Why data sharing is important: a case study using proprietary mutagenicity and skin sensitisation data

Dr. Donna Macmillan
Scientist

42nd ICGM – 11th November 2015

donna.macmillan@lhasalimited.org

Talk structure

(1) Why data sharing is important and how data is used

(2) Case study using Ames data (mutagenicity)

(3) Case study using LLNA data (skin sensitisation)

(4) Conclusions

Why is data sharing important?

- Encourages collaboration which benefits the scientific community
- Gaps in the chemical space covered by *in silico* models can exist
  - Derek Nexus alerts are built mainly on public data
- By donating proprietary data, these gaps can be filled
  - Model chemical space unique to each member
  - Can improve predictivity in the chemical space most important to members
  - Generalise models for mutual benefit

How is data shared?

- In-house
  - Allowing Lhasa access to confidential data set to be used in-house by scientist(s) at Lhasa
  - Faster results/feedback
  - Relatively easy to carry out
- On-site
  - Allowing Lhasa access to confidential data set only at member's site
  - Slightly slower overall process
  - Requires more organisational input/admin from Lhasa

How does it work?

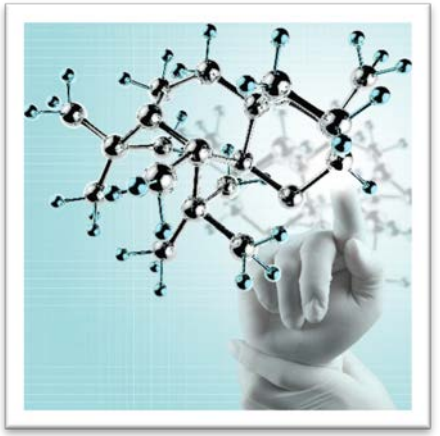
- **Prior to data sharing**
- Information about member data given to determine value to Lhasa
  - Size of data set
  - Number of FN, FP
- Contract/confidentiality agreements negotiated
- **During data sharing project**
  - No formal supervision required from member
  - Regular meetings with member and Lhasa scientist take place
    - Report/update written as per meeting
    - Discuss progress and adapt work as required

How does it work?

- **During data sharing project (continued)**
- Meetings with Lhasa colleagues also take place
  - Provide additional expertise/support
    - Discuss scope/mechanism/alert modifications
  - Signed off by member before sharing of any structures
    - Alert style substructures
- Real-time development of local models during project
  - e.g. custom KB
  - Allows continuous review of model performance

What kind of data do we need?

- Structure
  - SD file
  - SMILES
- Experimental results
  - Ames - strain data
  - LLNA - EC3 data
  - Binary results (positive/negative)
  - Complete experimental output useful
- Molecular properties/physicochemical parameters can be calculated by Derek Nexus/other programs



How do we use member data?

- Check that the data is complete
  - Curated if required
- Analyse the data
  - Whole data set
  - False negatives (FN)
  - False positives (FP)
- Analysis usually carried out using cluster analysis
  - By-eye analysis may be easier for smaller data sets
- Create new alerts and/or alert modifications
  - Implemented into Derek Nexus if public data/mechanistic rationale supports alert

Logistics of on-site visit

- Supplied by member
  - Desk space
  - Internet connection
  - Member data(!)
- Supplied by Lhasa
  - Scientist
  - Laptop

A case study…mutagenicity

Mutagenicity in Derek Nexus

- 122 mutagenicity alerts
- 25% of alerts contain proprietary data
- Comprehensive coverage of endpoint
  - Aromatic amines and boronic acids are still of significant interest and require refinement
- Derek Nexus performance against public data is very good

	Metrics (%)					Results				
Data set	**Se**	**Sp**	**PP**	**NP**	**Acc**	**TP**	**FP**	**TN**	**FN**	**Total**
Public	83	75	79	79	79	2908	762	2247	595	6512

