Methods Section for the disproportionality paper

1. Background: Disproportionality Analysis and Spontaneous Reports

Disproportionality analysis methods for drug safety surveillance represent the primary class of analytic methods for analyzing data from spontaneous report systems (SRSs). SRSs receive reports that comprise of one or more drugs, one or more adverse events (AEs), and possibly some basic demographic information (in addition to narrative and text data). Table 1 below shows a conceptual representation of a typical SRS entry.

Table 1: A conceptual representation of a typical entry in an SRS database

Age	*Sex*	*Drug 1*	*Drug 2*	…	*Drug 15000*	*AE 1*	*AE 2*	…	*AE 16000*
42	Male	No	Yes	…	No	Yes	No	…	Yes

Disproportionality analysis methods include the multi-item gamma-Poisson shrinker (MGPS), proportional reporting ratios (PRR), reporting odds ratios (ROR), and Bayesian confidence propagation neural network (BCPNN). The methods search SRS databases for "interesting" associations and focus on low-dimensional projections of the data, specifically 2-dimensional contingency tables. Table 2 shows a typical table.

Table 2: A fictitious 2-dimensional projection of an SRS database

	AE j = Yes	*AE j = No*	*Total*
Drug *i* = Yes	w_{00}=20	w_{01}=100	120
Drug *i* = No	w_{10}=100	w_{11}=980	1080
Total	*120*	*1080*	1200

The basic task of a DPA method then is to rank order the tables in order of "interestingness." Different DPA methods focus on different statistical measures of association as their measure of "interestingness". MGPS focuses on the "reporting ratio" (RR). The RR for the drug *i* – adverse event *j* combination (RR_{ij}) is the observed number of occurrences of the combination (20 in the example above) divided by the expected number

of occurrences. MGPS computes the expected value under a model of independence. Specifically, in the example above, overall, AE *j* occurs in 10% of the reports (120/1200). Thus, if drug *i* and adverse event *j* are statistically independent, 10% of the reports containing drug *i* should include AE *j*, that is 12 reports in this case. Thus the RR for this example is 20/12 or 1 2/3; this combination occurred about 67% more often than expected.

Natural (though not necessarily unbiased) estimates of various probabilities emerge from tables like Table 2. For example, one might estimate the conditional probability of AE *j* given drug *i* by $w_{00}/ w_{00}+ w_{01}$ (i.e. 20/120 in the example above). That is, the observed fraction of drug *i* reports that listed AE *j*. Table 3 below lists the formulae for the various measures of association in common use, along with their probabilistic interpretation. Here "$\neg drug$" for example denotes the reports that did not list the target drug. PRR is the "Proportional Reporting Ratio", ROR is the "Reporting Odds Ratio," and IC is the "Information Component" used by BCPNN. [2,3,6]

Table 3: Common measures of association for 2 X 2 tables in SRS analyses

Measure of Association	*Formula*	*Probabilistic Interpretation*
RR Reporting Ratio	$\frac{w_{00} * (w_{00} + w_{01} + w_{10} + d)}{(w_{00} + w_{10}) * (w_{00} + w_{01})}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae)}$
PRR Proportional Reporting Ratio	$\frac{w_{00} / (w_{00} + w_{01})}{w_{10} / (w_{10} + w_{11})}$	$\frac{\Pr(ae \mid drug)}{\Pr(ae \mid \neg drug)}$
ROR Reporting Odds Ratio	$\frac{w_{00} / w_{10}}{w_{01} / w_{11}}$	$\frac{\Pr(ae \mid drug)/\Pr(\neg ae \mid drug)}{\Pr(ae \mid \neg drug)/\Pr(\neg ae \mid drug)}$
Information Component	$\log_2 \frac{w_{00} * (w_{00} + w_{01} + w_{10} + w_{11})}{(w_{00} + w_{10}) * (w_{00} + w_{01})}$	$\log_2 \frac{\Pr(ae \mid drug)}{\Pr(ae)}$

All four of these measures make sense – in each case, a particular drug that is more likely to cause a particular AE than some other drug will typically receive a higher score. Similarly, if an AE and a drug are stochastically independent, all measures will return a null value. However, all four are subject to sampling variability, i.e. a different set of AE reports from the same "population" will not give exactly the same value of the measure of association.

