Journal of Clinical Epidemiology 112 (2019) 77−86

Journal of Clinical Epidemiology

REVIEW

Leveraging the entire cohort in drug safety monitoring: part 1 methods for sequential surveillance that use regression adjustment or weighting to control confounding in a multisite, rare event, distributed data setting

Jennifer C. Nelson[a,b,*], Ernesto Ulloa-Pérez[a,b], Jennifer F. Bobb[a,b], Judith C. Maro[c,d]

[a]*Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA*
[b]*Department of Biostatistics, University of Washington, Seattle, WA, USA*
[c]*Department of Population Medicine, Harvard Medical School, Boston, MA, USA*
[d]*Harvard Pilgrim Health Care Institute, Boston, MA, USA*

Accepted 4 April 2019; Published online 18 May 2019

Abstract

Objective: Study designs involving self-controlled or exposure-matched samples are commonly used to monitor postmarket vaccine and drug safety, and they use a subset of the available larger cohort. This article overviews group sequential methods designed for observational data safety monitoring that use the whole exposed and unexposed cohorts by implementing regression adjustment or weighting to control confounding.

Methods: We summarize what is known about the performance of ''whole cohort'' methods in multisite health plan data networks such as the Sentinel System of the Food and Drug Administration, where outcomes are rare, individual-level patient data cannot be pooled across sites, site heterogeneity is large, and data are dynamically updated over time.

Results: Group sequential estimation and testing methods that use regression or weighting can flexibly handle electronic health care data's unpredictability, including an uncertain rate of new product uptake, variable composition of the population over time, and data changes due to dynamic administrative updates. Regression and weighting methods generally have higher power, faster signal detection, and fewer practical challenges compared with some design-based confounder adjustment methods.

Conclusion: Group sequential regression adjustment and weighting approaches are feasible and underused in practice. They leverage more information than designs that involved sampling and increase power to detect rare adverse effects without increasing bias.

Keywords: Active surveillance; Distributed databases; Drug safety; Group sequential; Pharmacoepidemiology; Statistical methods

1. Introduction

Group sequential monitoring of clinical trials has been used for decades to conduct routine evaluations of evidence over time that can more rapidly detect treatment effects compared with one end-of-study analysis [1]. Recently, sequential methods have been modified for observational safety surveillance using electronic health care record data, including the Vaccine Safety Datalink [2] and the Food and Drug Administration's Sentinel System [3,4]. The goal is faster identification of rare side effects in the early postapproval period as new vaccine or drug uptake occurs. In nonexperimental settings like these, careful attention to confounding is needed, and these newer sequential approaches offer numerous options: self- or historical-control designs [5−7], exposure matching [7], stratification

Conflicts of interest statement: none.

Funding: This work was funded by the Sentinel System, a project sponsored by the U.S. Food and Drug Administration (FDA), United States to support monitoring the safety of FDA-regulated medical products. The Sentinel System is one piece of the Sentinel Initiative, a multifaceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) contract number HHSF223201400030I.

* Corresponding author: Biostatistics Unit, Kaiser Permanente Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, USA. Tel.: +1-206-287-2004; fax: +1-206-287-2871.

E-mail address: Jen.Nelson@kp.org (J.C. Nelson).

https://doi.org/10.1016/j.jclinepi.2019.04.012
0895-4356/

What is new?

Key findings

- Group sequential methods, which involve estimating and testing drug-outcome associations over time as new drug uptake occurs, can facilitate more rapid identification of safety concerns compared with a traditional one-time retrospective analysis.
- Group sequential methods that use regression or weighting to control confounding use information from the entire available cohort and can increase power and speed in detecting rare events compared with designs that sample a subset of participants. Estimating a risk difference, instead of a ratio measure, can yield further efficiency gains in the rare outcome setting.

What this adds to what was known?

- Group sequential methods from clinical trials have been adapted for observational drug safety monitoring of rare adverse events in multisite, distributed, electronic health data settings like FDA's Sentinel System. These adaptations permit confounder control, are designed to flexibly accommodate unpredictability in health plan data, such as an uncertain or slow new product uptake rate and unknown composition of the new user population over time, and have been shown to be feasible in this new setting.

What is the implication and what should change now?

- Compared with design-based confounder adjustment methods that involve sampling, strategies applied at the analysis phase, like regression or weighting, permit greater analytic flexibility, and result in fewer practical challenges when applied sequentially in health plan data settings where data change over time due to dynamic updating.

[8], regression [9], and inverse probability weighting [10]. In practice, vaccine surveillance using health care databases has typically involved self- or historical-control designs [11−20]. For drugs, cohort designs with propensity score matching are common [21−24]. Reviews of electronic health record−based vaccine surveillance studies that use sequential methods have been published [25,26], and their use in drug safety continues to emerge. In a related paper in this issue (insert in depth paper hyperlink here), Cook et al. present an in-depth real-world example application of methodologies outlined in this paper to further demonstrate and compare their use in practice.