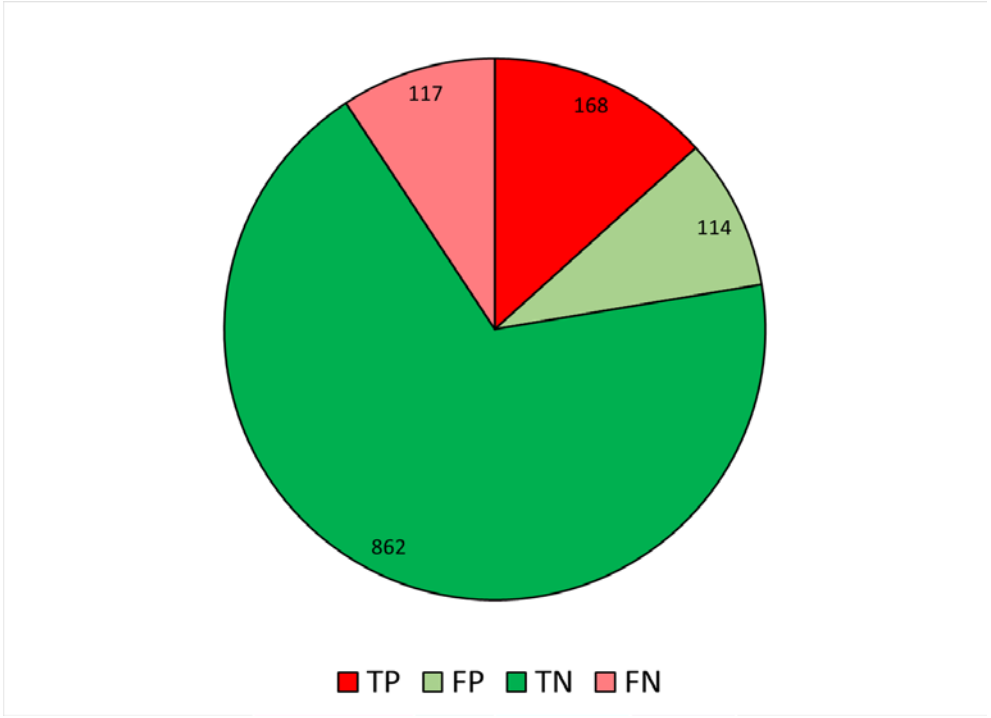
Member data set - Data

- Data curation
  - Conservative call assigned to compounds with discordant results
  - Duplicate compounds removed from data set
  - Lhasa MolID was given to each compound to anonymise the data
- Data composition
  - 1261 compounds with Ames test data
    - 1244 with Standard Ames (2-5 strains)
    - 17 with Mini Ames

Lhasa MolID	TA98 S9+	TA98 S9-	TA100 S9+	TA100 S9-
Lhasa_001	1	0	1	0
Lhasa_002	0	0	0	0
Lhasa_003	1	1	1	1
Lhasa_004	0	0	1	1

Member data set - Performance

- 1261 compounds
- Mainly negative results
  - Bias = 77% negative
- 114 FP
- 117 FN

