SAS®-Makro zur automatisierten Änderungskontrolle von SAS®-Datensätzen und SAS®-Outputs

Claudia Meurer
Accovion GmbH
Helfmann-Park 10
65760 Eschborn
claudia.meurer@accovion.com

Gabi Lückel
Accovion GmbH
Helfmann-Park 10
65760 Eschborn
gabi.lueckel@accovion.com

Zusammenfassung

Die Erstellung qualitativ hochwertiger und robuster statistischer Analysen von klinischen Studien innerhalb enger Zeitvorgaben ist die wichtigste Aufgabe und gleichzeitig immer wieder eine Herausforderung für die Biometrie. Während die klinischen Daten erst nach und nach verfügbar sind, der statistische Analyseplan und die Tabellenlayouts noch diskutiert werden, müssen bereits die notwendigen SAS®-Programme geschrieben werden. Nach Datenbankschluss, also auf Basis der finalen klinischen Daten, bleiben in der Regel nur noch wenige Tage Zeit, um die Analysen fertig zu stellen. Die Sicherung der Qualität muss in jedem Fall gewährleistet sein, auch bei noch so engen Timelines. Die Validierung von SAS®-Programmen und SAS®-Outputs ist entscheidend für die Qualitätssicherung, aber auch sehr zeitaufwändig, da es sich für die Auswertung einer klinischen Studie meist um eine Vielzahl von SAS®-Programmen, eine zweistellige Zahl von Analysedatensätzen sowie Hunderte von mehrseitigen Tabellen, Listen und Grafiken handelt. In der Praxis ist es kaum vermeidbar, dass bis kurz vor Abgabe der finalen Ergebnisse aus diversen Gründen noch Änderungen an SAS®-Programmen vorgenommen werden. Auch nach dem finalen Programmlauf können Änderungswünsche des Kunden eine neue Generierung der Outputs notwendig machen. Eine vollständige Revalidierung aller bereits geprüften Outputs ist ohne technische Unterstützung praktisch unmöglich. Trotzdem ist die Revalidierung unerlässlich, um zum Beispiel unbeabsichtigte Änderungen zu entdecken. Bei Accovion wurde ein SAS®-Makro entwickelt, das

- bestehende SAS®-Analysedatensätze (ADS) und Tabellen/Listen/Grafiken (TLG) archiviert
- alle SAS®-Programme ausführt, die die ADS sowie alle TLG innerhalb eines Projektes generieren
- eine programmatische Änderungskontrolle zwischen neuen und archivierten ADS und TLG durchführt und in Statusberichten darstellt

Schlüsselwörter: Änderungskontrolle, automatisiert, Validierung, Makro, Datensatz, Output

1 Einleitung

Für die Analyse einer klinischen Studie fallen typischerweise Hunderte von SAS®-Outputs an, dazu gehören Analysedatensätze, Tabellen, Listen und Grafiken. Die Entwicklung der SAS®-Programme und deren Validierung beginnen schon weit vor Datenbankschluss, damit die Ergebnisse so bald wie möglich nach Datenbankschluss validiert

vorliegen. Jede Änderung an Daten oder SAS®-Programmen soll oder kann unbeabsichtigt zu einem veränderten SAS®-Output führen. Deshalb ist es wichtig, die Output-Änderungen schnell und automatisch zu identifizieren. Dazu wurde bei Accovion ein SAS®-Makro entwickelt, das mit einem einzigen Lauf alle Änderungen aufzeigt und dokumentiert. Das Makro RMC (**R**erun Programs and **M**ark **C**hanges) kombiniert den automatischen Neulauf aller SAS®-Programme mit der Möglichkeit, die vorher erstellten SAS®-Outputs zu sichern und diese mit den neu erstellten zu vergleichen. RMC liefert als Ergebnis eine Übersicht aller SAS®-Outputs, die sich geändert haben, und listet die Änderungen im Einzelnen an. Sind alle SAS®-Programme eines Projektes bereits validiert, ist RMC ein ideales Tool, um nachfolgende Änderungen an den SAS®-Outputs zu dokumentieren. Alle Änderungen können leicht nachvollzogen werden, ungewollte Änderungen werden sofort entdeckt. Andererseits ist auch dokumentiert, wenn sich im Vergleich zur vorhergehenden Version nichts geändert hat.

