Impact of Clinical Center Effects on Objective Response Rate

Fang Liu[1], Cong Chen[1], Wen Li[1]
[1]Merck & Co., Inc., 351 North Sumneytown Pike, North Wales, PA, USA, 19454

Abstract

It is well-known that a trial could end up with different conclusions due to difference in the selected clinical centers. However, the impact of center effects on trial endpoints is not studied adequately. We quantify the impact of center effects on Objective Response Rate (ORR) in early oncology single-arm trials, which are conducted at multiple small and heterogeneous centers. Based on the variance formula for ORR we derived after adjusting the center effects, we provide guidance on minimizing the center effects during the trial design stage, by considering how many centers to be selected, how to distribute patients among centers and how to set a center enrollment cap for a trial. The conclusion can be applied directly to the clinical trials with binary endpoints other than ORR and also shed light on clinical trials with different endpoints.

Key words: Multi-centre trials, center effects, between-center variation, clinical center selection, enrollment cap

1. Background

Clinical trials often recruit participants from multiple clinical centers to expedite the enrollment and enhance the generalizability of results by including a wider range of population groups. Due to factors such as differing patient characteristics, methods of measuring or recording data, processes of care, or training of staff, patient outcomes sometimes vary by center [1-3]. Therefore, the conclusion of a trial could be difference due to difference in the selected centers. Because of this, many trials attempt to minimize the impact of any between-center variations on the trial results, either during the design stage (by stratifying on center in the randomization process), or during analysis stage (by accounting for center effects in the analysis model) [2]. Though randomization stratified by center can reduce the center effects, it is not clear how much center effects can be reduced. In addition, it is not applicable to single-arm trials. There are numerous methods proposed in the literature to account for the center effects in the analysis stage. However, adjusting for center effects in the model can often be problematic, particularly when there are a large number of centers compared to the overall sample size. In trials with binary endpoints, too few patients or events per center can lead to biased estimates [4] or inflated type I error rate for some analysis methods [5].

In this article, we will investigate the impact of center effects in single-arm early oncology trials, in which objective response rate (ORR) is the primary binary endpoint. In early oncology trials, it is common that only one or two patients are enrolled in certain clinical centers. Therefore, analysis models accounting for center effects could cause problem and are not often used. Instead, Clopper-Pearson confidence interval (CI) [6, 7] is used for ORR in early oncology trials.

Figures 1 and 2 show the ORR of each clinical center from two early oncology trials. The radius of the circle is proportional to the number of subjects enrolled in each center. The ORR varies across centers in both trials, even for centers with relative more subjects. For example, in Figure 1, the ORRs of clinical centers 12 and 13 are less than 20%, while the ORRs of clinical centers 14 and 15 are around 35%.