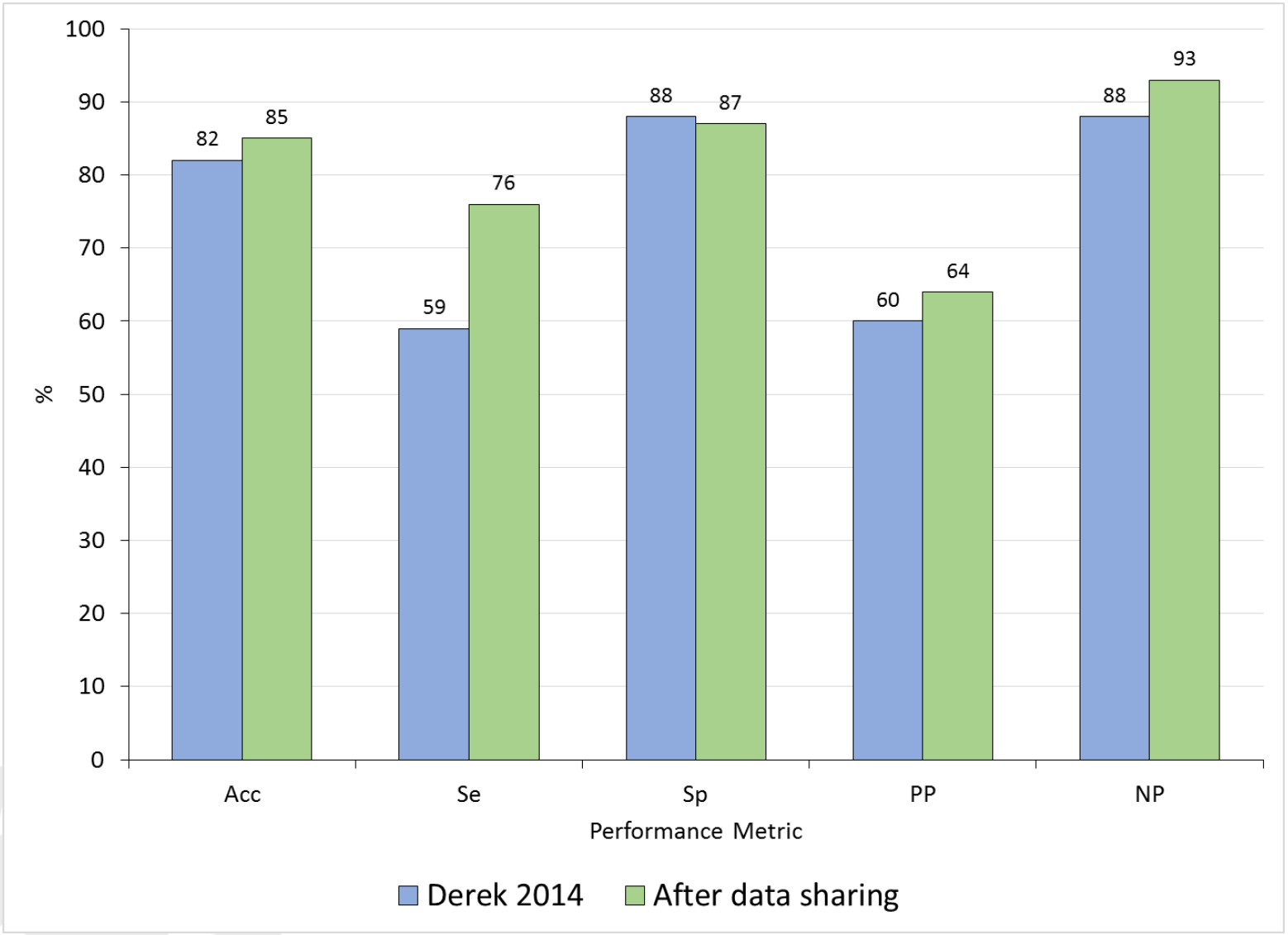


	Mutagenicity									
	Metrics (%)					Results				
Data set	Se	Sp	PP	NP	Acc	TP	FP	TN	FN	Total
Public	83	75	79	79	79	2908	762	2247	595	6512
Member	59	88	60	88	82	168	114	862	117	1261

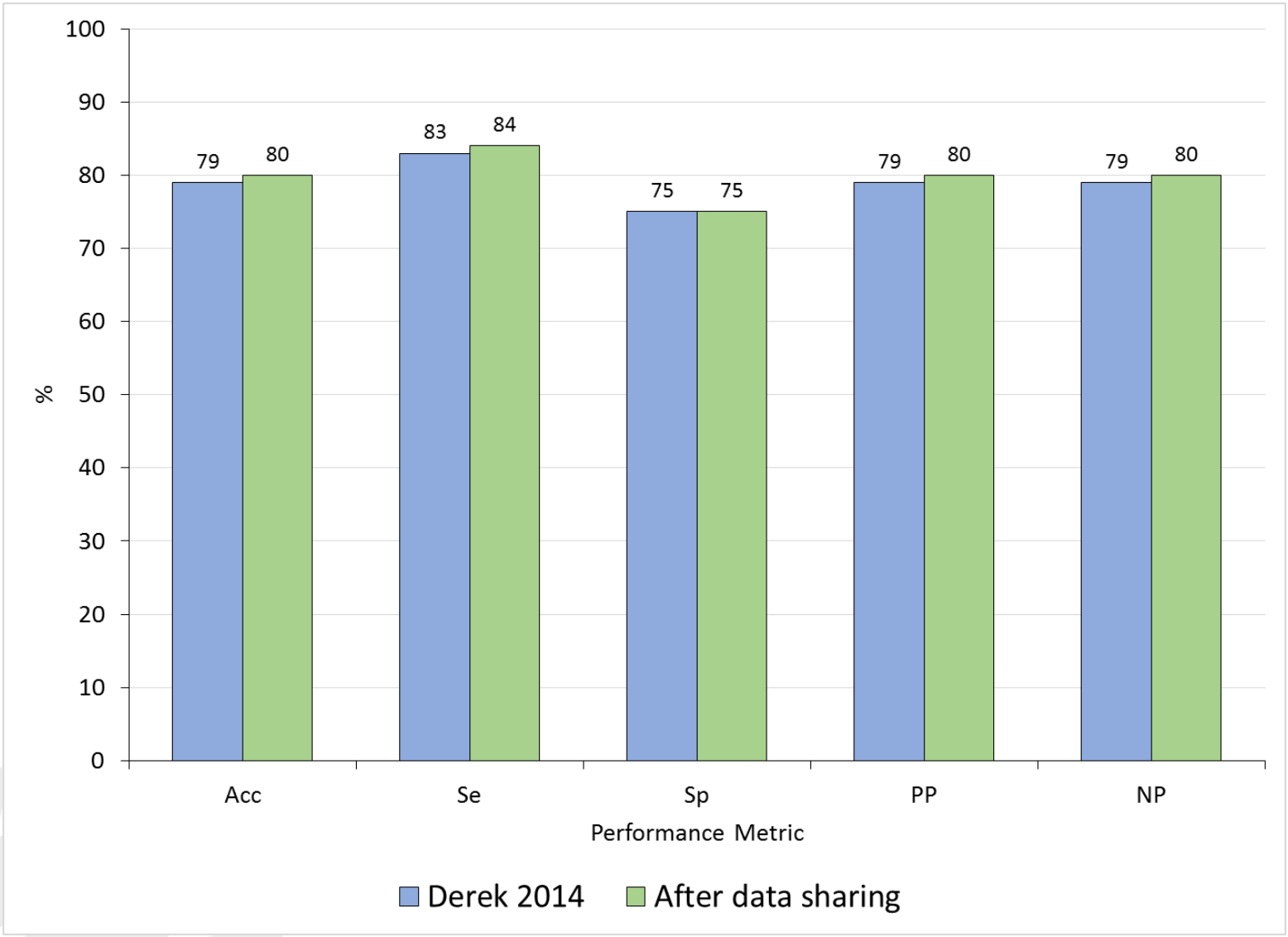
Member data set – New/modified alert summary