2 Entstehung des Makros

Das SAS®-Makro RMC ist zur Unterstützung der täglichen Arbeitsabläufe entstanden und immer weiter entwickelt worden. Die erste Fassung beinhaltete den Start aller SAS®-Programme eines Projektes in einem Batch-Lauf (der berühmte „grüne“ Knopf) und die Suche nach signifikanten Schlüsselwörtern („ERROR“, „WARNING“, „0 observations“...) in den entstandenen SAS®-Logfiles. In der nächsten Entwicklungsstufe des Makros wurden alte Output-Versionen mit den neu erstellten Outputs abgeglichen. In der täglichen Anwendung kamen immer neue Anforderungen der Benutzer (SAS®-Programmierer und Biostatistiker) hinzu, die nach und nach umgesetzt wurden.

3 Systemvoraussetzungen

Das Makro RMC wurde bei Accovion unter Unix in SAS® entwickelt. Im RMC werden SAS®-Programmcode und Unix-spezifische Kommandos kombiniert. Beispiele für im RMC angewendete Unix-Kommandos:

- *sas* test.sas - um ein SAS®-Programm auszuführen
- *grep* test.log „XYZ“ - um nach bestimmten Schlüsselwörtern oder Strings in den Outputs zu suchen
- *diff* test1.lst test2.lst - um zwei Outputs miteinander zu vergleichen und sich die Unterschiede auflisten zu lassen

Die Unix-Kommandos werden mittels „x“-Kommando aus SAS® heraus gestartet. Das RMC-Makro ist auf die Accovion-spezifische Unix-Verzeichnisstruktur angepasst. Es ist wichtig, dass der User Schreibrechte auf bestimmte Verzeichnisse hat, um das RMC-Makro fehlerfrei laufen zu lassen.

4 Funktionalität

Das Makro besteht aus drei unabhängigen Modulen:

- Archivierung der SAS®-Analysedatensätze (ADS) und Tabellen, Listen, Grafiken (TLG)
- Erstellung der neuen ADS und TLG
- Änderungskontrolle durch Abgleich der vorhergehenden mit den neuen Outputs

Diese drei Module generieren kurze und präzise Zusammenfassungen, die sogenannten „RMC Reports“. Hieraus wird ersichtlich, ob die gewünschten Output-Änderungen erfolgt sind oder dass keine Änderungen zu Vorversionen auftraten. Diese Reports können später auch zur Dokumentation, dass alles geprüft und für valide befunden wurde, herangezogen werden. Die Module des RMC-Makros können unabhängig voneinander oder in verschiedenen Kombinationen gestartet werden.