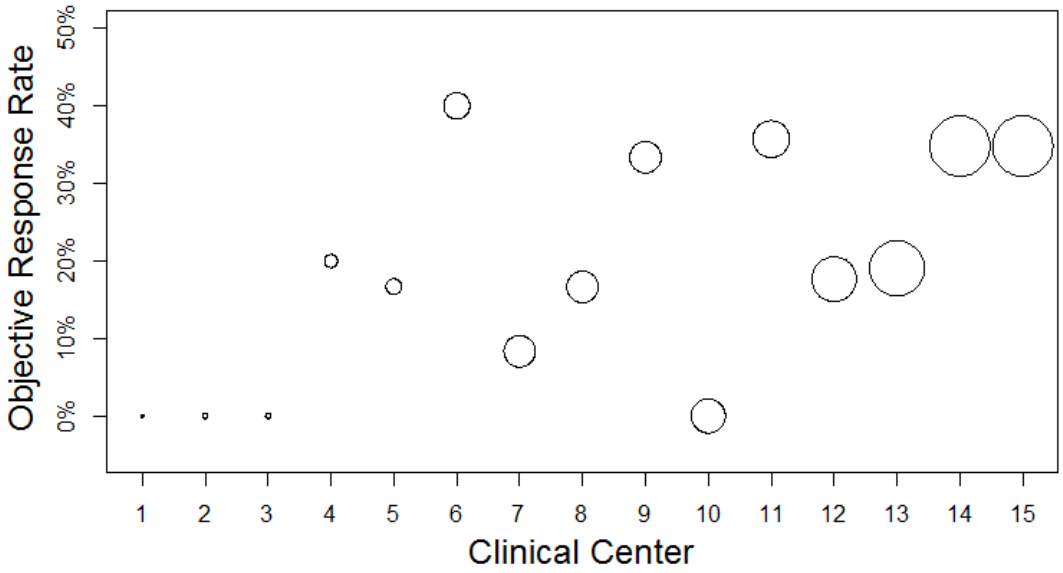

Figure 1: ORR by clinical center from example trial 1

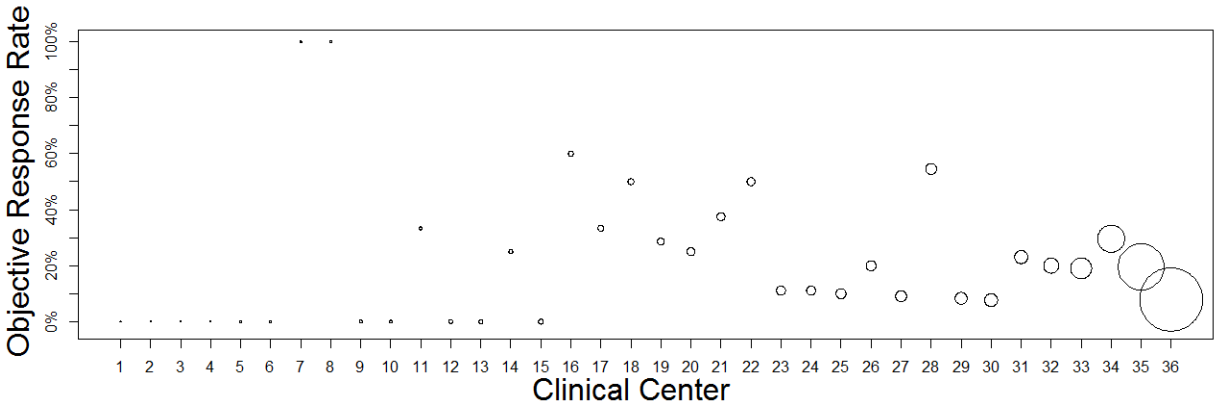

Figure 2: ORR by clinical center from example trial 2

We estimated the between-center standard deviation (SD) and coefficient of variance (CV) in ORR from nine early oncology trials using generalized linear mixed model, as provided in Table 1. The between-center coefficient of variance estimates vary from 0% to 43.9%.

From above real trial examples, we noticed that the center effects could potentially be very large. However, how significantly the center effects could impact the ORR estimate? Can we reduce the center effects in the design stage in the single-arm trials? Will the statistical inference be valid after minimizing the center effects? We try to answer these questions in the article.

The rest of the article is organized as follows. In section 2, we derive the ORR variance formula after adjusting the center effects. A simulation study is performed in Section 3 to evaluate the confidence interval coverage using the derived ORR variance formula. In Section 4, we provide guidance on minimizing the center effects in the design stage, including how many centers to be selected, how to distribute patients among centers and how to set a center enrollment cap in a trial. We conclude this article with a brief discussion in Section 5.

Table 1: Between-center SD and CV estimates in ORR from Nine Trials

Example Trials	Number of Centers	Number of responses	Number of subjects	ORR Estimate*	Between-center SD in ORR*	Between-center CV in ORR
Trial 1	15	41	173	23.1%	5.7%	24.7%
Trial 2	36	71	356	21.2%	9.3%	43.9%
Trial 3	52	30	259	10.9%	4.4%	40.4%
Trial 4	10	15	25	60.0%	0.0%	0.0%
Trial 5	18	8	31	25.7%	3.3%	12.8%
Trial 6	9	4	38	13.2%	0.0%	0.0%
Trial 7	26	27	101	27.8%	10.1%	36.3%
Trial 8	29	10	55	18.1%	2.2%	12.2%
Trial 9	16	34	193	18.2%	6.5%	35.7%

* Estimated from generalized linear mixed model using PROC Glimmix with clinical center as a random effect.

SD: Standard Deviation; CV: Coefficient of Variation.