- 5 new alerts
  - Amine (x4)
  - Boronic acid
- 4 modifications to existing alerts
  - Azide, hydrazoic acid or azide salt
  - Alkyl aldehyde
  - Arylhydrazine
  - Arylboronic acid or derivative
- 4 potential new alerts/alert modifications require more data/mechanistic support

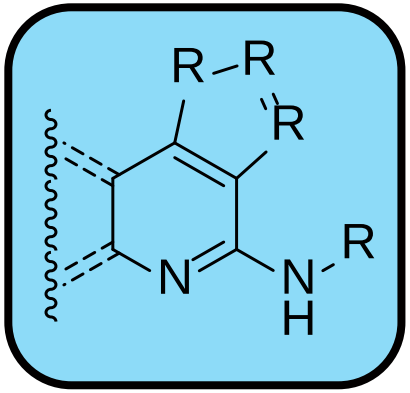
Results - Member data - Mutagenicity



Results - Public data - Mutagenicity

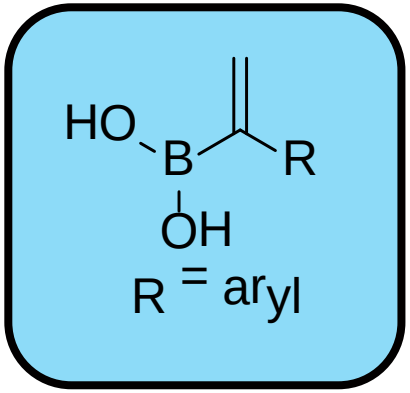


Alert example - aryl amine



- New aryl amine alert
  - 46 compounds in member data set
    - 26 positive in Ames
    - 20 negative in Ames
    - 0% PP in Derek 2014
  - After alert implementation
    - 63% PP in new KB

Alert example - styrene boronic acid



- New boronic acid alert
  - 2 compounds in member data set
    - Both positive in Ames
  - After alert implementation
    - 100% PP in new KB
  - Mechanism still under investigation
    - Originally thought that only aryl boronic acids were Ames positive

A case study…skin sensitisation

Skin sensitisation in Derek Nexus

- 80 skin sensitisation alerts
- Good coverage
  - Ongoing KB development work on this endpoint
  - Using proprietary data assists in making these improvements more relevant to member chemical space
- Performance against public data is good

	Metrics (%)					Results				
Data set	**Se**	**Sp**	**PP**	**NP**	**Acc**	**TP**	**FP**	**TN**	**FN**	**Total**
Public	77	70	73	76	74	1020	385	910	296	2611