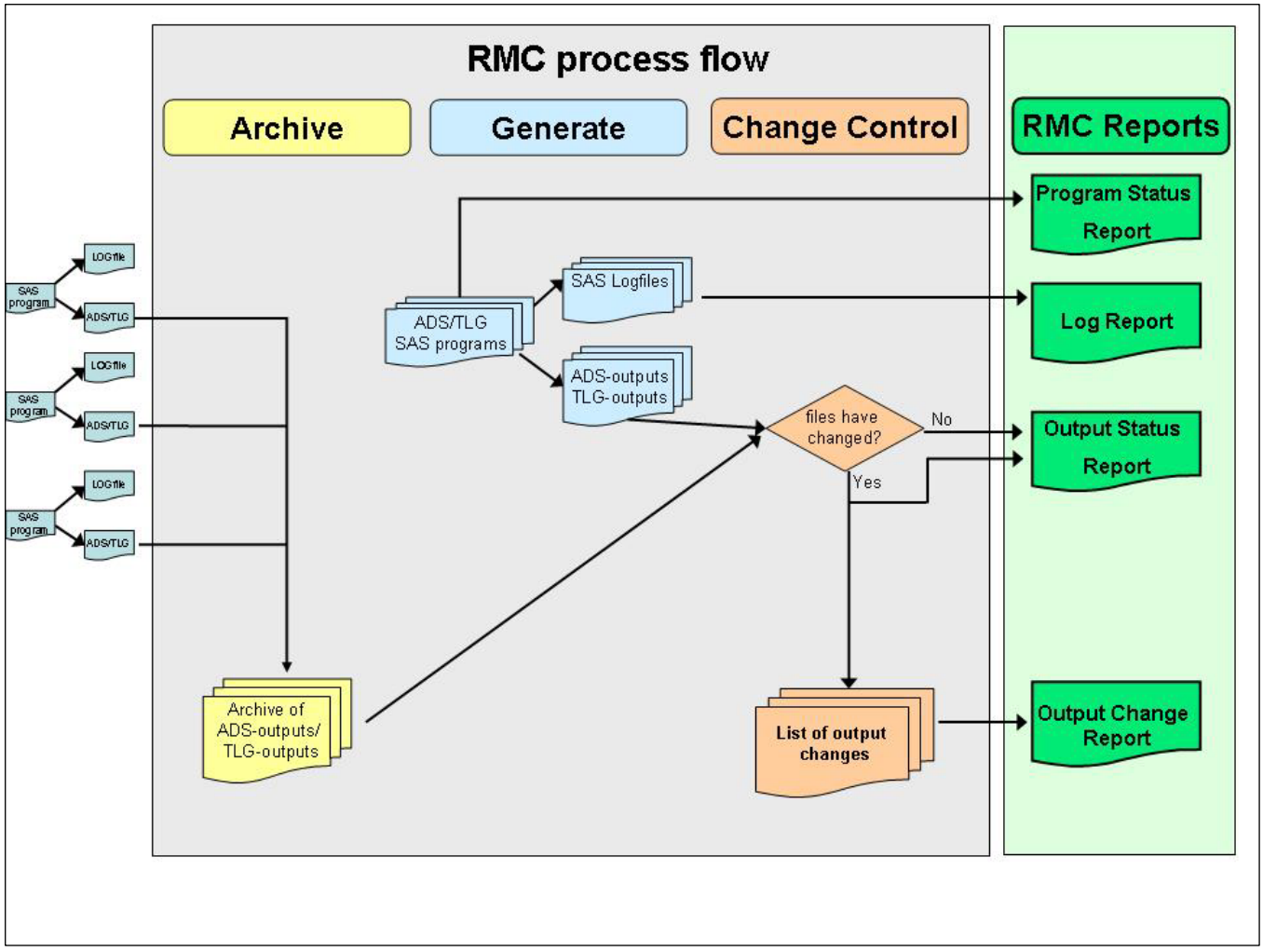

Abbildung 1: RMC Prozessablauf

RMC unterstützt die Archivierung, Erstellung und den Vergleich von SAS®-Datensätzen genauso wie von Tabellen, Listen und (bis zu einem gewissen Maße auch) Grafiken. Bis auf wenige Ausnahmen ist die Programmlogik der drei Module für ADS und

TLG gleich und wird deshalb hier allgemein beschrieben. Jedes Modul wird durch die Angabe verschiedener Optionen (Makroparameter) gesteuert und ist im Folgenden einzeln beschrieben.

4.1 Archivierung

Die Makroparameter SAVE_ADS und SAVE_TLG steuern die Möglichkeit, existierende Versionen von ADS und/oder TLG zu speichern. Gegen diese vorhergehende Versionen kann später der Abgleich mit den neu erstellten Versionen erfolgen. Dabei werden auf dem Server automatisch Unterverzeichnisse erstellt, die im Namen das aktuelle Datum mit Uhrzeit und den Zusatz „ADS“ oder „TLG“ enthalten. Dieses Vorgehen erlaubt einen sicheren und chronologischen Archivierungsprozess alter Versionen, ein zufälliges Überschreiben der Verzeichnisse ist ausgeschlossen.

