RURAL design, estimated sample size, and statistical power offered by base cohort (N=4600)

Background. People living in rural communities in the South live shorter and less healthy lives than those residing elsewhere in the United States. The high rural mortality and morbidity does not spare any racial or ethnic group. The NHLBI funded the RURAL Study to investigate the basis of the very high rural burden of heart, lung, blood, and sleep (HLBS) diseases in the rural Southeast US. We designed the cohort envisioning a sample size of 4,600 participants (age 25-64 years, 52% women; 65% White, 33% Black, 2% Hispanic) recruited from ten economically challenged rural counties in four states (KY, AL, MS, and LA). We targeted six higher-risk and four lower-risk rural counties 'paired within a state' for their degree of poverty, race/ethnic composition, and population sizes. RURAL deploys a self-contained mobile examination unit (MEU) with a CT scanner and has designed a baseline Exam to characterize the local built, social, and economic environments; assess familial, lifestyle factors, and medical history; assay standard HLBS risk factors, including genetic risk; evaluate lung function via spirometry; measure subclinical disease burden (CT scan for coronary artery calcium [CAC] and the ankle-brachial index); set up mHealth with 'take-home' smartphones and wearable activity monitors (Fitbits) to assess physical activity and sleep patterns; build bio- and data-repositories, and robust community collaborations. RURAL participants are under regular surveillance of recruited participants to identify and adjudicate new-onset HLBS disease events. Our central hypothesis is that differences in the HLBS risk among people living in these ten rural Southern counties arise from the synergistic interactions of diverse exposures. Due to the COVID-19 pandemic, recruitment of the cohort was slowed. As a result, at the end of the 6-year grant (U01HL146382), April 2025, RURAL will have recruited about 3600 participants. An extension of the grant and a new contract funded by the NHLBI will support the completion of the recruitment of the RURAL cohort and its first baseline Exam, achieving a **total sample size of approximately 4600 as initially envisioned. The contract will fund a second examination of RURAL (Exam 2) to re-examine approximately 4000 participants across the four states in the ten counties** (from which the RURAL cohort was recruited) using the MEU.

RURAL has recruited to-date 3159 participants across six Counties in three states (**Table 1**).

Table 1. RURAL Cohort Enrollment by County/Parish and Demographic Characteristics, as of January 22, 2025

	Alabama		**Mississippi**		**Louisiana**	
	Dallas County	**Wilcox County**	**Oktibbeha County**	**Panola County**	**Assumption Parish**	**Franklin Parish**
Target	693	181	718	617	421	353
Recruited	503	397	757	628	459	371
Women n, (%)	335 (67)	295 (74)	486 (64)	442 (70)	313 (68)	242 (65%)
Age n, (%)						
25-34	66 (13)	70 (18)	106 (14)	91 (15)	59 (13)	51 (14)
35-44	96 (19)	87 (22)	171 (23)	135 (22)	129 (28)	112 (30)
44-54	153 (30)	109 (28)	203 (27)	197 (31)	136 (30)	94 (25)
55-64	188 (37)	131 (33)	277 (37)	205 (33)	135 (29)	114 (31)
Race n, (%)						
Black	362 (72)	355 (89)	338 (45)	337 (54)	80 (17)	82
White	135 (27)	39 (10)	384 (51)	286 (46)	373 (81)	286
Other	6 (1)	3 (1)	35 (5)	5 (1)	6 (1)	2

The target sample sizes for Kentucky Counties are in **Table 2**. It is anticipated that after completion of recruitment in Kentucky, RURAL will have a base cohort size that is or exceeds 4600, the sample size used for the power calculations that follow.

Table 2. RURAL: Selected Counties in Kentucky (Participants to be recruited)

County	**Population Age 25-64**	**% Race Mix* [%W, %B, %H, %O]**	**Target Sample Size**	**Projected Sample by Race**			
				White	**Black**	**Hispanic**	**Other**
Boyle	14712	[86.2, 8.9, 2.8, 2.0]	523	452	45	15	11
Perry	14289	[96.4, 1.5, 0.7, 1.4]	509	491	8	3	7
Garrard	9341	[95.3, 1.9, 1.8, 1.0]	332	317	6	6	3
Breathitt	7092	[98.1, 0.2, 0.6, 1.1]	253	248	1	1	3
*W, white; B, Black; H, Hispanic ethnicity							

Statistical power calculations for statistical analyses of the full RURAL cohort.

The statistical power calculations below are intended to guide ancillary studies at RURAL Exam 2.

We calculated the minimal detectable odds ratios for a difference in the incidence of clinical events (CVD or LD events) between subgroups (e.g., sex, race, high-risk counties). Also, we calculated minimal detectable standardized associations (e.g., z-values, odds ratios) for temporal change in both continuous and binary variables (e.g., a comparison of risk factors, subclinical disease variables, and clinical outcomes between Exam 1 vs. 2) and how these serial changes may be moderated by binary variables (e.g., effect modification by sex, race, high- vs. low-risk Counties). **Table 3** provides an overview of the expected proportions of participants in RURAL subgroups used to calculate the minimal detectable standardized associations.