2. Variance and confidence interval of ORR after adjusting center effects

Consider an early stage single-arm oncology trial with K clinical centers and N subjects. Let n_k be the number of subjects enrolled at Center k, where $k = 1,2,\dots,K$, and $\sum_{k=1}^{K} n_k = N$. Let P be the true ORR of the treatment and P_k be the true ORR at Center k. σ represents the between-center standard deviation in ORR. It is reasonable to assume $P_k = P + \varepsilon$, where ε is a random variable with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. Thus, $E(P_k) = P$ and $Var(P_k) = \sigma^2$. Let R_{ki} denote the response of subject i at Center k ($R_{ki} = 1$ means responder; otherwise $R_{ki} = 0$) and $R_{ki}|P_k \sim Bernoulli(P_k)$. Let Y_k denote the total number of responses at Center k, $Y_k = \sum_{i=1}^{n_k} R_{ki}$ and $Y_k|P_k \sim Binomial(n_k, P_k)$. The true ORR ($P$) is estimated as $\hat{P} = \frac{\sum_{k=1}^{K} Y_k}{N}$.

The variance of ORR estimate after adjusting center effects can be written as

$$Var(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{N} + \frac{(\sum_{k=1}^{K} n_k^2 - N)}{N^2}\sigma^2. \qquad (1)$$

The derivation of Equation (1) is provided in the Appendix. The variance of ORR contains two parts, one part ($\frac{\hat{P}(1-\hat{P})}{N}$) is from the binomial distribution and the other part ($\frac{(\sum_{k=1}^{K} n_k^2 - N)}{N^2}\sigma^2$) is from the between-center variation. For Equation (1), when $\hat{P}$ and N are fixed, it is easy to prove that

- the variance of ORR is the largest with $K = 1$ and $n_k = N$, which means that all subjects are enrolled in one center;
- the variance of ORR is the smallest with $K = N$ and $n_k = 1$, which means that each center enrolls only 1 subject and there are N centers.;
- if number of centers K is fixed, the variance of ORR is minimized when n_k is the same across K centers and Equation (1) can be simplified as

$$Var(\hat{P}) = \frac{\hat{P}(1-\hat{P})}{N} - \frac{1}{N}\sigma^2 + \frac{1}{K}\sigma^2.$$

From Equation (1), we can construct ORR CI using the extended Clopper-Pearson method [8] or normal approximation when the sample size is large. The extended Clopper-Pearson CI was proposed for over-dispersed binary data, when the binary data (such as ORR) shows more variation than estimated by the binomial distribution (e.g., $\frac{\hat{P}(1-\hat{P})}{N}$) [8]. Let λ denote the variance inflation factor, which can be estimated as

$$\hat{\lambda} = Var(\hat{P})/(\hat{P}(1-\hat{P})/N)$$

and, per definition, an estimate of the effective sample size is N/λ. The extended Clopper-Pearson $100(1-\alpha)\%$ CI of $\hat{P}$ after adjusting the center effects is

$$\left(\ 1 - BetaInv(\tfrac{\alpha}{2}, \tfrac{N-Y}{\lambda}, \tfrac{Y}{\lambda}+1),\ 1 - BetaInv(1-\tfrac{\alpha}{2}, \tfrac{N-Y}{\lambda}+1, \tfrac{Y}{\lambda})\ \right),$$

where Y is the total number of responses in the trial and $Y = \sum_{k=1}^{K} Y_k$.

The $100(1-\alpha)\%$ Wald CI of $\hat{P}$ after adjusting the center effects based on asymptotic normality is

$$\hat{p} \pm z_{\alpha/2}\sqrt{Var(\hat{P})},$$

where $z_{\alpha/2}$ denote the $1-\alpha/2$ quantile of the standard normal distribution.

The concept of 'effective sample size' is well-known within the survey sampling community. Although it lacks a unique definition in the statistical literation, effective sample size is generally used as a measure of the equivalent number of independent samples [8]. In the next section, We evaluate the ORR confidence interval coverage accounting for center effects using simulations.

3. Simulation study

3.1 Simulation Set-up

In the simulation studies, we consider single-arm oncology trials with 50 or 200 subjects. For illustration purpose, the true ORR is set as 20% and between-center SD is set as 7%. Thus, the between-center CV is 35%, which is reasonable based on the nine example trials as described in Section 1. In each of the simulated trials, subjects are evenly allocated to 1, 2, 5, 10, 20, 25 or 50 clinical centers. As we discussed in Section 2, when the number of clinical centers is fixed, the variance of ORR is the smallest if subjects are equally allocated across centers. Therefore, the CI coverage in our simulation represents the optimal coverage that we can achieve in a real trial with the same number of clinical centers.