Parameter	Funktionalität	Mögliche Werte
SAVE_ADS	Sichern der Datensätze durch Verschieben in ein neu erstelltes Archivierungsverzeichnis	YES/NO
SAVE_TLG	Sichern der Tabellen, Listen, Grafiken durch Verschieben in ein neu erstelltes Archivierungs- verzeichnis	YES/NO
KEEP_ADS	Definiert einzelne Datensätze, die nicht ins Archivierungsverzeichnis verschoben werden sollen	*<data set(s)>*

Bei der Archivierung werden alle vorliegenden ADS/TLG in das Archivierungsverzeichnis verschoben, nicht kopiert. Das bedeutet auch, dass die Outputs der nachfolgend ablaufenden SAS®-Programme in einem nun „leeren“ Verzeichnis abgelegt werden. Damit wird noch einmal geprüft, ob voneinander abhängige SAS®-Programme in der richtigen Reihenfolge laufen, wenn zum Beispiel die Existenz eines SAS®-Datensatzes für die Erstellung eines anderen Voraussetzung ist. Außerdem werden nicht mehr notwendige Outputs nicht neu generiert. Möchte man allerdings einzelne existierende ADS behalten, kann man diese mit dem Makroparameter KEEP_ADS spezifizieren. Das ist dann sinnvoll, wenn ein ADS nicht neu erstellt wird, zum Beispiel bei Format- oder Randomisierungsdatensätzen.

4.2 Erstellen von ADS/TLG

Eine weitere Funktionalität des RMC-Makros ist die kontrollierte Ausführung verschiedener SAS®-Programme im Batch-Lauf. Dies erhöht die Transparenz der Auswertung, da alle SAS®-Programme, die zu einem Projekt laufen sollen, in der festgelegten Reihenfolge angegeben sind. Mit den Makroparametern RUN_ADS und RUN_TLG wird gesteuert, ob die angegebenen SAS®-Programme ausgeführt werden sollen.

Parameter	Funktionalität	Mögliche Werte
RUN_ADS	Ausführen aller SAS®-Programme, die ADS erstellen	YES/NO
RUN_TLG	Ausführen aller SAS®-Programme, die TLG erstellen	YES/NO

Dieses Modul des RMC-Makros findet bereits in der Phase der frühen Programmentwicklung Verwendung. So kann „auf Knopfdruck“ die Erstellung der neuen ADS/TLG gestartet werden, wenn neue Rohdaten vorliegen.

```
*----------------------------------------;
* ADS programs                           ;
*----------------------------------------;
** Note, that programs have a defined sequence to be called **;
** NOTE: ADSL has to run as first program **;
   %runpg(prg= der/adsl);
   %runpg(prg= der/adcm);
   %runpg(prg= der/addm);
   ...
   %runpg(prg= der/admh);
```

Abbildung 2: Beispiel für die kontrollierte Ausführung mehrerer SAS®-Programme

Das Makro RUNPG wird innerhalb des RMC-Makros aufgerufen. Es startet das entsprechende SAS®-Programm und erstellt ein SAS®-Logfile, das automatisch mithilfe von Unix-Befehlen nach kritischen Schlüsselwörtern, wie „ERROR“, „WARNING“, „duplicate“, „missing“ und relevanten „NOTE“s durchsucht wird. Diese Suchergebnisse werden im „Log Report“ zusammengefasst und müssen vom SAS®-Programmierer im Einzelnen geprüft werden.

```
********************/.../der/adsl.log********************
---------- ERROR ----------------------------------
---------- FATAL ----------------------------------
---------- WARNING --------------------------------
---------- INFO: (user defined messages) ------------
********************/.../der/addm.log********************
---------- ERROR ----------------------------------
1320: ERROR: Variable VISITNUM not found.
---------- FATAL ----------------------------------
---------- WARNING --------------------------------
---------- NOTE: Variable ... is uninitialized ------
1410:NOTE: Variable age_group is uninitialized.
---------- NOTE: Missing values were generated ... --------
624:NOTE: Missing values were generated as a result of performing
an operation on missing values.
---------- INFO: (user defined messages) ------------
2556:INFO: Age does not meet the inclusion criteria:
subjid=0002489 age=17
```

Abbildung 3: Auszug aus dem „Log Report“

Wie man dem „Log Report“ entnehmen kann, ist das SAS®-Programm *adsl.sas* fehlerfrei gelaufen, während *addm.sas* vom SAS®-Programmierer noch überarbeitet werden muss.