Although a matched cohort design is advantageous for many safety questions [27], one drawback is that it only samples a subset of patients and does not use readily available information from the larger cohort, particularly if 1:1 matching is performed. Similar concern has been noted in the analysis of matched case−control data nested within a larger cohort [28]. In postmarketing, assessing rare events for which statistical power was limited in prelicensure trials is a particular priority. Thus, excluding data by design may be especially problematic because (i) new drug uptake can be small relative to a large comparator population, and so 1:1 sampling could exclude many comparable patients and (ii) safety outcomes are often rare, and so each lost event represents a higher fraction of the (already limited) total available information. Even in large databases, serious adverse events may be few and a reduction of information could slow or preclude safety signal detection.

In this paper, we overview group sequential methods that have been adapted from clinical trials and use the entire available exposed and unexposed cohort of interest. They do so by incorporating confounding control at the analysis phase using either regression adjustment [9] or inverse probability weighting [10]. Regression adjustment is accomplished using a generalized estimating equations framework to reduce modeling assumptions. We refer to these methodologies, respectively, as group sequential generalized estimating equations (GS GEE) regression and group sequential inverse probability weighted (GS IPTW) regression. We summarize what is known about the performance of these "whole cohort" sequential methods from both simulation studies and data applications in the observational safety setting for which they were developed: multisite, health care data networks like Sentinel. Key setting features are rare outcomes, a distributed data environment where individual-level patient data cannot typically be pooled across health plan sites, large site heterogeneity, and data that are dynamically updated over time as new information is captured for health plans' administrative and clinical purposes.

2. "Whole cohort" group sequential regression methods

Sequential methods control the overall false positive (type 1) error rate across all analyses conducted over time at a prespecified level and permit early rejection of the null hypothesis (of no exposure-outcome association) based on preset signaling rules. The methods reviewed herein involve both (1) risk estimation of the association between an exposure and outcome that adjusts for confounders and (2) hypothesis testing, which constructs a "safety signal" threshold at multiple surveillance time points indicating statistical significance of the estimated effect. We summarize the methods' key features (Table 1), advantages and limitations

(Table 2), performance, and how they fit into a broader class of methods that one can envision for distributed observational data settings (Table 3). Detailed method descriptions have been published previously [9,10], and the companion paper in this issue provides a detailed illustration of their implementation in a multisite safety evaluation of the combination measles-mumps-rubella-varicella vaccine using the Food and Drug Administration Sentinel data (insert hyperlink for in-depth example paper here).

2.1. Group sequential regression adjustment for odds ratio or relative risk estimation (GS GEE)

As surveillance or study data accrue, GS GEE estimates an exposure-outcome association using regression adjustment for confounders at multiple sequential time points. It can accommodate adverse events that occur relatively quickly (e.g., within weeks) after new exposure initiation by using logistic regression to estimate an odds ratio. It can also be used for events that require longer follow-up by using Poisson regression to estimate a relative risk that accounts for differences in exposed person-time and censoring due to loss to follow-up or medication switching [9]. To protect privacy, the individual-level data at each site (Appendix Table 1) are first deidentified via aggregation: the numbers of outcomes and new users (or person-time) are totaled within groups defined by categories of exposure and confounders (Appendix Table 2) or categories of a propensity score. The grouped data are then combined across sites and used in a single model with flexible regression adjustment for categories of the confounders or propensity score and for site. Interactions with site or other confounders may be included to address heterogeneity. The use of a generalized estimating equations framework provides for robust standard errors [29]. The application paper in this issue illustrates this logistic regression modeling approach in Section 2.3.1, modeling the probability of an acute seizure outcome with adjustment for categories of age, sex, and site among a pediatric cohort of measles-mumps-rubella-varicella vaccine recipients (insert in-depth paper hyperlink here).

2.2. Group sequential inverse probability of treatment weighting linear regression for risk difference estimation (GS IPTW)

This ‘‘whole cohort’’ sequential method estimates a multivariable adjusted risk difference and uses inverse probability of treatment weighting based on a propensity score to adjust for confounding [10]. It extends the nonsequential methodology proposed by Lunceford and Davidian [30] and adapts it to a distributed data setting. Like sequential logistic regression, it applies when adverse events occur relatively quickly after new exposure initiation. To estimate a risk difference, linear regression is implemented with a binary outcome. Instead of grouping data centrally across sites, the inverse probability of treatment weighting regression approach uses meta-analysis. Individual-level data at each site are used to fit separate propensity score models and propensity score–weighted linear regression models for risk differences. Only the site-specific risk differences and their standard errors are combined centrally (Appendix Table 3) using Mantel-Haenszel methods to compute an overall site-stratified risk difference estimate. Robust standard errors are computed to avoid using incorrect model-based estimates (i.e., to avoid a normal or binomial assumption) of the variability. Standard errors are also computed to correctly propagate the uncertainty associated with estimating the propensity score, which can differ by site especially when site sample sizes vary widely [10]. The exact formulation of the variance estimator is detailed by Lunceford and Davidian [30]. Section 2.3.2 of the application paper in this issue specifies how propensity score models are fit and how GS IPTW is applied using data from multiple Sentinel sites to sequentially evaluate the association between receipt of measles-mumps-rubella-varicella vaccine and seizure risk (insert in-depth paper hyperlink here).