3.2 Simulation Results

Results based on 10,000 simulations are presented in Table 2. The 95% CI of ORR are calculated using six methods,

- Clopper-Pearson CI and Wald CI without adjusting center effect,
- extended Clopper-Pearson CI and Wald CI accounting for center effects with true between-center SD (σ) from simulation set-up,
- extended Clopper-Pearson CI and Wald CI accounting for center effects with between-center SD estimated from the generalized linear mixed model.

In general, the coverages of CIs accounting for center effects become closer to the nominal coverage (95%), comparing to the CIs without adjusting center effects. When there is only one clinical center in a trial, the coverages of CIs accounting for center effects are the same as CIs without adjusting center effects, as the between-center SD cannot be estimated with one clinical center. If the true between-center SD is known, the coverages of CIs (shaded in gray) are always close to 95%. When there are less than 5 clinical centers in a trial with 50 subjects, or less than 20 clinical centers in a trial with 200 subjects, the coverages of CIs accounting for center effects with estimated between-center SD are lower than the coverages of CIs using the true between-center SD, which means the between-center SD is underestimated when there is a limited number of clinical centers in a trial.

Clopper-Pearson CI is known as a conservative method, as the coverage level is usually higher than the nominal level. However, when there are less than 5 clinical centers in a trial with 50 subjects, and less than 20 centers in a trial with 200 subjects, the coverages of Clopper-Pearson CIs are less than 95%. Therefore, if a trial has a limited number of clinical centers, Clopper-Pearson CI could become less conservative. However, when there are 10 or more clinical centers in a trial with 50 subjects, or more than 50 clinical centers in a trial with 200 subjects, the coverages of Clopper-Pearson CIs become close to 95%. This suggests that increasing the number of clinical centers can reduce the impact of the center effects, so that Clopper-Pearson CI can provide valid statistical inference on ORR.

Table 2: 95% ORR Confidence Interval Coverage

Sample Size	No. of Clinical Centers	Clopper-Pearson CI	Extended Clopper-Pearson CI		Wald CI		
			With estimated between-center SD*	With true between-center SD	Without adjusting center effects	With estimated between-center SD*	With true between-center SD
50	1	82.0%	82.0%	96.3%	76.7%	76.7%	94.7%
	2	89.3%	92.7%	96.0%	85.2%	89.8%	95.6%
	5	93.8%	95.7%	96.1%	90.7%	93.2%	95.5%
	10	95.3%	96.7%	96.5%	92.4%	93.4%	92.4%
	25	96.2%	96.3%	96.2%	93.9%	94.1%	93.9%
	50	96.5%	96.5%	96.5%	94.3%	94.3%	94.3%
200	1	56.6%	56.6%	95.1%	53.2%	53.2%	95.2%
	2	69.6%	84.0%	95.3%	66.9%	81.1%	94.9%
	5	81.7%	91.1%	94.9%	80.2%	89.5%	94.5%
	10	87.9%	93.2%	95.1%	87.1%	91.9%	93.7%
	20	91.1%	92.7%	95.6%	90.9%	92.1%	94.5%
	50	93.1%	93.2%	94.5%	92.9%	92.8%	93.8%

P(True ORR)=20%; σ (Between-center SD in ORR) =7%.

*Estimated from generalized linear mixed model using PROC Glimmix with center as a random effect.

4. Application to Trial Design

In this section, we discuss how to minimize the center effects in the trial design stage using a hypothetic trial with 50 subjects and a true ORR of 20%. We mainly focus on three aspects: how many centers should a trial have, how to allocate subjects across centers, and how to set up a center enrollment cap.

4.1 How many centers should a trial have?

Figure 3 shows the association between the relative effective sample size (defined as effective sample size/number of subjects enrolled*100) and number of clinical centers. When the between-center SD in ORR is 7%, the relative effective sample size increases from 40% to 89% with number of centers increasing from 1 to 10. The relative effective sample size is higher when the between-center SD is smaller. Similarly to Section 3, we assume that subjects are evenly distributed among the clinical centers in this figure. For a trial with 50 subjects, we should consider to have at least 10 centers, to achieve around 90% efficiency in sample size. When the number of centers becomes more than 10, the relative effective sample size becomes more stable and there is no much gain in the relative effective sample size.