SAS®-Programme werden im Laufe des Validierungsprozesses bei Accovion üblicherweise in einer Entwicklungsumgebung (Unterverzeichnis auf der Unix) erstellt, bevor sie dann in einem anderen Unterverzeichnis validiert werden. Ist der Validierungsprozess abgeschlossen, wird das SAS®-Programm in die Produktivumgebung geschoben. Das Makro RUNPG durchsucht diese drei Unterverzeichnisse nach dem angegebenen SAS®-Programm, im „Status Report“ des RMC-Makros erscheint dann der Status jedes SAS®-Programms.

```
Status                  Program name

Productive              der/adsl.sas
                        der/adcm.sas
                        ...
Under development       der/addm.sas

Under validation        der/adae.sas

Program missing         der/adpv.sas
```

Abbildung 4: Auszug aus dem „Program Status Report“

4.3 Änderungskontrolle

Der Schwerpunkt des RMC-Makros liegt auf der programmatischen Unterstützung der SAS®-Programmierer bei der Änderungskontrolle von SAS®-Outputs.
RMC dokumentiert die Änderungen zwischen den zu vergleichenden Outputs. Ob diese Änderungen korrekt sind, kann nur der verantwortliche SAS®-Programmierer feststellen. Der SAS®-Programmierer kann im „Output Change Report" gezielt nachschauen, wo sich etwas geändert hat. Wenn die SAS®-Programme und Outputs bereits validiert waren, sollten keine Änderungen nach einem erneuten Batch-Lauf auftreten. Wird dies vom RMC auch so dokumentiert, kann der SAS®-Programmierer alle neu erstellten ADS/TLG als validiert betrachten.
Unter anderem werden folgende Makro-Parameter von RMC benutzt, um Versionen von ADS/TLG zu vergleichen:

Parameter	Funktionalität	Mögliche Werte
COMP_ADS	Vergleich aller Datensätze	YES/NO
COMPDIR_ADS	Definiert ein Unix-Verzeichnis, auf dem die Datensätze abgelegt sind, mit denen die neu erstellten Datensätze verglichen werden sollen. Ist hier kein Unix-Verzeichnis angegeben, dient zum Vergleich das automatisch erstellte Archivierungsverzeichnis.	<*directory*>
COMP_TLG	Vergleich aller TLG	YES/NO
COMPDIR_TLG	Definiert ein Unix-Verzeichnis, auf dem die TLG abgelegt sind, mit denen die neu erstellten TLG verglichen werden sollen. Ist hier kein Unix-Verzeichnis angegeben, dient zum Vergleich das automatisch erstellte Archivierungsverzeichnis.	<*directory*>
BLOCK	Spezifiziert einen Character String, mit dem der Vergleich auf einen Teil der TLG beschränkt werden kann. (Bsp.: BLOCK= ae führt das RMC nur für TLG aus, die mit „ae" beginnen)	<*subset*>
OUTNAME	Spezifiziert einen Character String, mit dem die Bezeichnungen der generierten Reports ergänzt werden. (Bsp.: OUTNAME= aenew erzeugt u.a. rmc_tlg_comp_aenew_status.lst)	<*char*>

Neue ADS/TLG werden üblicherweise mit vorhergehenden Versionen verglichen. Wenn man neue gegen bereits validierte Versionen, die in einem speziellen Unterverzeichnis abgelegt sind, vergleichen will, kann man dieses Verzeichnis mit den Makrovariablen COMPDIR_* definieren (zum Beispiel test_study/validated/… oder test_study/backup_20090121/…).

4.3.1 Änderungskontrolle bei Datensätzen

Der Abgleich von Datensätzen erfolgt innerhalb des RMC-Makros mit SAS®-Code, genauer gesagt mit PROC COMPARE. RMC liest alle einzelnen SAS®-Outputs von PROC COMPARE und extrahiert die relevanten Informationen. Diese werden zusammengefasst und in einen „Output Status Report“ geschrieben.
Das folgende Beispiel zeigt den „Output Status Report“ mit der Anzahl der Variablen im vorherigen (OLDVAR) und im neuen Datensatz (NEWVAR), die Anzahl der Beobachtungen im vorherigen (OLDOBS) und neuen Datensatz (NEWOBS). Außerdem werden die Anzahl der Variablen (COMVAR) und die Anzahl der Beobachtungen (COMOBS), die in beiden Datensätzen vorhanden sind, angegeben.

