGNORE
DisMod: if group exists, group_review and specificity must also exist		group	-identifies redundant data across rows for a certain source -I do not know if this is important for ST-GPR	IGNORE
DisMod: optional, but can only be 0, 1, or null; if group_review exists, group		group_review	-1 = visible for modelling -0 = not visible for modelling -I do not know if this is important	IGNORE

and specificity must also exist			for ST-GPR	
	ihme_loc_id	ihme_loc_id		
DisMod: must exist in the data set and are allowed to contain nulls	input_type_id	input_type		
both: must exist in the data set and cannot be null	is_outlier	is_outlier		
both: “must exist in the data set and must match values in the Epi Extraction Template”	location_id	location_id		
	location_name	location_name		
	lower	lower		
both: “must exist in the data set and must match values in the Epi Extraction Template”	measure	measure		
	measure_adjustment	measure_adjustment		
	measure_issue	measure_issue		
		modelable_entity_name	-not used in GBD 2016 and later -- instead use bundle_id and bundle_name -I think we can ignore	IGNORE
	nclust		-no idea	??
both: must exist	nid	nid		

in the data set and cannot be null Values in this column must be present in the GHDx				
	Notes	note_modeler note_sr		
	nstrata		-no idea	??
	old_outlier		-the fact that this is old outlier makes me think we can ignore, but dont know	IGNORE
	orig_year_end		-dont know	IGNORE
	orig_year_start		-dont know	IGNORE
	origin_id	origin_id		
	origin_seq	origin_seq		
		outlier_note	-Im not sure if outliers should be carried over to ST-GPR, but can ignore	IGNORE
		page_num	-not important	IGNORE
	parent_var		-not sure what this is	??
		prenatal_fd	-- likely case specific	IGNORE
		prenatal_fd_definition	-- likely case specific	IGNORE
		prenatal_top	-- likely case specific	IGNORE
		Previously_uploaded	-just saying whether data is old or new -can ignore	IGNORE

DisMod: “must exist in the data set and must match values in the Epi Extraction Template”	recall_type_id	recall_type		
DisMod: must exist in the data set and are allowed to contain nulls	recall_type_value	recall_type_value		
		registry_id	-havent seen this before in bundle, I think we can ignore	IGNORE
		replace_me	-havent seen this before in bundle, I think we can ignore	IGNORE
	representation_level		-not sure what this is	??
DisMod: “must exist in the data set and must match values in the Epi Extraction Template”	representative_id	representative_name		
		response_rate	-response rate of survey -if no equivalent column in st-gpr, we can ignore	IGNORE
ST-GPR: must exist in the data and are allow	sample_size	sample_size		
DisMod: must exist in the data set and are allowed to contain nulls	sampling_type_id	sampling_type		

both: must exist in the data set and are allowed to contain nulls A null value in the `seq` column indicates new data that should be inserted into the database.	seq	seq		
both: “must exist in the data set and must match values in the Epi Extraction Template”	sex	sex		
	sex_issue	sex_issue		
		short_registry_name	-havent seen this before in bundle, I think we can ignore	IGNORE
	smaller_site_unit	smaller_site_unit		
DisMod: “must exist in the data set and must match values in the Epi Extraction Template”	source_type_id	source_type		
DisMod: if specificity exists, group and group_review must also exist		specificity	-free text field, identifies subsets of groups of data -I do not know if this is important for ST-GPR	IGNORE
	standard_error	standard_error		
		table_num	-not important	IGNORE
	stgpr_bundle		-we can	IGNORE

			probably keep this as is, but also dismod doesnt need	
	survey_module		-Just more data info, I think we can ignore	IGNORE
	survey_name		-Just more data info, I think we can ignore	IGNORE
DisMod: must exist in the data set and are allowed to contain nulls	uncertainty_type_id	uncertainty_type		
	uncertainty_type_value	uncertainty_type_value		
	underlying_field_citation_value	underlying_field_citation_value		
both: must exist in the data set and are allowed to contain nulls; nonnull values must be present in the GHDx	underlying_nid	underlying_nid		
DisMod: “must exist in the data set and must match values in the Epi Extraction Template”	unit_type	unit_type		
	unit_type_id		-all values are 1 -im not sure what this is and how it’s different than unit_type	IGNORE
DisMod: “must exist in the data set and must match values in	unit_value_as_published	unit_value_as_published		

the Epi Extraction Template”				
	upper	upper		
DisMod: “must exist in the data set and must match values in the Epi Extraction Template”	urbanicity_type	urbanicity_type		
	urbanicity_type_id		-all values are 1 -im not sure what this is and how it’s different than urbanicity_type	IGNORE
	V1		-no idea	IGNORE
STGPR: must exist in the data and cannot be null	val	mean		
	var		-no idea	IGNORE
STGPR: must exist in the data and allowed to contain nulls	variance		-don’t think this exists in dismod	Must add to convert to ST-GPR; no dismod equivalent
		verification_UseRawBundle	-no idea, but havent seen this before so I think we can ignore	IGNORE
	who_reference_number		-I feel like we can ignore	IGNORE
both: must exist in the data set and cannot be null	year_end	year_end		
ST-GPR: must exist in the data	year_id		-Im not sure what this is, but	Must add to convert to

set and cannot be null			there does not seem to be an equivalent in dismod	ST-GPR; no dismod equivalent
	year_issue	year_issue		
both: must exist in the data set and cannot be null	year_start	year_start		

Ryan's notes on main issues:
1. stupid stuff like name changes for columns that mean the same thing like for dismod i think the column with the number in it is called "mean" whereas in STGPR it's called "val"

2. stupid stuff like I think one uses ID numbers to refer to things like urbanicity, recall type, survey type where the other uses an actual name like "rural", "cross-sectional survey"

3. the 'measure' options are different I think. This most commonly means that in STGPR, I have to model in 'proportion', and then dismod it's 'prevalence'

Rose: handling ages differently-- "For example, let's say you have a source in a dismod bundle with age_start 6 and age_end 12. When you convert that to age_group ID, do you use age 5-9? Or age 5-9 AND 10-14? What about 0-99? It doesn't make sense for a range that big to just duplicate for every age group, so you have to age split. How and where do you convert/split? You'll have to create a set of rules for conversions like this, which will likely depend on the cause, and should be customizable if your goal is to create a function for broad use"

Questions

- Should we just not put in any covariates when converting ST-GPR to DisMod