2.3. Hypothesis testing

Customary test statistics are computed for both methods: a generalized score statistic for sequential regression adjustment using GS GEE [27] and a standardized Wald statistic (risk difference/sqrt(variance[risk difference]) for GS IPTW regression. Exact permutation testing is used to account for rare events and avoid incorrect large sample normality assumptions. A unifying family of sequential designs from group sequential randomized trials is used to compute the signaling threshold at each analysis time point and maintain the overall desired type 1 error [31]. This methodology gives investigators flexibility to choose the testing frequency (e.g., equally spaced over time or not), a calendar-time (e.g., monthly) versus information-time scale (e.g., number of new users since the previous analysis), and the signal threshold level at each analysis (e.g., constant over time [32] or higher at early versus later analyses [33]) based on scientific rationale specific to the safety question [34]. Planned sequential thresholds can also be modified over time to correctly control type 1 error while accounting for unpredictable changes in the data [35]. For instance, the unifying family approach can use the most up-to-date data on drug uptake and population characteristics at each analysis and correct the thresholds accordingly. Finally, unlike alpha-spending methods that define thresholds based on type 1 error, unifying family thresholds are on the scale of the risk measure (e.g., the relative risk) [36], which facilitates making design decisions about what threshold value should constitute a concerning level of excess risk using the interpretable

Table 1. Key features of the "whole cohort" sequential methods that implement either regression adjustment (for odds ratio and relative risk estimation) or inverse probability of treatment weighted linear regression (for risk difference estimation)

Method	Confounding adjustment	Target of inference	Data setting	How privacy is preserved
GS GEE (logistic model)	Regression adjustment for categorical levels of a propensity score or for categorical levels of individual confounders	Odds ratio (OR)	Binary outcome and exposure	Aggregate individual-level data at each health plan site and combine aggregated data for central analysis
GS GEE (Poisson model)		Relative risk (RR)	Person-time outcome and exposure	
GS IPTW (weighted linear model)	Inverse probability of treatment weighting (IPTW) with weights based on a propensity score	Risk difference (RD)	Binary outcome and exposure	Only the RD and its standard error from each health plan site is combined centrally. Meta-analytic Mantel Haenszel methods are used to form a combined site-stratified estimate.

Linear regression with binary outcomes for risk difference (RD) estimation is possible but has not been developed due to potential model instability. Weighted regression methods that estimate relative quantities (e.g., odds ratio or relative risk) using inverse probability of treatment weighting are also possible but not included for similar reasons.

quantity of interest [37]. Sequential design planning decisions, their rationale, and hypothesis testing details are considered in the context of the measles-mumps-rubella-varicella vaccine safety example in Section 2.2 of the related application paper in this issue (insert in depth paper hyperlink here).

2.4. Summary

Observational safety surveillance using health care databases raises a unique combination of challenges for sequential multivariable exposure-outcome estimation and testing: confounding, distributed data, rare events, dynamic data, and site heterogeneity. Both the sequential regression adjustment using GS GEE and sequential weighting methods using GS IPTW have been tailored to address these issues: (1) they employ analysis-based confounder control methods that facilitate use of the entire cohort, (2) they use deidentified data because individual-level data cannot generally be pooled across sites, (3) they perform exact testing to handle rare outcomes, (4) they adapt a sequential boundary methodology from clinical trials that allows flexible corrections over time to respond to dynamically updated outcome, exposure, and confounder data as it is captured by the health plan systems, and (5) they address heterogeneity across sites. It is also useful to note how these existing methods fit into a broader framework (Table 3) and can be easily extended. For instance, GS IPTW regression could be applied using aggregated data combined across sites (Appendix Table 2) if centralized analyses are preferred or if there are not enough data at some sites to fit site-specific risk difference models required for meta-analyses.

3. Performance of "whole cohort" group sequential regression methods

Because sequential methods have been primarily used in trials, relatively little is known about the sequential application and performance of regression adjustment using GS GEE and weighting using GS IPTW regression methods in an observational health care database safety setting like Sentinel. In this section, we summarize the current knowledge from simulation studies and data applications conducted sequentially in this setting.