This may be particularly the case with large sparse databases. Due to the Law of Large Numbers, this statistical variability diminishes as the sample size increases. In the SRS context, however, the count in the "w_{00}" cell is often small, leading to substantial variability (and hence uncertainty about the true value of the measure of association) despite the often large numbers of reports overall.

PRR and ROR do not address the variability issue whereas MGPS and BCPNN adopt a Bayesian approach to address the issue. GPS places a prior distribution on RRs that encapsulates a prior belief that most RRs are close to the average value of all RR's (i.e., close to 1). Only in the face of substantial evidence from the data does MGPS return an RR estimate that is substantially larger than one. Thus, for example, an RR of 1,000 that derives from an observed count of w_{00}=1 might result in a MGPS RR estimate (Empirical Bayesian Geometric Mean or EBGM) of 1.5 (i.e. the crude RR is shrunk towards a value of 1) whereas an RR of 1,000 that derives from an observed count of w_{00}=100 might result in a EBGM RR estimate of close to 1,000. For the specific Bayesian setup that MGPS uses, observed counts in excess of 10 result in RR estimates that typically receive essentially no shrinkage, although in practice larger differentials have been observed depending on the thresholds used. [5,7,8]

The EBGM score is the mean of the posterior distribution of the true RR. Other summaries are possible. For example, DuMouchel mentions "EB05". [9] This is the 5th percentile of the posterior distribution – meaning that there is a 95% probability that the "true" RR exceeds the EB05. Since EB05 is always smaller than EBGM this, in a sense, adds extra shrinkage and represents a more conservative choice than EBGM.

2. Computing the Disproportionality Metrics

Given a two-by-two table such as Table 2, the subsections below provide formulae for the various disproportionality metrics.

2.1 Proportional Reporting Ratio

$$PRR = \frac{w_{00}/w_{00} + w_{01}}{w_{10}/w_{10} + w_{11}}$$

2.2 Reporting Odds Ratio

$$ROR = \frac{w_{00}/w_{10}}{w_{01}/w_{11}}$$

2.3 MGPS

Let $w_{00}(i,j)$ denote the w_{00} entry for the two-by-two table for the ith drug and the jth condition. Assume that each $w_{00}(i,j)$ is a draw from a Poisson distribution with mean $\mu(i,j)$. Let $\mu(i,j) = \lambda(i,j)*E(i,j)$, where $E(i,j)=w_{0+}(i,j)*w_{+1}(i,j)/w_{++}(i,j)$, i.e., the expected value of $w_0(i,j)$ under independence and is assumed to be known. The goal is to estimate the values of the λ's. A $\lambda(i,j)$ far from one supports the notion that drug i and condition j are not independent. MGPS is a Bayesian procedure and starts with a particular five-parameter prior distribution for the collection of λ's:

$$\pi(\lambda;\alpha_1,\beta_1,\alpha_2,\beta_2,P) = Pg(\lambda;\alpha_1,\beta_1) + (1-P)g(\lambda;\alpha_2,\beta_2)$$

where $g(\lambda;\alpha,\beta)$ denotes a gamma density with α/β. The "EBGM" measure is defined as:

$$EBGM(i,j) = 2^{EB\log_2(i,j)}$$

where:

$$EB\log_2 = \left(Q_w\left[\psi(\alpha_1 + w_{00} - \log(\beta_1 + E)\right] + (1-Q_w)\left[\psi(\alpha_n + w_{00} - \log(\beta_2 + E)\right]\right)/\log(2)$$

$$Q_w = Pf(w_{00};\alpha_1,\beta_1,E)/\left[Pf(w_{00};\alpha_1,\beta_1,E) + (1-P)f(w_{00};\alpha_2,\beta_2,E)\right], \text{ and}$$

$$f(w_{00};\alpha,\beta,E) = (1+\beta/E)^{-w_{00}}(1+E/\beta)^{-\alpha}\Gamma(\alpha+w_{00})/\Gamma(\alpha)n!.$$

MGPS uses an empirical Bayes approach and chooses α_1, β_1, α_2, β_2, and P to maximize:

$$\prod_{i,j} Pf(w_{00}(i,j);\alpha_1,\beta_1,E(i,j)) + (1-P)f(w_{00}(i,j);\alpha_2,\beta_2,E(i,j)).$$

For further details see reference [9].