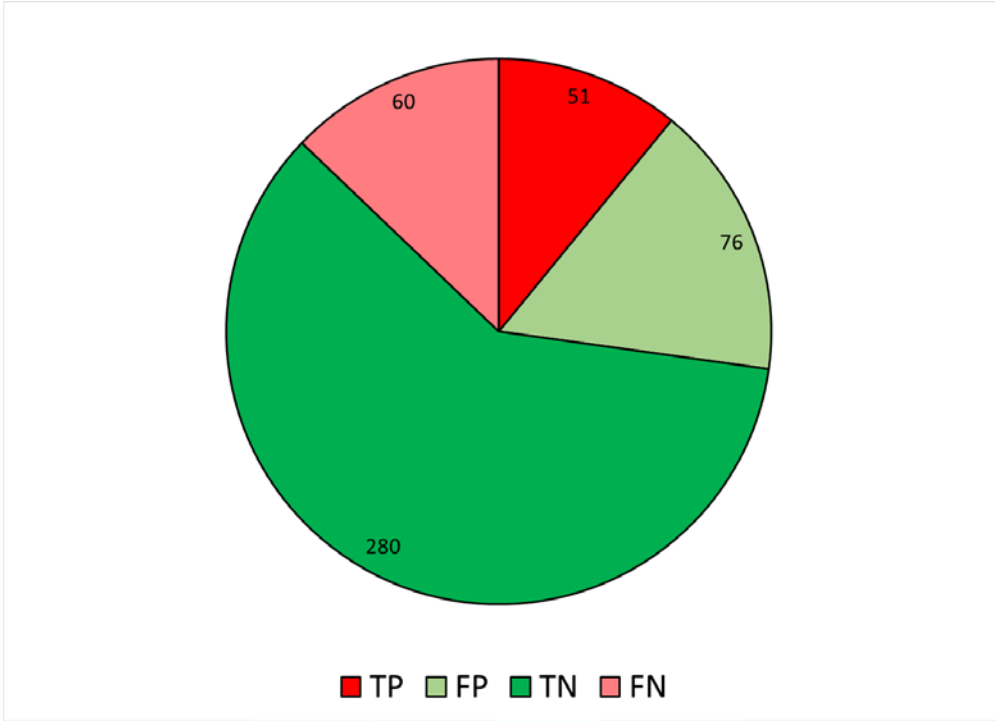
Member data set - Data

- Data curation
  - Conservative call assigned to compounds with discordant results
  - Duplicate compounds removed from data set
  - Lhasa MolID was given to each compound to anonymise the data
- Data composition - skin
  - 467 compounds with LLNA data
    - 346 with standard LLNA
    - 121 with 1% LLNA

Lhasa MolID	LLNA	EC3
Lhasa_001	0	
Lhasa_002	1	15
Lhasa_003	1	1.8
Lhasa_004	1	0.24

Member data set - Performance

- 467 compounds
- Mainly negative results
  - Bias = 76% negative
- 74 FP
- 62 FN