3.1. Performance of testing approaches used by the whole cohort sequential regression methods

The two fundamental components of both the GS GEE regression and GS IPTW regression approaches are exact testing and a unifying family approach to compute sequential thresholds. Zhao et al. used a simulation evaluation and a vaccine data example to provide the first proof-of-concept combining these tools [38]. Nelson et al. demonstrated their flexibility in an unpredictable observational setting by applying them within the Vaccine Safety Datalink's dynamic distributed network [19]. This 2008–2011 study prospectively monitored the safety of a new vaccine among 149,337 children and addressed many unanticipated circumstances that can commonly occur when using electronic health care record data for safety surveillance [35]. For example, there was unanticipated differential vaccine uptake by age group and site. Furthermore, an unexpectedly large amount of previously missing data accrued at one time point due to an unforeseen data quality issue at one site that was later corrected, which resulted in one planned

Table 1 *(Continued)*

How site heterogeneity is addressed	Test statistic	Approach for rare events	Sequential threshold	Dynamic data updating
Health plan site is included as confounder in the model. Interactions between site and exposure or site and confounders can be included as needed.	Score test	Exact methods, using permutation testing	Unifying family, with thresholds computed via simulation. This flexible framework allows an investigator to choose the desired testing frequency and the signal threshold level at each time point from a wide range of options and based on scientific, ethical, and statistical factors.	Estimation: Regression models simply utilize the updated outcome, exposure, and confounder data in the regression model. Testing: Sequential thresholds can be updated over time to reflect unexpected differences between planned and actual sequential testing assumptions
A separate RD model is fit at each site allowing the exposure effect to vary flexibly by site. The combined RD estimate is site stratified and accounts for differences in propensity score variability by site (e.g., due to sample size differences by site)	Wald test			

analysis being skipped. Nelson et al. showed that the unifying family sequential testing framework could successfully address these unpredictable real-world data challenges while maintaining the desired type 1 error by recomputing the sequential thresholds at each analysis time point [19] using the most up-to-date observed vaccine uptake and confounders at each analysis [35].

3.2. Performance of regression adjustment using GS GEE for odds ratio and relative risk estimation and testing

The first studies to examine the performance of sequential regression adjustment method using GS GEE were done in Mini-Sentinel [9,39,40]. One was a proof-of-concept for implementing sequential Poisson regression using simulated data that emulated a chronically used drug with long-term event follow-up and censoring [39]. GS GEE using Poisson regression adjustment with exact testing and unifying family thresholds was compared to key existing methods: (1) the conditional sequential sampling procedure, which uses stratification to control confounding, exact testing, and alpha-spending [8] and (2) Poisson regression adjustment using established randomized trial testing methods: a large sample Wald test and alpha-spending [36]. All methods successfully detected a simulated true relative risk of two using data from five sites with rare events (i.e., 5% outcome prevalence) at the second analysis time point [39].

A follow-up simulation study more comprehensively evaluated sequential regression using GS GEE compared with these same alternative methods, with logistic instead of Poisson regression and varying outcome prevalence (1%, 5%, 10%), exposure prevalence (10%, 50%), and confounding strength (four scenarios with increasing confounding complexity) [9]. All three methods maintained the prespecified type 1 error, but the conditional sequential sampling procedure had lower power and slower signal detection times. In a reanalysis of 19,264 recipients of a pentavalent combination diphtheria and tetanus toxoids and acellular pertussis adsorbed, inactivated poliovirus, and Haemophilus b conjugate vaccine and 248,923 comparators from a multisite Vaccine Safety Datalink study [19], all three methods successfully detected the previously suggested fever signal (odds ratio = ~2), but the conditional sequential sampling procedure P was slower to signal (month 31) than both GS GEE regression with exact testing and unifying family thresholds (month 18) and the alpha-spending regression method (month 20) [9].

One additional study assessed the performance of GS GEE using sequential logistic regression in a multisite, observational rare event setting compared with exposure matching [41]. The matching approach used 1:1 propensity score matching, conditional logistic regression and, like sequential regression adjustment using GS GEE, it used exact testing (with a likelihood ratio test) and a unifying family boundary. Two trial-based alpha-spending approaches were also compared, one that used the whole cohort with regression adjustment and one that used a matched design. In simulations with quarterly tests, all methods had acceptable type I error. But, on average, regression adjustment using either GS GEE or trial-based alpha-spending was 10% more powerful, about 2 months faster at detecting true safety signals, and less prone to implementation difficulties than both matching methods.

Table 2. Advantages and limitations of ''whole cohort'' sequential methods that implement either regression adjustment (for odds ratio and relative risk estimation) or inverse probability of treatment weighted linear regression (for risk difference estimation)