2.4 BCPNN

The BCPNN method estimates the Information Component (IC) as:

$$IC(i,j) = \log_2 \frac{w_{00}(i,j)+1/2}{E(i,j)+1/2}$$

For details see reference [1,2].

Many DPA analyses consider stratified versions of these metrics, stratifying by age, sex, and year of report, for example. See Appendix A for further details and specific formulae for stratified metrics.

3. Applying DPA to Longitudinal Data

In the context of spontaneous report systems, some authors use the term "signal of disproportionate reporting" (SDR) when discussing associations highlighted by DPA methods (Hauben et al., 2005, Hauben and Reich, 2005). In reality, most SDRs that emerge from spontaneous report databases represent noise because the reports are associated with treatment indications (i.e., confounding by indication), co-prescribing patterns, co-morbid illnesses, protopathic bias, channeling bias, or other reporting artifacts, or, the reported adverse events are already labeled or are medically trivial. In this sense, SDRs *generate* hypotheses. Furthermore, spontaneous report databases present a number of well-documented limitations such as under-reporting, over-reporting, and duplicate-reporting, they fail to provide a denominator – how many individuals are actually consuming drug, and generally have limited temporal information with regard to duration of exposure and the time order of exposure and condition (Hauben et al., 2005). The richer context of longitudinal data (such as claims databases or electronic health records) affords the possibility of more refined analysis to address some of these artifacts. Nonetheless, given the wide acceptance of DPA methods in pharmacovigilance, application of DPA methods to longitudinal data may prove useful.

A key step in the application of DPA methods to any data is the mapping of the data into drug-condition two-by-two tables. With longitudinal data many choices present themselves. In this paper we consider three particular approaches, "distinct patients," "SRS," and "Modified SRS."

In what follows we will illustrate the approaches using the example of Figure 1. Figure 1 shows three patients. Patient 1 consumed drug A during two separate drug eras. The patient experienced condition X three times during these eras, twice during the first era and once during the second. Patient 2 also had three drug eras but with three separate drugs, A, B, and

C. Finally Patient 3 had two overlapping drug eras, one with drug B and one with drug C. The patient experienced condition O while taking both B and C, and conditions O and X after the drug eras.

Note we treat conditions as if they occur at distinct moments in time. In fact the data may contain condition "eras" and what we are utilizing is the timestamp of the beginning of the era. Drug eras, on the other hand, play an important role in our approach. A drug era represents a continuous period of drug usage, possibly augmented with an additional off-drug period. We refer to the optional off-drug period as a surveillance window and discuss this further below. In practice, defining the on-drug portion of the drug era itself requires design decisions. For example, should two 30-day prescriptions with a one-day gap between the two prescriptions be considered one drug era or two? We return to this issue later.

We now consider three different approaches to constructing the two-by-two table for drug A and condition X. Appendix B provide a formal mathematical description.

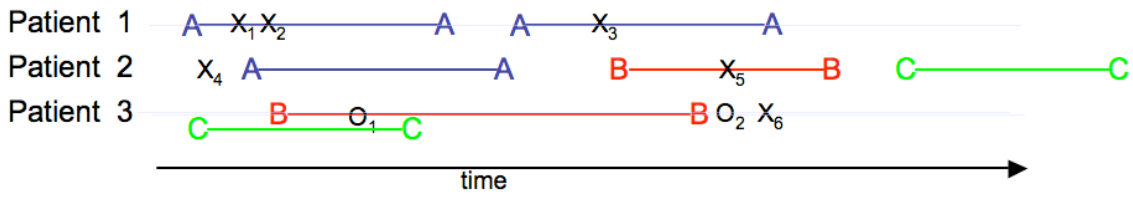

Figure 1. A longitudinal dataset with three patients, three distinct drugs (A, B, and C) and two distinct conditions (X and O).