Data Set Overview

FILE	OLDVAR	NEWVAR	COMVAR	OLDOBS	NEWOBS	COMOBS	EQUAL	CHANGE	STATUS
ADAE	69	75	69	208	379	208	Y	*	
ADEF	26	26	26	456	456	456	Y		
ADDM	69	69	69	484	484	484	N	*	
ADPV									NEW
...									

Abbildung 5: Auszug aus dem „Output Status Report“ für ADS

Die Spalte EQUAL gibt an, ob es zwischen den Variablen und Beobachtungen, die in beiden Datensätzen vorhanden sind, Unterschiede gibt oder nicht. Die Spalte CHANGE markiert mit einem Stern, wenn es Unterschiede bei Variablen, Beobachtungen oder Dateninhalten zwischen den beiden Datensätzen gibt. Auch wenn alle Datensätze eines Projektes kurzfristig neu generiert werden müssen, kann man sich anhand des „Output Status Report“ schnell einen Überblick über alle Änderungen verschaffen.
Inhaltliche Änderungen an den Daten (im obigen Beispiel für Datensatz ADDM) werden im „Output Change Report“ dokumentiert. Hier sind detailliert die Unterschiede zwischen den Datensätzen angegeben:

```
          Comparison of data sets against data sets in
                 /sas_data4/VALIDATION/RMC/dds

                      The COMPARE Procedure
              Comparison of OLDDDS.ADDM with DDS.ADDM
                          (Method=EXACT)

                        Data Set Summary
Data set              Created          Modified  NVar    NObs
OLDDDS.ADDM  11SEP08:13:20:51  11SEP08:13:20:51    20     440
DDS.ADDM     11SEP08:13:20:51  11SEP08:13:20:51    20     440
...
Number of Variables in Common: 20.
...
Number of Observations in Common: 440.
...
Number of Variables Compared with All Observations Equal: 19.
Number of Variables Compared with Some Observations Unequal: 1
Total Number of Values which Compare Unequal: 421.
...

Variables with Unequal Values

Variable  Type  Len  Ndif   MaxDif
AGE       NUM     8   421    0.695

            Value Comparison Results for Variables
                 Base    Compare
Obs               AGE       AGE       Diff.     % Diff
      1       58.0000    57.7522    -0.2478    -0.4272 ...
```

Abbildung 6: Auszug aus dem „Output Change Report“ für ADS

4.3.2 Änderungskontrolle bei TLG

Beim Abgleich von Tabellen und Listen mit dem RMC-Makro werden als erstes die Existenz der alten und neuen Versionen verglichen. Für jedes File wird geprüft, ob es als alte und neue Version vorhanden ist. Die Ergebnisse dieses Abgleichs werden wie folgt dokumentiert:

```
Comparison of outputs against outputs in $SAS.../out/validated

File Overview - Equal Files
Filename    File Status    Equal
dm001t             Y
ae023l             Y

File Overview - Changed, New, and Missing Files
Filename       File Status    Equal
dm001t_test       MISSING
dm011t              NEW
ae002l                          N
```

Abbildung 7: Auszug aus dem „Output Status Report“ für TLG

Aus dem „Output Status Report“ ist ersichtlich, dass die vorherige Version dm001t_test nicht neu erzeugt wurde, dass eine neue Tabelle dm011t hinzugekommen ist, die es vorher noch nicht gab, und dass es Unterschiede für das File ae002l zwischen der vorherigen und der aktuellen Version gibt. Oftmals ist mit bloßem Auge nicht erkennbar, wo die Unterschiede zwischen den Outputs sind. Vor allem bei großen Dokumenten ist diese Arbeit manuell nicht zu leisten, weil sie viel zu zeitaufwändig wäre. Das RMC vergleicht die beiden Outputs Zeile für Zeile, indem es das Unix-Kommando *diff* benutzt. Die Ergebnisse dieses Abgleichs werden in einen SAS®-Datensatz geschrieben, um diese dann weiter bearbeiten zu können. Dazu wird das SAS®-Makro DELETE_LINES aufgerufen. Hiermit werden alle Zeilen gelöscht, die in beiden Versionen identisch sind. Außerdem werden Zeilen gelöscht, die irrelevante Änderungen beinhalten. Was in diesem Zusammenhang vom SAS®-Programmierer als irrelevant definiert wird, kann im SAS®-Makro DELETE_LINES angegeben werden. Üblicherweise sind dies zum Beispiel die Fußnoten der TLG, die den Text „Accovion“ und das Erstellungsdatum enthalten, oder die Kopfzeilen, in denen neben Kunden- und Projektname auch der Hinweis DRAFT oder FINAL angegeben ist.

```
%MACRO DELETE_LINES;
   if index(compress(upcase(diff)),"ACCOVION") gt 0 then delete;
   if index(compress(diff),"&company") gt 0 then delete;
   ...
%MEND DELETE_LINES
```