Method	Advantages	Limitations
GS GEE (logistic or Poisson regression model to estimate a relative measure: odds ratio [OR] or relative risk [RR])	• Uses adverse event data from all eligible cohort members (i.e., from all exposed and comparators), maximizing power • Estimates an RR or OR, quantities well understood by epidemiologists • A flexible adjustment method that - Can adjust for individual confounders or for a summary score (e.g., categories or smooth function of the propensity score or disease risk score) - can assess heterogeneity by including interactions (e.g., with health plan site) - can easily conduct subgroup analyses • Is broadly applicable for short-term (binary) outcomes and exposures as well as chronic medication use and long-term (person-time) outcomes, which is a common setting for most drugs • Easy to include dynamic data updates over time. Modest updates to exposure and confounder data should yield modest changes to estimates/tests (since such updates don't alter who is in the cohort or what outcomes are informative)	• Requires an adequate number of events in exposed and unexposed groups, which are often rare (for the outcome model) • If adjustment for individual confounders is desired, - Number of confounders is limited if events are rare - Number of confounders is limited by the need to ensure deidentified aggregated data • Need to remove/exclude outlying observations (analogous process to trimming for propensity score matching and weighting approaches) • Current methods have not explored direct modeling of an RD measure for binary outcomes using linear regression
GS IPTW (inverse probability of treatment weighted linear regression to estimate a risk difference [RD])	• Uses adverse event data from all eligible cohort members, maximizing power • Estimates a risk difference (RD) quantity that is often of interest to policy-makers • A flexible adjustment method that - can easily conduct subgroup analyses - can adjust for a large number of confounders, even with rare events - strongly controls for site heterogeneity via a meta-analytic approach that allows exposure effects to vary by site • Directly accounts for differences in propensity score variability across sites • Uses an RD, which may be more stable and powerful than methods that make relative comparisons of risk • Easy to include dynamic data updates over time. Modest updates to exposure and confounder data should yield modest changes to estimates/tests (since such updates don't alter who is in the cohort or what outcomes are informative)	• Requires an adequate number of exposed and unexposed patients (for propensity score estimation) • Less well known compared to other confounder adjustment methods • Is currently designed for short-term exposure and acutely occurring outcomes (i.e., is not applicable for chronic drugs and long-term follow-up) • Need to trim or restrict the cohort to avoid including patients very unlikely or likely to receive exposure (i.e., those with large propensity score weights that can unduly inflate the variance)

Implementation difficulties were defined to occur when a planned analysis had to be skipped because of lack of accrual of informative new outcome data since the previous test. This problem intensified for matching methods as the frequency of testing increased because inherently fewer new users were available for matching at each new analysis (e.g., 15 out of 75 weekly analyses were skipped in one scenario). When reanalyzing data from the Vaccine Safety Datalink combination diphtheria and tetanus toxoids and acellular pertussis adsorbed, inactivated poliovirus, and Haemophilus b conjugate safety study [19], the GS GEE regression method successfully signaled (week 86), but the matching methods failed to signal before the end of surveillance (week 125). Estimated odds ratios also varied across the methods, with the adjusted odds ratios from GS GEE regression analyses stabilizing after week 71 (1.6–1.8) and the conditional odds ratios from matched analyses steadily decreasing (from 1.6 to 1.2) [41].

3.3. Performance of weighting using GS IPTW regression for risk difference estimation and testing

GS IPTW regression has been evaluated in simulation studies and data examples within Mini-Sentinel's

Table 3. General framework defining a class of methods for multivariable inference in a distributed observational setting

			Target of inference			
Use of information	**Confounder adjustment strategy**	**Distributed data method**	**Odds ratio**	**Relative risk**	**Risk difference**	**Hazard ratio**
Whole cohort	Regression	Data aggregation	GS GEE	GS GEE		
		Meta-analysis				
	Inverse weighting	Data aggregation				
		Meta-analysis			GS IPTW	
	Stratification	Data aggregation				
		Meta-analysis				
Subcohort	Matching	Data aggregation				
		Meta-analysis				

distributed network, including the application paper presented in this issue (insert in depth paper hyperlink) [10,42,43]. In the original report introducing this meta-analytic, stratified, inverse probability of treatment weighting approach, its performance was compared with two "gold standard" methods that allow individual-level data to be pooled across sites [10]: (1) a single, pooled data linear regression model adjusting for individual confounders and using robust standard errors and (2) a single, pooled data propensity score–weighted linear regression model with site included in the propensity score (i.e., nonstratified inverse probability of treatment weighting). Pooling is not typically feasible in a distributed data setting, but it should theoretically yield "gold standard" estimates because it can leverage the most granular level of information. A simulation study assessed method performance under a wide variety of scenarios for outcome prevalence (0.1%, 1%, 5%), exposure prevalence (25%, 50%), site sample sizes (equal vs. unequal), and strength of confounding by site. The somewhat surprising main finding was that GS IPTW (i.e., a stratified inverse probability of treatment weighted estimate) was just as powerful as nonstratified inverse probability of treatment weighting, even when no confounding by site was present and one would expect a nonstratified estimate to do better [10]. In an analysis of data from a prior Vaccine Safety Datalink study of the measles-mumps-rubella-varicella combination vaccine and seizure, both the stratified GS IPTW and nonstratified inverse probability of treatment weighting method identified the known signal at the second analysis after 6,944 measles-mumps-rubella-varicella vaccine doses. Regression adjustment for individual confounders did not signal until the 8th analysis and 47,831 measles-mumps-rubella-varicella vaccine doses [10].