3.1 Distinct Patients

In the "distinct patients" approach to table construction, $w_{00}+w_{01}+w_{10}+w_{11}$ (denoted w_{++}) equals the total number of patients in the database. w_{00} is the number of patients that had a drug A era and experienced condition X during a drug A era. w_{01} is the number of patients that had a drug A era and did not experience condition X during a drug A era. w_{10} is the number of patients that did not have a drug A era but experienced condition X. w_{11} is the number of patients that did not have a drug A era and never experienced condition X. Thus, for the example of Figure 1, w_{00}=1 (patient 1), w_{01}=1 (patient 2), w_{10}=1 (patient 3), and w_{11}=0. Note that w_{++}=3.

3.2 SRS

The second approach attempts to mimic what SRS reports the longitudinal data would generate. w_{00} is the number of distinct X conditions the occur during drug A eras. w_{01} is the number of distinct non-X conditions that occur during drug A eras. w_{10} is the number of distinct X conditions that occur during non-A drug eras. w_{11} is the number of distinct non-X conditions that occur during non-A drug eras. Thus, for the example of Figure 1, w_{00}=3 ($A+X_1$, $A+X_2$, $A+X_3$), w_{01}=0, w_{10}=1 ($B+X_5$), and w_{11}=2 ($B+O_1$,$C+O_1$).

3.3 Modified-SRS

The final approach attempts to mimic what SRS reports the longitudinal data would generate plus some additional counts that attempt to patch obvious weaknesses of the SRS approach. Specifically, in the SRS approach, drug eras in which no conditions occur are ignored, thereby discarding potentially useful information favoring the safety of a drug. Similarly, conditions that occur while no drug eras are active are also ignored. This also discards information that might exonerate a drug. The modified SRS approach counts "non-event" drug eras and "non-drug" conditions. In this approach, as with SRS, w_{00} is the number of distinct X conditions the occur during drug A eras. w_{01} however, is the number of distinct non-X conditions that occur during drug A eras plus the number of A eras in which no events occur. w_{10} is the number of distinct X conditions that occur outside drug A eras. w_{11} is the number of distinct non-X conditions that occur during non-A drug eras plus the number of non-A drug eras with no conditions plus the number of non-X conditions with no drug era. Thus, for the example of Figure 1, w_{00}=3 ($A+X_1$, $A+X_2$, $A+X_3$), w_{01}=1 (patient 2's A era), w_{10}=3 (X_4, $B+X_5$, X_6), and w_{11}=4 (patient 2's C era, $B+O_1$,$C+O_1$, O_2).

Our application of DPA to longitudinal data also makes a distinction between incident and prevalent conditions. The incident case only considers the first occurrence of each event, whereas the prevalent case (considered in the above example) considers all occurrences. Thus, for the example above, the incident analysis would proceed as above but only consider the first event of each type. Figure 2 illustrates the modified dataset used in an incident analysis. Note that our use of the term "incident" does necessarily coincide with standard

use in epidemiological practice/ In particular, we not require an event-free "clean" period prior to first condition occurrence.

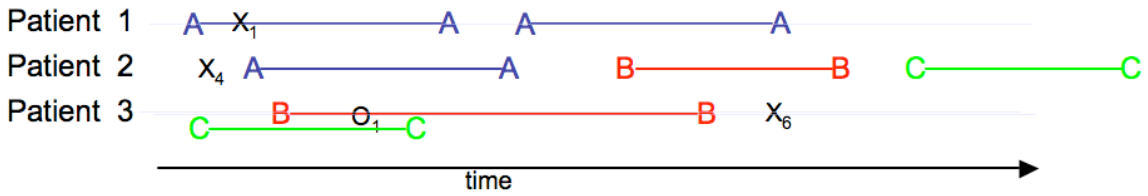

Figure 2. A longitudinal dataset with three patients, three distinct drugs (A, B, and C) and two distinct conditions (X and O). Incident conditions only.

References

1 NOREN, G.N., BATE, A., HOPSTADIUS, J., STAR, K., EDWARDS, I.R. (2008). Temporal pattern discovery for trends and transient effects: Its application to patient records.

2 BATE A, LINDQUIST M, EDWARDS IR et al.: A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol (1998) 54(4):315-321.

3 EVANS SJ, WALLER PC, DAVIS S: Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf (2001) 10(6):483-486.

4 FRAM D, ALMENOFF J, DUMOUCHEL W: Empirical Bayesian data mining for discovering patterns in post-marketing drug safety. (Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2003).