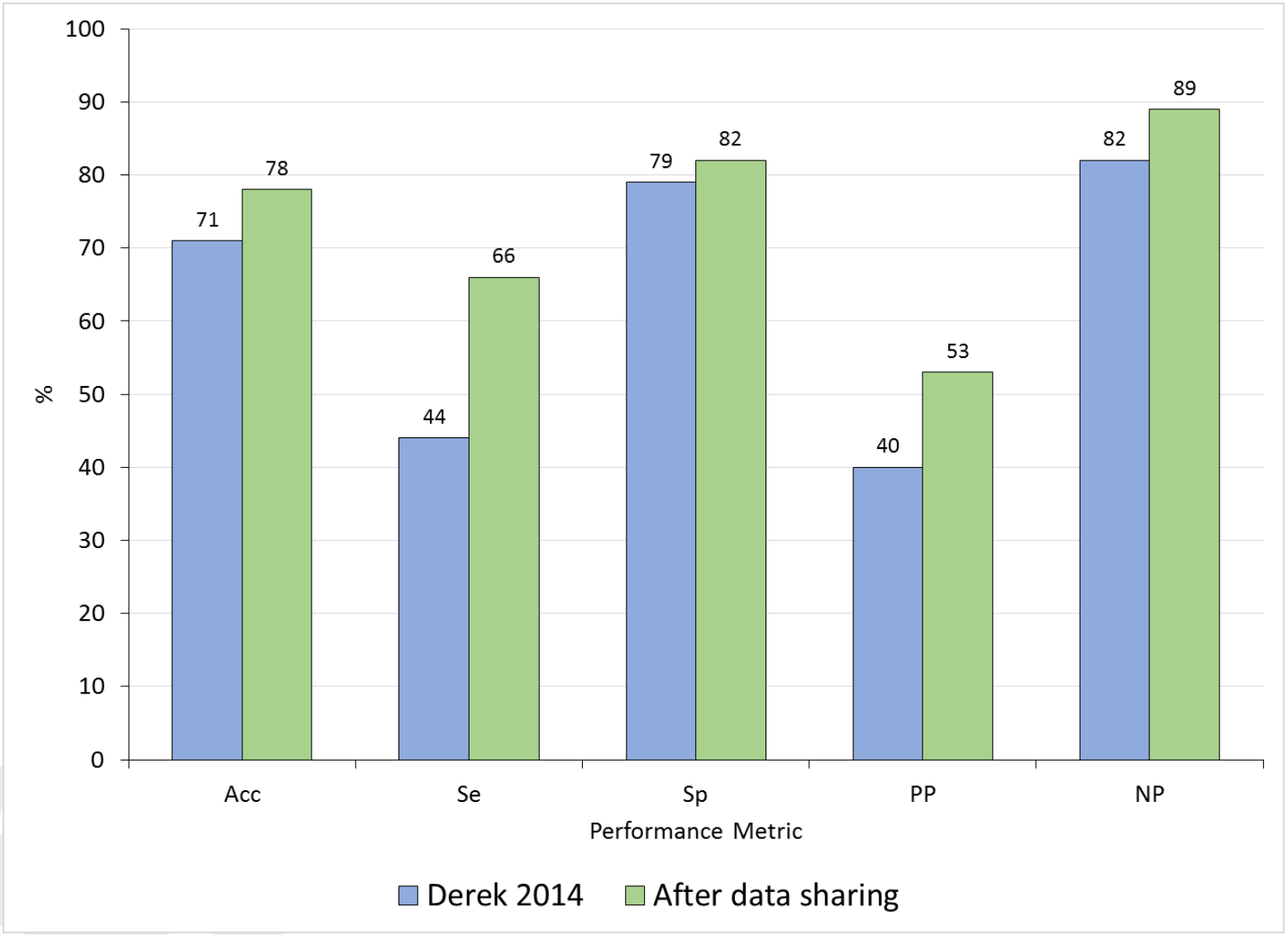


	Skin									
	Metrics (%)					Results				
Data set	Se	Sp	PP	NP	Acc	TP	FP	TN	FN	Total
Public	77	70	73	76	74	1020	382	910	296	2611
Member	44	79	40	82	71	49	74	282	62	467