In the companion application paper in this issue (insert in depth paper hyperlink) and in two more exhaustive Mini-Sentinel technical reports [42,43], exposure-outcome pairs with known positive associations were used to compare the performance of GS IPTW regression for risk difference estimation with estimation of a relative measure of risk (i.e., odds ratio or relative risk) using GS GEE regression adjustment. The paper in this issue presents results (in Section 3) from a sequential reanalysis of the association between the measles-mumps-rubella-varicella vaccine and seizure risk using data from four Mini-Sentinel data partners from February 2006 to July 2008 [42]. Both methods successfully detected the known safety signal, but GS IPTW method signaled at the fourth planned analysis after 1.75 years and 17,376 doses of measles-mumps-rubella-varicella vaccine while GS GEE logistic regression did not signal until the eighth analysis after 2.75 years and 48,233 doses. A similar reanalysis of the known positive association between the use of angiotensin-converting-enzyme inhibitors and the risk of angioedema was conducted using four data partners from June 2008 to December 2012 [43]. Although comparisons were limited by differences in the data availability at each site, the GS IPTW risk difference method signaled after 43,636 new users were observed while GS GEE logistic regression signaled after 111,329. In both examples, the risk difference approach needed considerably less (35%–40%) information to detect the known signal.

3.4. Summary

Sequential regression adjustment using GS GEE and sequential weighting using GS IPTW regression methods have performed well compared with other available methods that might be selected for use in sequential, rare event, observational surveillance [9,10,39–43]. With respect to power and signal detection speed, sequential regression adjustment using GS GEE typically outperformed approaches that used design-based confounder adjustment strategies (i.e., exposure matching and the conditional sequential sampling procedure with stratification) and performed similarly to or slightly better than established trial-based methods that make large sample assumptions and use alpha-spending. GS IPTW regression estimation of a risk difference has been shown to be equally strong or stronger than "gold standard" risk difference estimation approaches that leverage pooled individual level data. And, it was found to be more powerful and

yield substantially faster signal detection than sequential regression adjustment using GS GEE that estimates a relative measure of risk (i.e., odds ratio or relative risk).

4. Conclusions

We summarized existing knowledge about group sequential methods for safety surveillance that use the ''whole cohort'' by incorporating confounding control using regression adjustment [9] or inverse probability of treatment weighting [10]. Early studies of their performance show strong promise for their use, but gaps remain. More prospective implementation experience in distributed systems is needed to fully understand the impact of dynamic data updating on these approaches. Simulation studies are needed that provide clearer practical guidance on when, based on outcome prevalence, it is important to use the more computationally expensive exact sequential methods to ensure valid inference as opposed to using trial-based large sample techniques, which are simpler to implement but make incorrect assumptions when events are rare. In addition, there has not been a comprehensive simulation evaluation to assess site heterogeneity, propensity score model misspecification, or outcome misclassification. Finally, gaining a more comprehensive understanding of sequential analyses that directly monitor for safety signals on the risk difference scale would be particularly useful because policy-makers can more readily use a risk difference to compare risk and benefit, and the risk difference is also less variable to estimate (than relative measures) when events are rare, yielding higher power and faster signal detection.

This literature suggests some caution when using matching to control confounding in this setting. Practically, as the frequency of sequential testing is increased, there are inherently fewer new users available to be matched at each new analysis time point, which can reduce the number and/or quality of matches. If conditional models are used to account for the matching (i.e., where only discordant matched sets contribute information), as testing frequency increases more planned sequential analyses must be skipped because of a lack of new informative outcome data. Design-based confounder adjustment strategies like propensity score matching can also constrain downstream analysis options, for example, making subgroup analyses by age or gender less straightforward. And, estimates from an exposure matched design generalize to the population with a covariate distribution that is similar to the exposed group, which is a moving target and not a stable quantity over time. Regression (or stratification) methods avoid many of these implementation issues and can also increase power.

Several enhancements to these existing sequential regression approaches are worth pursuing. More flexible survival regression methods, such as Cox models designed for distributed data settings using regression or stratification, are needed to assess longer-term event follow-up. Sequential survival methods that estimate measures other than hazard ratios (e.g., survival curves) should also be pursued because relative measures can be unstable with rare events and proportional hazards assumptions may be too restrictive. It may also be useful to explore sequential methods that focus more heavily on estimation than testing, ''signaling'' when the effect estimate stabilizes to a prespecified degree (indicating that further surveillance would likely not yield additional information) as opposed to when it exceeds a predefined threshold of concern. Some of this work is currently being pursued [44].

CRediT authorship contribution statement

Jennifer C. Nelson: Conceptualization, Methodology, Investigation, Resources, Visualization, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Ernesto Ulloa-Pérez:** Conceptualization, Methodology, Investigation, Resources, Visualization, Writing - original draft, Writing - review & editing. **Jennifer F. Bobb:** Conceptualization, Methodology, Investigation, Resources, Visualization, Writing - original draft, Writing - review & editing. **Judith C. Maro:** Conceptualization, Methodology, Investigation, Resources, Visualization, Writing - original draft, Writing - review & editing, Funding acquisition.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.04.012.