5 HAUBEN M, ZHOU X: A brief primer on automated signal detection Quantitative methods in pharmacovigilance: focus on signal detection. Ann Pharmacother (2003) 37(7-8):1117-1123.

6 SZARFMAN A, MACHADO SG, O'NEILL RT: Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf (2002) 25(6):381-392.

7 HAUBEN M, REICH L, ZHOU X: Safety related drug-labelling changes: findings from two data mining algorithms. Drug Saf (2004) 27(10):735-744.

8 HAUBEN M, ZHOU X: Trimethoprim-induced hyperkalaemia -- lessons in data mining. Br J Clin Pharmacol (2004) 58(3):338-339.

9 DUMOUCHEL W: Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat (1999) 53(3):170-190.

APPENDIX A

$i-$ stratum number, $i = 1, 2, ..., M$.

PRR:

$$PRR\_score = \frac{\sum_i W_{i00} * (W_{i10} + W_{i11}) / N_i}{\sum_i W_{i10} * (W_{i00} + W_{i01}) / N_i};$$

PRR05:

$$PRR05\_score = PRR * \exp(-1.645 * \sigma),$$

where $\sigma^2 = \frac{\sum_i ((W_{i00} + W_{i01})(W_{i10} + W_{i11})(W_{i00} + W_{i10}) - W_{i00} * W_{i10} * N_i) / N_i^2}{\sum_i W_{i00}(W_{i10} + W_{i11}) / N_i \sum_i W_{i10}(W_{i00} + W_{i10}) / N_i};$

ROR:

$$ROR\_score = \frac{\sum_i W_{i00} W_{i11} / N_i}{\sum_i W_{i10} W_{i01} / N_i};$$

ROR05:

$$ROR05\_score = ROR * \exp(-1.645 * \sigma),$$

where $\sigma^2 = \frac{\sum_i (W_{i00} + W_{i11}) W_{i00} W_{i11} / N_i^2}{2(\sum_i W_{i00} W_{i11} / N_i)^2} +$

$$\frac{\sum_i (W_{i10} + W_{i11}) W_{i01} W_{i10} + (W_{i01} + W_{i10}) W_{i00} W_{i11} / N_i^2}{2 \sum_i W_{i00} W_{i11} / N_i \sum_i W_{i01} W_{i10} / N_i} +$$

$$+ \frac{\sum_i (W_{i01} + W_{i10}) W_{i01} W_{i10} / N_i^2}{2(\sum_i W_{i01} W_{i10} / N_i)^2}$$

SIGNED CHI-SQUARE:

$$Signed\_Chi\_square\_score = sign(\sum_i W_{i00} - \sum_i (W_{i00} + W_{i01})(W_{i00} + W_{i10}) / N_i) *$$

$$* \frac{\left(\sum_i W_{i00} - \sum_i (W_{i00} + W_{i01})(W_{i00} + W_{i10}) / N_i\right)^2}{\sum_i \left((W_{i00} + W_{i01})(W_{i10} + W_{i11})(W_{i00} + W_{i10})(W_{i01} + W_{i11}) / (N_i^2 (N_i - 1))\right)};$$

BCPNN:

$$BCPNN\_score = \log_2\left(\frac{\sum_i W_{i00} + 1/2}{\sum_i \frac{(W_{i00} + W_{i01})(W_{i00} + W_{i10})}{N_i} + 1/2}\right);$$

BCPNN05:

$BCPNN05\_score = \log_2(z)$, where z is a solution to the following equation:

$$\int_0^z \frac{(e + 1/2)^{n+1/2}}{\Gamma(n + 1/2)} \lambda^{n+1/2-1} e^{-(n+1/2)\lambda} d\lambda = 0.05,$$

where $e = \frac{(W_{i00} + W_{i01})(W_{i00} + W_{i10})}{N_i}, n = \sum_i W_{i00};$