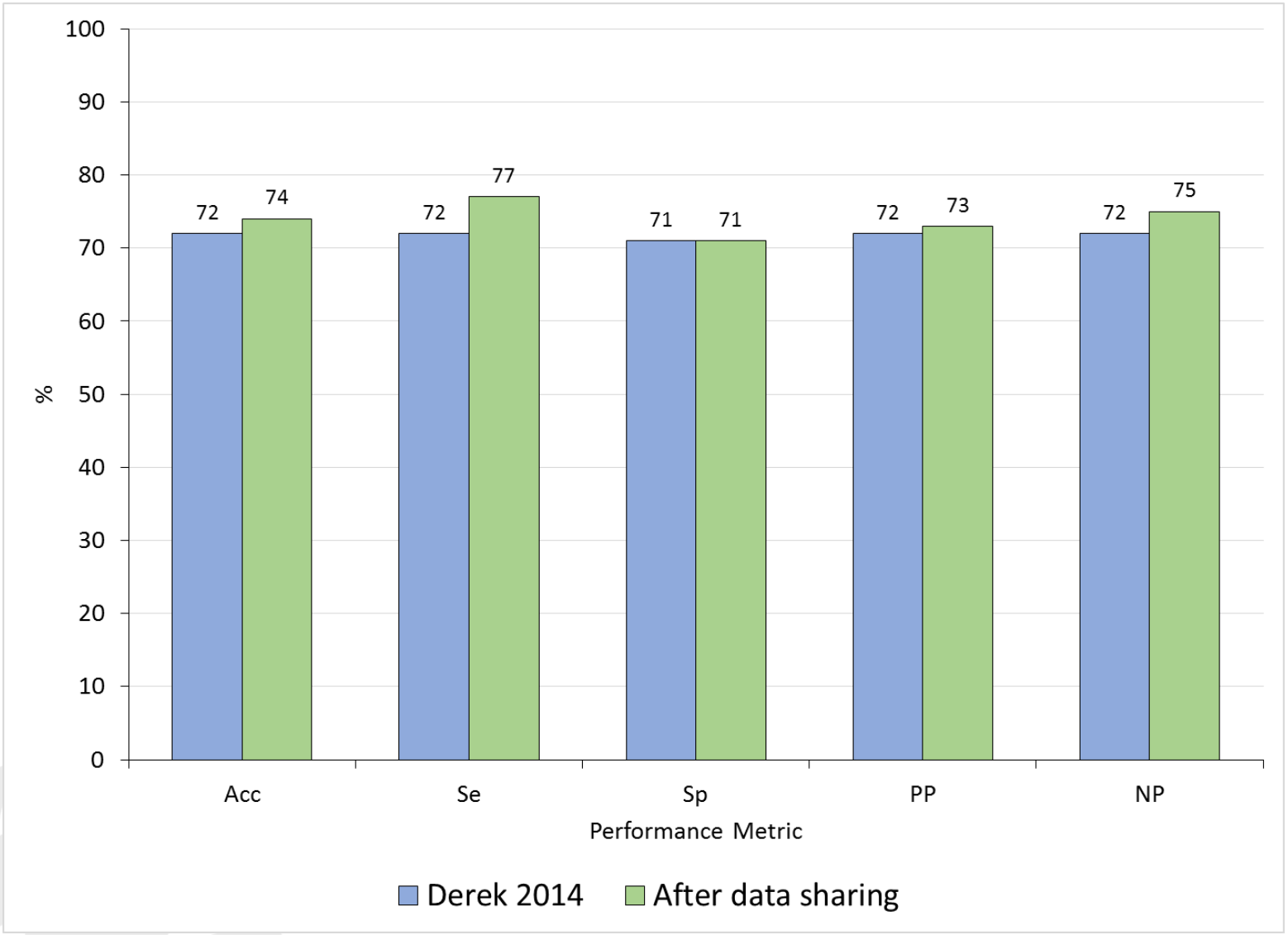
Member data set – Alert summary

- 7 new alerts
  - Aziridine
  - Activated heterocycle
  - Phosphoryl azide
- 5 modifications to existing alerts
  - Activated *N*-heterocycle
  - Activated pyridine, quinoline or isoquinoline (x 2)
  - Substituted phenol or precursor
  - Imine or alpha,beta-unsaturated imine
- 5 potential new alerts/alert modifications require more data

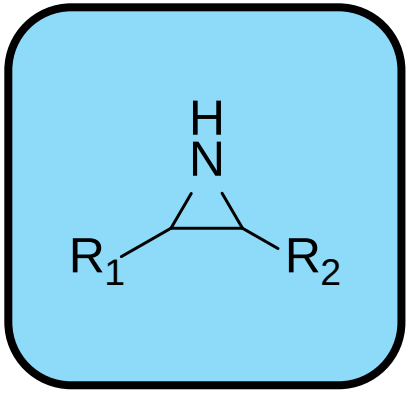
Results - Member data



Results - Public data

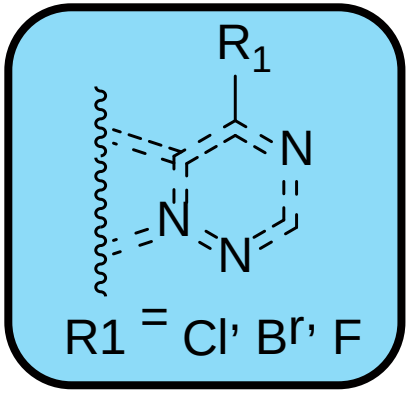


Alert example - aziridine



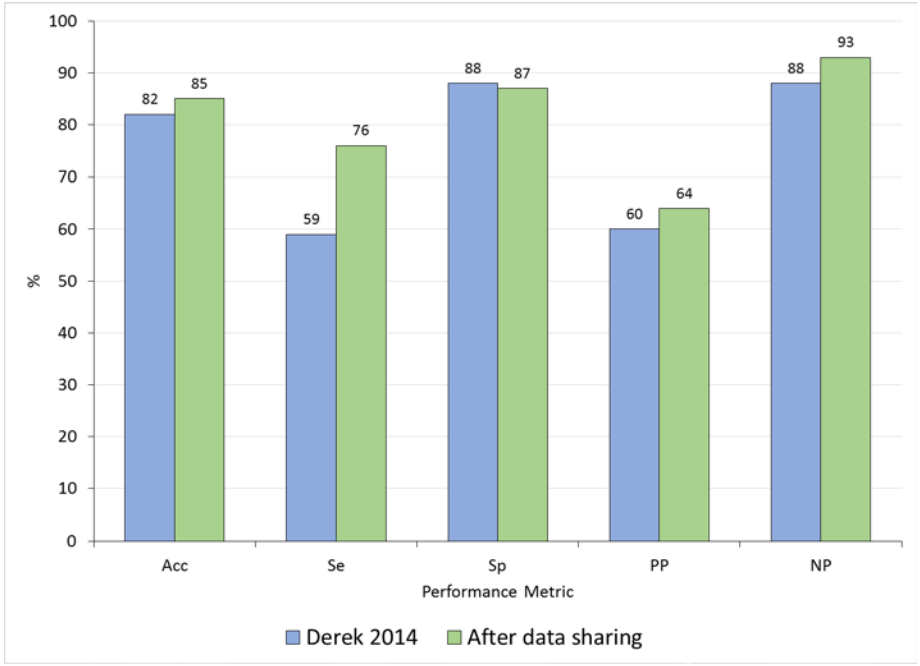
- New aziridine alert
  - 1 compound in member data set
    - Strong sensitiser in LLNA
  - 2 FN in public data set were predicted correctly after implementation of new alert
    - Sensitiser in humans