References

[1] Whitehead J. The design and analysis of sequential clinical trials, revised. 2nd ed. Chichester, West Sussex: John Wiley & Sons Ltd; 2000:135−81.
[2] Baggs J, Gee J, Lewis E, Fowler G, Benson P, Lieu T, et al. The Vaccine Safety Datalink: a model for monitoring immunization safety. Pediatrics 2011;127(Suppl 1):S45−53.
[3] Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the sentinel system–a national resource for evidence development. N Engl J Med 2011;364:498−9.
[4] Nguyen M, Ball R, Midthun K, Lieu TA. The Food and drug Administration's post-licensure rapid immunization safety monitoring program: strengthening the federal vaccine safety enterprise. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):291−7.
[5] Hocine MN, Musonda P, Andrews NJ, Tubert-Bitter P, Farrington CP. Sequential case series for pharmacovigilance. J R Stat Soc Ser A Stat Soc 2009;172:213−36.
[6] Shih MC, Lai TL, Heyse JF, Chen J. Sequential generalized likelihood ratio tests for vaccine safety evaluation. Stat Med 2010;29:2698−708.
[7] Kulldorff M, Davis RL, Kolczak M, et al. A maximized sequential probability ratio test for drug and vaccine safety surveillance. Seq Anal 2011;30:58−78.