EBGM05:

is a solution to the following equation:

$$\int_0^{EBGM05\_score} \pi(\lambda, \alpha_1 + n, \beta_1 + e, \alpha_2 + n, \beta_2 + e, Q_n) d\lambda = 0.05,$$

where

$$\pi(\lambda, \alpha_1 + n, \beta_1 + e, \alpha_2 + n, \beta_2 + e, Q_n) =$$

$$= Pg(\lambda, \alpha_1, \beta_1) + (1 - P)g(\lambda, \alpha_2, \beta_2),$$

$$g(\lambda, \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda},$$

$$Q_n = \frac{Pf(n, \alpha_1, \beta_1, e)}{Pf(n, \alpha_1, \beta_1, e) + (1-P) f(n, \alpha_2, \beta_2, e)},$$

$$f(n, \alpha, \beta, e) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha) n!} (1 + \frac{\beta}{e})^{-n} (1 + \frac{e}{\beta})^{-\alpha},$$

$$e = \frac{(W_{i00} + W_{i01})(W_{i00} + W_{i10})}{N_i}, n = \sum_i W_{i00};$$

APPENDIX B

This appendix provides a formal mathematical definition of the alternative methods of constructing two-by-two tables from longitudinal data for DPA.

Notation

Let $y_{ict} = 1$ if patient *i* has condition *c* at time *t*, and $y_{ict} = 0$ otherwise, *i*=1,…,*I*, *c*=1,…*C*, and *t*=1,…,*T*. Let $x_{idt} = 1$ if patient *i* "takes" drug *d* at time *t*, and $x_{idt} = 0$ otherwise, *i*=1,…,*I*, *d*=1,…*D*, and *t*=1,…,*T*. This may include the user-defined off-drug "surveillance" window. Let $y^*_{ict} = 1$ if $y_{ict} = 1$ and $y_{ics} = 0$ for all $s < t$, 0 otherwise. Let $z_{it} = 1$ if patient *i* has coverage at time *t*, 0 otherwise.

Let D_{id} be the set of all ordered pairs (*r*,*s*), *r*,*s* ∈ {1,…,*T*}, where $x_{idr} = 1$ and (*r*=1 or $x_{id(r-1)}$=0), $x_{ids} = 1$ and (*s*=*T* or $x_{id(s+1)}$=0), and $y_{ict} = 0$ for all *c* and all *t* ∈ [*r*,…,*s*], and $z_{it} = 1$ for all *t* ∈ [*r*,…,*s*]. This defines condition-free periods of continuous drug exposure.

Define *I*(*x*) = 1 if *x*>0, 0 otherwise.

Prevalent Conditions, Distinct Patients

$$w_{00} = \sum_i I(\sum_t x_{idt} y_{ict})$$

$$w_{01} = \sum_i \left\{ I(\sum_t z_{it} x_{idt}) - I(\sum_t x_{idt} y_{ict}) \right\}$$

$$w_{10} = \sum_i \left\{ (1 - I(\sum_t z_{it} x_{idt})) \times I(\sum_t z_{it} y_{ict}) \right\}$$

$$w_{11} = \sum_i \left\{ (1 - I(\sum_t z_{it} x_{idt})) \times (1 - I(\sum_t z_{it} y_{ict})) \right\}$$

Prevalent Conditions, SRS

$$w_{00} = \sum_i \sum_t x_{idt} y_{ict}$$

$$w_{01} = \sum_i \sum_t \sum_{c' \neq c} x_{idt} y_{ic't}$$

$$w_{10} = \sum_i \sum_t \sum_{d' \neq d} x_{id't} y_{ict}$$

$$w_{11} = \sum_i \sum_t \sum_{d' \neq d} \sum_{c' \neq c} x_{id't} y_{ic't}$$

Prevalent Conditions, Modified-SRS

$$w_{00} = \sum_i \sum_t x_{idt} y_{ict}$$

$$w_{01} = \sum_i \left\{ \sum_t \sum_{c' \neq c} x_{idt} y_{ic't} + |D_{id}| \right\}$$

$$w_{10} = \sum_i \left\{ \sum_t \left\{ \sum_{d' \neq d} x_{id't} y_{ict} + y_{ict} z_{it} (1 - I(\sum_d x_{idt})) \right\} \right\}$$

$$w_{11} = \sum_i \left\{ \sum_t \sum_{d' \neq d} \sum_{c' \neq c} x_{id't} y_{ic't} + \sum_{d' \neq d} |D_{id'}| + \sum_t \sum_{c' \neq c} z_{it} y_{ict} (1 - I(\sum_d x_{idt})) \right\}$$

For incident conditions replace y_{ict} in the above definitions by y^*_{ict}.