Abbildung 8: Auszug aus dem SAS®-Makro DELETE_LINES

Alle relevanten Änderungen werden im „Output Change Report“ gelistet. Sind keine Unterschiede zwischen den Versionen vorhanden, bleibt dieser Report leer.

```
Comparison of outputs against outputs in
  /sas_data4/VALIDATION/RMC/out/validated
                              Changed Files - Details

------------------ Filename=ae002l ----------------------------

Old/
New                                    Differences

O    0009   10/F/    Asthma/    13APR2000/ 04MAY2008/   Mild  PER/
N    0009   10/F/    Asthma/    13APR2008/ 04MAY2008/   Mild  PER/
```

Abbildung 9: Auszug aus dem „Output Change Report“ für TLG

Der inhaltliche Abgleich von Grafiken ist als .eps oder .cgm nicht möglich. Um sicherzustellen, dass sich eine Grafik nicht geändert hat, könnte man die der Grafik zugrundeliegenden Daten in einem Datensatz abspeichern oder mit PROC PRINT in einem SAS®-Listfile ablegen und diese dann mit der neuen Version vergleichen.

5 Weiterentwicklung und Herausforderungen

Ungeachtet der Möglichkeiten, die das RMC-Makro bisher schon bietet, wird es auf Wunsch der SAS®-Programmierer und Biostatistiker immer weiter entwickelt.
Die Änderungskontrolle ist bisher auf SAS®-Datensätze und SAS®-Listfiles ausgelegt. Jedoch ist die steigende Nachfrage nach rtf-Dokumenten die Überlegung wert, ob man das RMC dementsprechend erweitert.

6 Zusammenfassung

RMC unterstützt die Auswertung klinischer Studiendaten von Beginn der Programmerstellung bis zum finalen Lauf aller statistischen Studienergebnisse. Die Versionskontrolle von SAS®-Programmen in der Pharmazeutischen Industrie ist bereits gängige Praxis, der Fokus des RMC-Makros dagegen liegt auf der Änderungskontrolle der mit SAS® erzeugten Outputs (Datensätze, Tabellen, Listen, Grafiken). Zusätzlich ermöglicht RMC den wiederholten Lauf aller SAS®-Programme zu einem Projekt „auf Knopfdruck“. Die zusammenfassenden Reports, die RMC erstellt, zeigen den Status jedes Programms und weisen auf Fehler/Notes im SAS®-Log hin.
In der Validierungsphase eines Projektes ermöglicht RMC die Archivierung vorhergehender Outputs, den Neulauf aller SAS®-Programme und die automatische Änderungskontrolle. Mit Hilfe der erzeugten Reports kann der SAS®-Programmierer die Änderungen schnell nachvollziehen und verifizieren, beziehungsweise als unbeabsichtigte Änderungen identifizieren. Die von RMC erzeugten Reports können zusammen mit anderen Validierungsdokumenten abgelegt werden. Speziell am Ende eines Projektes, wenn die Validierung abgeschlossen ist und die Auswertung finalisiert ist, können nachträgliche Änderungen ganz genau anhand der RMC-Reports nachvollzogen werden. Erfolgt am

Ende eines Projektes der finale Lauf aller SAS®-Programme, stellt RMC sicher, dass sich an den validen Outputs nichts geändert hat. Auf diese Weise ist gewährleistet, dass unbeabsichtige Änderungen nicht übersehen werden.