[8] Li L. A conditional sequential sampling procedure for drug safety surveillance. Stat Med 2009;28:3124−38.
[9] Cook AJ, Wellman RD, Nelson JC, Jackson LA, Tiwari RC. Group sequential method for observational data by using generalized estimating equations: application to the Vaccine Safety Datalink. J R Stat Soc Ser C Appl Stat 2015;64(2):319−38.
[10] Cook AJ, Wellman RD, Marsh T, Tiwari RC, Nguyen MD, Russek-Cohen E, et al. Statistical methods for estimating causal risk differences in the distributed data setting for postmarket safety outcomes. 2012. Available at http://www.mini-sentinel.org/methods/methods_development. Accessed October 5, 2016.
[11] Lieu TA, Kulldorff M, Davis RL, Lewis EM, Weintraub E, Yih K, et al. Real-time vaccine safety surveillance for the early detection of adverse events. Med Care 2007;45:S89−95.
[12] Yih WK, Nordin JD, Kulldorff M, Lewis E, Lieu TA, Shi P, et al. An assessment of the safety of adolescent and adult tetanus-diphtheria-acellular pertussis (Tdap) vaccine, using active surveillance for adverse events in the Vaccine Safety Datalink. Vaccine 2009;27:4257−62.
[13] Belongia EA, Irving SA, Shui IM, Kulldorff M, Lewis E, Yin R, et al. Vaccine Safety Datalink Investigation: real-time surveillance to assess risk of intussusception and other adverse events after pentavalent, bovine-derived rotavirus vaccine. Pediatr Infect Dis J 2010;29:1−5.
[14] Daley MF, Yih WK, Glanz JM, Hambidge SJ, Narwaney KJ, Yin R, et al. Safety of diphtheria, tetanus, acellular pertussis and inactivated poliovirus (DTaP-IPV) vaccine. Vaccine 2014;32:3019−24.
[15] Lee GM, Greene SK, Weintraub ES, Baggs J, Kulldorff M, Fireman BH, et al. H1N1 and seasonal influenza vaccine safety in the vaccine safety datalink project. Am J Prev Med 2011;41:121−8.
[16] Tse A, Tseng HF, Greene SK, Vellozzi C, Lee GM, VSD Rapid Cycle Analysis Influenza Working Group. Signal identification and evaluation for risk of febrile seizures in children following trivalent inactivated influenza vaccine in the Vaccine Safety Datalink Project, 2010-2011. Vaccine 2012;30:2024−31.
[17] Greene SK, Kulldorff M, Yin R, Yih WK, Lieu TA, Weintraub ES, et al. Near real-time vaccine safety surveillance with partially accrued data. Pharmacoepidemiol Drug Saf 2011;20:583−90.
[18] Tseng HF, Sy LS, Liu IL, Qian L, Marcy SM, Weintraub E, et al. Postlicensure surveillance for pre-specified adverse events following the 13-valent pneumococcal conjugate vaccine in children. Vaccine 2013;31:2578−83.
[19] Nelson JC, Yu O, Dominguez C, Cook AJ, Peterson D, Greene SK, et al. Adapting group sequential methods to observational post-licensure safety surveillance: results of a pentavalent combination DTaP-IPV-Hib (Pentacel) vaccine safety study. Am J Epidemiol 2013;177:131−41.
[20] Yih WK, Lee GM, Lieu TA, Ball R, Kulldorff M, Rett M, et al. Surveillance for adverse events following receipt of pandemic 2009 H1N1 vaccine in the Post-Licensure Immunization Safety Monitoring (PRISM) system, 2009-2010. Am J Epidemiol 2012;175:1120−9.
[21] Fireman B, Toh S, Butler MG, Go AS, Joffe HV, Graham DJ, et al. A protocol for active surveillance of acute myocardial infarction in association with the use of a new antidiabetic pharmaceutical agent. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):282−90.
[22] Carnahan RC, Gagne JJ, Nelson JC, Fireman B, Wang S, Shoaibi A, et al. Mini-sentinel prospective routine observational monitoring program tools (prompt): rivaroxaban surveillance. 2015. Available at http://www.mini-sentinel.org/work_products/Assessments/. Accessed October 5, 2016.
[23] Gagne JJ, Glynn RJ, Rassen JA, Walker AM, Daniel GW, Sridhar G, et al. Active safety monitoring of newly marketed medications in a distributed data network: application of a semi-automated monitoring system. Clin Pharmacol Ther 2012;92(1):80−6.
[24] Gagne JJ, Rassen JA, Choudhry NK, Bohn RL, Patrick AR, Sridhar G, et al. Near-real-time monitoring of new drugs: an application comparing prasugrel versus clopidogrel. Drug Saf 2014;37:151−61.
[25] Yih WK, Kulldorff M, Fireman BH, Shui IM, Lewis EM, Klein NP, et al. Active surveillance for adverse events: the experience of the Vaccine Safety Datalink project. Pediatrics 2011;127(Suppl 1):S54−64.
[26] Leite A, Andrews NJ, Thomas SL. Near real-time vaccine safety surveillance using electronic health records − a systematic review of the application of statistical methods. Pharmacoepidemiol Drug Saf 2016;25(3):225−37.
[27] Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf 2010;19:858−68.
[28] Breslow NE, Amroim G, Pettinger MB, Rossouw J. Using the whole cohort in the analysis of case-control data: application to the Women's Health Initiative. Stat Biosci 2013;5(2). https://doi.org/10.1007/s12561-013-9080-2.
[29] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13−22.
[30] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med 2004;23:2937−60.
[31] Kittelson JM, Emerson SS. A unifying family of group sequential test designs. Biometrics 1999;55:874−82.
[32] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191−9.
[33] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549−56.
[34] Emerson S, Kittelson JM, Gillen D. Tutorial in biostatistics: frequentist evaluation of group sequential clinical trial designs. Stat Med 2007;26:5047−80.
[35] Nelson JC, Cook AJ, Yu O, Zhao S, Dominguez C, Fireman B, et al. Challenges in the design and analysis of sequentially-monitored post-licensure safety surveillance studies using observational health care utilization data. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):62−71.
[36] Lan K, DeMets D. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659−63.
[37] Nelson JC, Wellman R, Yu O, Cook AJ, Maro JC, Ouellet-Hellstrom R, et al. A synthesis of current surveillance planning methods for the sequential monitoring of drug and vaccine adverse events using electronic health care data. EGEMS 2016;4(1):1219.
[38] Zhao S, Cook AJ, Jackson LA, Nelson JC. Statistical performance of group sequential testing methods for post-licensure medical product safety surveillance: a simulation study. Stat Interfaces 2012;5:381−90.
[39] Cook AJ, Wellman RD, Marsh T, Li L, Heckbert S, Heagerty P, et al. Statistical approaches to group sequential monitoring of post-marketing safety surveillance data: current state of the art for use in the Mini-Sentinel surveillance project. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):72−81.
[40] Cook AJ, Wellman RD, Tiwari RC, Li L, Heckbert S, Marsh T, et al. Statistical approaches for group sequential monitoring of postmarketing safety surveillance data. An Overview and simulation evaluation of methods applicable to the Mini-Sentinel surveillance pilot. 2011. Available at http://www.mini-sentinel.org/methods/methods_development. Accessed October 5, 2016.
[41] Stratton K, Cook AC, Jackson LA, Nelson JC, et al. Simulation study comparing exposure matching with regression adjustment in an observational vaccine safety setting with group sequential monitoring. Stat Med 2015;34:1117−33.

[42] Cook AJ, Wellman RD, Shaoibi A, Tiwari RC, Boudreau D, Goonesekera S, et al. Demonstrate feasibility of new Mini-Sentinel group sequential monitoring methods in a distributed setting by implementing them in practice. 2014. Available at http://www.mini-sentinel.org/methods/methods_development. Accessed October 5, 2016.

[43] Nelson JC, Boudreau D, Wellman R, Yu O, Cook AJ, Maro J, Ouellet-Hellstrom R, et al. Improving safety surveillance planning methods for routine assessments that use regression adjustment or weighting to control confounding. 2016. Availale at http://www.mini-sentinel.org/methods/methods_development. Accessed October 5, 2016.

[44] Cook AJ, Wellman RD, Shoaibi A, Tiwari RC, Heckbert SR, Li L, et al. Safety signaling methods for survival outcomes to control for confounding in the Mini-Sentinel Distributed Database. 2015. Available at http://www.mini-sentinel.org/methods/methods_development. Accessed October 5, 2016.