Table 3. **Expected subgroup percentages in the RURAL cohort after Exam 1 completion**

Age categories				Sex		Race			Higher HLBS risk* County residence	
25-34	35-44	45-54	55-64	Men	Women	White	Black	Other	Yes	No
14	24	29	33	33	67	63	35	2	54	46

*based on HLBS mortality data from the CDC averaged over 5 years

Further, we calculated the minimal detectable standardized associations for different sample sizes that will correspond with cohort attrition rates of 10%, 20%, and 30% since attrition of the recruited sample in our rural areas is currently unclear. Our calculations of minimally detectable associations are overly conservative because we expect to have a 5% larger cohort size than the originally planned 4600 participants.

Under assumptions of homogeneity (in attrition and variability) and for full cohort data analyses, we show that RURAL has enough power to detect serial changes in different measures and assess effect modification by covariates of temporal changes.

Table 4 provides *minimal detectable odds ratios* (MDOR, calculated using Pearson's chi-square statistic) for testing differences in incidence rates for events (CVD or lung disease events) between two subgroups of specific sizes (α=0.05; β=0.80). The subgroup sizes are based on the attrition rates (assumed homogeneity across subgroups) and the percentages in the RURAL cohort (**Table 3**). **Table 4** shows that the MDORs for differences in incidence rates are almost all below value 2, even for low five-year incidence rates. Given that the percentages for white and black participants are similar to those for women and men, the MDORs reported in **Table 4** would also apply to comparing incidence rates between white and black participants. Thus, the RURAL sample is adequate to detect small changes in incidence rates of CVD or LD, even if there is a 30% cohort attrition. For example, if the incidence rate of CVD is 5.0% for women (over five years), an incidence rate of at least 8.6% can be detected in men, even if the attrition on follow-up is 20%. *Note that minimum detectable hazard ratios would be slightly better than the MDORs in* ***Table 4*** *when time-to-events are used instead of proportions of events*.

Table 4. MDORs for detecting differences in incidence rates between subgroups with different sample sizes

Subgroup	Attrition	Subgroup sizes	5-year incidence rate			
			1.5%	2.5%	5.0%	7.5%
Women vs. Men	0%	3066 vs 1534	1.85	1.64	1.44	1.36
	10%	2760 vs 1380	1.90	1.68	1.47	1.38
	20%	2453 vs 1227	1.96	1.72	1.50	1.41
	30%	2146 vs 1074	2.04	1.78	1.54	1.44
High-risk vs. Low-risk	0%	2470 vs 2130	1.81	1.61	1.42	1.34
	10%	2223 vs 1917	1.86	1.64	1.45	1.36
	20%	1976 vs 1704	1.92	1.69	1.48	1.39
	30%	1729 vs 1491	2.00	1.75	1.51	1.42

Minimal detectable standardized mean changes (MDSMC) between Exams 1 and 2 for a continuously distributed measure, expressed in total standard deviations (σ_{TOT}), are shown in **Table 5** for different cohort attrition rates and intraclass correlation coefficient (ICC) between measures (continuous variables) from RURAL Exam 1 and Exam 2. The total standard deviation combines intra- and inter-individual variation (at baseline), and the MDSMCs are calculated using an F-test (α=0.05; β=0.80). It follows that small changes in the average population value can be detected for numerical variables (e.g., risk factors levels) between Exam 1 and Exam 2. Longitudinal changes in the mean value of measures equivalent to ~7% of the baseline variation in the variables can be detected at Exam 2. For example, a change of 1.3 mmHg or more in the mean systolic BP between the two Exams would be detectable ($\sigma_{\text{TOT}} = 18.55$ mmHg). Furthermore, for a comparison of serial changes in levels of a risk factor between two subgroups (e.g., sex), the minimal detectable standardized difference in mean change (MDSDMC) is approximately 41% larger than the MDSMC listed in **Table 5** when the intra-individual variation within the two subgroups is identical, irrespective of the difference in subgroup sizes. For example, if no mean serial change in systolic BP is observed in women, a mean change in men of 1.84 mm Hg or more (across Exams) would be statistically significant (with 80% confidence) from the change observed in women.

Table 5. MDSMC for different attrition rates and ICCs

Cohort size	ICC (%)		
	1%	5%	10%
4600	0.059	0.057	0.056
4140	0.062	0.061	0.059
3680	0.066	0.064	0.062
3220	0.070	0.069	0.067

The *minimal detectable mean difference in probability* (MDMDP) between Exams 1 and 2 for a paired binary variable (such as the presence of CAC measured at Exam 1 and 2) is shown in **Table 6**. The difference is calculated using the McNemar statistic and it depends on the observed agreement between two waves (i.e., the probability that at both Exams, the same classification is observed: both present or both absent). A high agreement indicates a stronger dependency between the two Exams. The calculated difference in **Table 6** is an absolute difference in probabilities between two waves (and is not standardized). For example, if the cohort attrition is 30%, and the agreement between the two Exams is 80%, we can detect an absolute serial change in the prevalence of an event (e.g., a CAC score>0) of 2.22%.

Table 6. MDMDP for different attrition rates & observed agreement

Cohort size	**Observed agreement**			
	90%	80%	70%	60%
4600	1.32	1.86	2.28	2.62
4140	1.38	1.96	2.40	2.76
3680	1.48	2.08	2.54	2.94
3220	1.58	2.22	2.72	3.14