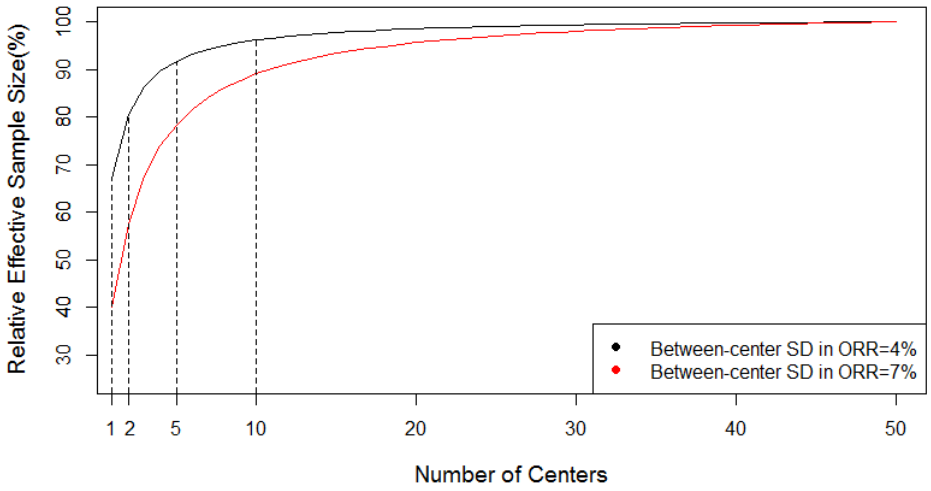

Figure 3: Relative Effective sample size

4.2 How to allocate subjects across centers?

Table 3 illustrates the relative effective sample size with different enrollment in each center. The optimal enrollment of the trial is 50 centers with one subject in each center, which has 100% relative effective sample size. However, this is the ideal scenario and it is not likely to be applied in a real trial. Another extreme case is having one center with all 50 subjects enrolled in this center, which results in the largest ORR variance and the relative effective sample size is small, e.g. only 40%.

When a trial has ten centers with five subjects in each center as recommended in Section 4.1, the relative effective sample size is 89% and it seems to be acceptable and feasible. However, if three of the ten centers enroll most of subjects (39 subjects), the relative effective sample size reduces to 77%, which is similar to the relative effective sample size of the five centers with 8-12 subjects per center. Note that when there are 4-6 subjects in each of the ten centers, the relative sample size decreases slightly (from 89.1% to 88.7%) comparing to a trial with five subjects in each of the ten centers. This gives us some flexibility in the enrollment. In summary, if there are a sufficient number of centers with patients relative balanced across the centers, the impact of center effects on ORR could be relative small.

Table 3: Effective sample size with different enrollment in each center

Number of Centers	Enrollment in each center	SD of $\hat{P}$	Effective Sample Size	Relative Effective Sample Size (%)
50	(1,1,…1)	0.057	50	100.0
1	50	0.090	20	40.0
10	(5,5…,5)	0.060	45	89.1
10	(1, 1, 1, 1, 2, 2, 3, 13, 12, 14)	0.064	39	77.3
10	(4,4,4,5,5,5,5,6,6,6)	0.060	44	88.7
5	(8,9,10,11,12)	0.064	39	78.0
N (Sample size)=50, P(True ORR)=20%, σ (Between-center SD in ORR) =7%. SD: Standard Deviation.				

4.3 How to set an enrollment cap for clinical centers?

Another question of interest is the enrollment cap of clinical centers. It is common that in a trial, a couple of centers (we call them 'super centers') can enroll much more subjects than other centers. Therefore, setting an enrollment cap can avoid extreme enrollment for these super centers. Table 4 provides the relative effective sample size for different enrollment caps in a trial with ten clinical centers. An enrollment cap of 50% can give at most 71% relative effective sample size, while an enrollment cap of 30% can have at most 84% relative effective sample size.

Table 4: Enrollment cap for clinical centers

Number of Centers	Enrollment of each center	SD of $\hat{P}$	Effective Sample Size	Relative Effective Sample Size	Enrollment Cap
10	(2,2,3,3,3,3,3,3,3,25)	0.067	36	71.6	50%
10	(4,4,4,3,3,3,3,3,3,20)	0.064	39	78.3	40%
10	(4,4,4,4,4,4,4,4,3,15)	0.062	42	84.0	30%
10	(5,5,5,5,5,5,5,5,5,5)	0.060	45	89.1	10%
N (Sample size)=50, P(True ORR)=20%, σ (Between-center SD in ORR) =7%. SD: Standard Deviation.					