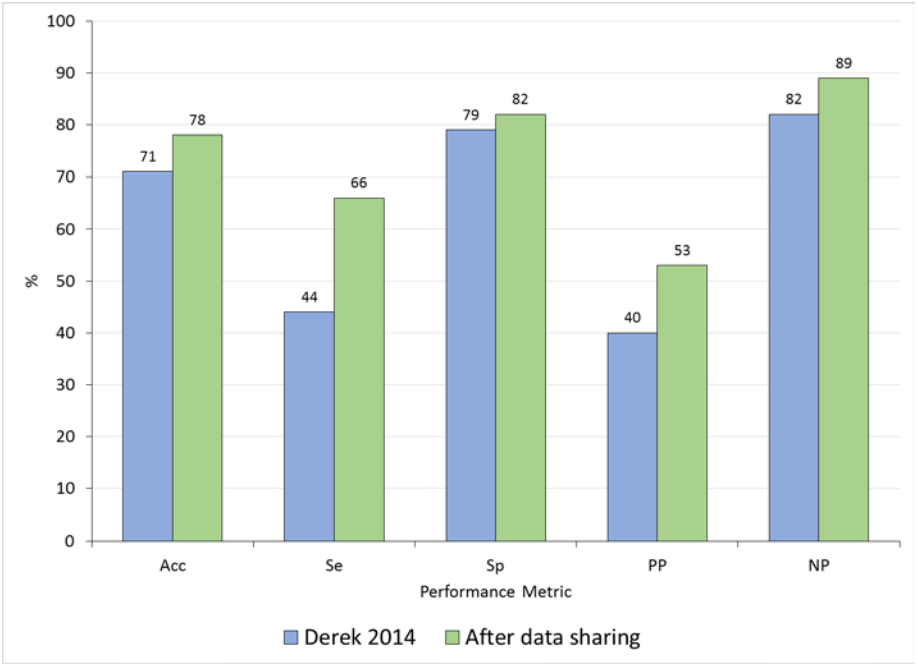
Alert example - activated heterocycle



- New activated heterocycle alert
  - 6 compounds in member data set
    - Extreme/strong sensitiser in LLNA
  - 5 FN in public data set were predicted correctly after implementation of new alert

Summary





- Data sharing greatly improves predictivity of member data
  - In particular, sensitivity can be improved without adversely affecting specificity
- Public data set predictivity is also improved
  - Increased chemical space coverage useful to all members

Conclusions

- Successful data sharing has led to improvements in mutagenicity/skin sensitisation chemical space coverage
  - Predictivity of (large) public data sets improved by a few percentage points
- Major improvements in predictivity of proprietary data
  - 17% and 22% increase in Se and 5% and 7% increase in PP for mutagenicity and skin sensitisation, respectively
- Benefits both Lhasa and all members
  - 21 alerts/alert modifications being implemented into Derek Nexus from the two member data sets shown
  - Released early 2016

Conclusions

- Collaborative publication in the pipeline
  - Joint poster abstracts submitted to SOT 2016
- The success of the data sharing project has led to other data sharing initiatives being organised with the member discussed and other members

If any members are interested in discussing a data sharing opportunity please contact our Business Development Director

liz.covey-crump@lhasalimited.org

Acknowledgements

- Steven Canipa
- Richard Williams
- Everyone at Lhasa Limited
- The member who donated data

Thank you for listening

Questions?

Lhasa Limited Registered Office
Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS
Registered Charity (290866)

+44 (0)113 394 6020
info@lhasalimited.org
www.lhasalimited.org