5. Conclusion and Discussion

Though center effects are well-recognized in clinical trial, the degrees of impacts on the trial results are not well studied. In this article, we derive a mathematic formula to evaluate the impact of the center effects on the ORR estimates in oncology trials. When there are large center effects, it could significantly impact the efficiency of a trial and Clopper-Pearson confidence interval becomes less conservative. We further provide some guidance to minimize the center effects at the trial design stage. For a single-arm oncology trial with 50 subjects, if there are ten centers (depending on perceived between-center variability) with patients relative balanced across the centers, the impact of center effects on ORR could be relative small and Clopper-Pearson CI could become appropriate. When equally allocation is not possible, we should consider adding an enrollment cap to avoid too many participants in the super centers.

Statisticians are seldom consulted at the trial design stage for the topics such as clinical center selection, number of patients per center or recruitment cap per center, as these are more like operational questions. This article provides some statistical insights on these topics by illustrating these insights through early oncology trials. Statisticians should take this opportunity to involve further in the trial design or setup, e.g. contributing to the operational topics on clinical center selection, number of patients per center or recruitment cap per center. The conclusions in this article can be directly applied to the clinical trials with the binary endpoints other than ORR. Similar ideas are under consideration for clinical trials with control arm or with continuous endpoints.

Acknowledgements

During the development of the methodology, we have received great helps from our colleagues at Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. We are especially thankful to Dr. Guanghan Liu, Dr. Meihua Wang and Dr. Victoria J. Plamadeala for their insightful suggestions to this article.

Reference

1. Kahan, B. C. and Harhay, M.O. Many multicentre trials had few events per centre, requiring analysis via random-effects models or GEEs. J Clin Epidemiol, 2015, 68 (12): 1504-1511.
2. Kahan, B. C. Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? BMC medical research methodology. 2014, 14 (1): 20.
3. Kahan, B. C. and Morris, P.T. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effect? Statistics in medicine. 2013; 32 (7): 1136-1149.
4. Agresti, A. and Hartzel, J. Strategies for comparing treatments on a binary response with multi-centre data. Statistics in medicine. 2000, 19 (8): 1115-1139.
5. Kahan, B.C. and Morris, P.T. Improper analysis of trials randomized using stratified blocks or minimization. Statistics in Medicine, 31 (4): 328-340.
6. Clopper, C.J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934, 26, 404-413.
7. Agresti, A. and Coull, B. Approximate is better than exact for interval estimation of binomial proportions. *American Statistician*, 1998, 52 (2), 119-126.
8. Chen, C. and Tipping, R.W. Confidence interval of a proportion with over-dispersion, *Biometrical Journal*, 2002, 44 (7), 877-886.

Appendix

The derivation of Equation (1) is described below.

Variance of R_{ki} (the response of subjects i at Center k) is

$$Var(R_{ki}) = E\big(Var(R_{ki}|P_k)\big) + Var\big(E(R_{ki}|P_k)\big)$$
$$= E\big(P_k(1-P_k)\big) + Var(P_k) = P - E\big(P_k^2\big) + \sigma^2$$
$$= P - (P^2 + \sigma^2) + \sigma^2 = P(1-P),$$

and the correlation between R_{ki} and R_{kj} (the response of subjects j at Center k) is

$$Cov\big(R_{ki}, R_{kj}\big) = E\left(Cov\big(R_{ki}, R_{kj}\big|P_k\big)\right) + Cov\left(E(R_{ki}|P_k), E\big(R_{kj}\big|P_k\big)\right)$$
$$= 0 + Cov\big(P_k,\ P_k\big) = \sigma^2.$$

So we can have

$$Var(Y_k) = Var\left(\sum_{i=1}^{n_k} R_{ki}\right) = n_k Var(R_{ki}) + 2\binom{n_k}{2} Cov\big(R_{ki},\ R_{kj}\big)$$
$$= n_k P(1-P) + n_k(n_k-1)\sigma^2.$$

Therefore,

$$Var\big(\hat{P}\big) = Var\left(\frac{\sum_{k=1}^{K} Y_k}{N}\right) = \frac{1}{N^2} Var\left(\sum_{k=1}^{K} Y_k\right)$$
$$= \frac{1}{N^2}\sum_{k=1}^{K}\left(n_k P(1-P) + n_k(n_k-1)\sigma^2\right)$$
$$= \frac{P(1-P)}{N} + \frac{\left(\sum_{k=1}^{K} n_k^2 - N\right)}{N^2}\sigma^2.$$